Journal of Language Evolution, 2018, 145–162
doi: 10.1093/jole/lzy007
Research article

Bayesian phylolinguistics reveals the internal structure of the Transeurasian family

Martine Robbeets[1]* and Remco Bouckaert[2]

[1]Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany and [2]Centre of Computational Evolution, University of Auckland, Auckland, New Zealand

*Corresponding author: robbeets@shh.mpg.de

Abstract

The historical connection between the Transeurasian languages, i.e. the Japonic, Koreanic, Tungusic, Mongolic, and Turkic languages, is among the most disputed issues of historical linguistics. Here, we will combine the power of classical historical-comparative linguistics and computational Bayesian phylogenetic methods to infer a phylogeny of the Transeurasian languages. To this end, we will use lexical etymologies supporting the reconstruction of proto-Transeurasian forms with meanings that belong to the Leipzig-Jakarta 200 basic vocabulary list. Our application of Bayesian phylogenetic inference to the classification of the Transeurasian languages is unprecedented. In addition to the methodological implications for Bayesian inference applied to proposed language phyla at relatively deep time depths and with relatively sparse sets of surviving daughter languages, our research has also factual implications for the existing theories of Transeurasian relationships. Our results move the field forward in that they provide a quantitative basis to test various competing hypotheses with regard to the internal structure of the Transeurasian family.

Key words: Bayesian phylolinguistics; pseudo Dollo model; phylogeny; basic vocabulary; Transeurasian; Altaic

1. Introduction

The term 'Transeurasian' refers to a group of geographically adjacent and structurally homogeneous languages across Eurasia that consists of up to five different families: the Turkic, Mongolic, Tungusic, Koreanic, and Japonic languages. It was coined by Johanson and Robbeets (2010: 1–2) to complement the traditional term 'Altaic', which we reserve for the unity of the Turkic, Mongolic, and Tungusic languages only. Figure 1 shows twenty-three contemporary Turkic languages, ten Mongolic languages, ten Tungusic languages, six Japonic languages in addition to Korean. These numbers approximate the number of languages per family recognized by Glottolog (i.e. twenty-seven Turkic, seventeen Mongolic, thirteen Tungusic, fifteen Japonic, and two Koreanic).

The question of whether these five groups descend from a single common ancestor has been the topic of a long-standing debate. As early as 1692, Nicolaes Witsen first mooted the contours of the Transeurasian language family, but Ramstedt is usually considered the founder of Transeurasian linguistics because he established a modern linguistic framework for Transeurasian comparison, supported by regular sound correspondences (1957) and morphological cognates (1952). While, until the late sixties, the field focused on the comparison of Turkic, Mongolic, and Tungusic on the one side (e.g. Poppe 1960, 1965, 1975) and of Korean and

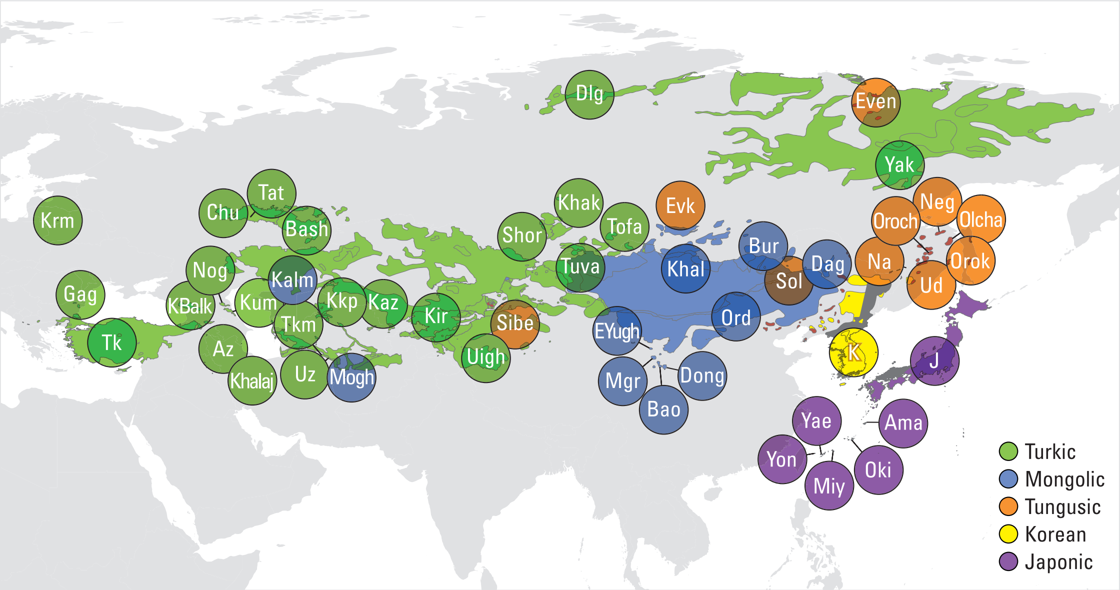

Figure 1. The distribution of the Transeurasian languages. Abbreviations for languages are explained as follows: Ama.: Amami; Az.: Azerbaijani; Bao.: Bao'an; Bash.: Bashkir; Bur.: Buriat; Chu.: Chuvash; Dag.: Dagur; Dlg.: Dolgan; Dong.: Dongxiang; EYugh.: Eastern Yughur; Even: Even; Evk.: Evenki; Gag.: Gagauz; J: Japanese; Kalm.: Kalmuk; KBalk.: Karachay-Balkar; Krm.: Karaim; Kkp.: Karakalpak; Kaz.: Kazakh; Khak.: Khakas; Khal.: Khalkha; Khalaj: Khalaj; Kir.: Kirgiz; K: Korean; Kum.: Kumyk; Ma.: Manchu; MK: Middle Korean; MMo.: Middle Mongolian; Miy.: Miyako; Mogh.: Moghol; Mgr.: Monguor; Na.: Nanai; Neg.: Negidal; Nog.: Nogai; Oki.: Okinawa; Olcha: Olcha; OJ: Old Japanese; OT: Old Turkic; Ord.: Ordos; Oroch: Oroch; Shor: Shor; Sibe: Sibe; Sol: Solon; Tat.: Tatar; Tofa.: Tofalar; Tk.: Turkish; Tkm.: Turkmen; Tuva: Tuva; Ud.: Udehe; Uigh.: Uighur; Uz.: Uzbek; Yae.: Yaeyama; Yak.: Yakut; Yon.: Yonaguni.

Japanese on the other (e.g. Martin 1966), in the seventies, Miller's (1971) monograph 'Japanese and the other Altaic languages' increased the scholarly interest in the overall comparison of these languages. Clauson (1956) and Doerfer (1963–1975) raised substantial criticism against the genealogical relatedness of these languages, which was mainly based on the alleged lack of basic vocabulary and the explanation of all correlations by borrowing. Starostin et al. (2003) resurrected scholarly interest in the Transeurasian unity, accumulating a body of evidence that was far more impressive in quantity and rich in empirical material than the number and scope of etymologies proposed previously. However, these new matches were, in their turn, criticized for reason of phonological, morphological, or semantic overpermissiveness, among others by Robbeets (2005), leaving room for a reduced core of reliable etymologies and by Vovin (2005, 2009, 2010) and Georg (2007), completely rejecting all evidence advanced so far. For an elaborate discussion of the history and the current state of the debate, see Robbeets (2017).

Robbeets (2005, 2015) has shown that even if the majority of support provided in the past is questionable, there is nonetheless a core of reliable evidence for the classification of Transeurasian as a valid genealogical grouping. In line with the requirements of the classical comparative method of historical linguistics, the evidence consists of regular sound correspondences, lexical etymologies including common basic vocabulary and shared verb morphology. As a result, the hypothesis that the Transeurasian languages are related is gradually gaining acceptance in the field (Gözaydin 2006; Rozycki 2006; Büyükmavi 2007; Décsy 2007; Kara 2007; Dybo 2016).

Whereas supporters of Transeurasian affiliation basically agree about the unity of the family, they do not necessarily coincide on its internal structure. Here, we set up four different hypotheses of classification that are representative of the variation in the different classifications proposed in the past. To this end, we will use contemporary and historical lexical data, which yield proto-Transeurasian reconstructions corresponding to an item on the Leipzig-Jakarta 200 basic vocabulary list. By applying Bayesian phylogenetic methods to the phylogeny of the Transeurasian languages, our aim is to infer which model is best supported by the data. In this way, we intend to provide a quantitative basis to determine the internal structure of the Transeurasian family.

2. Previous proposals of classification

Over the last century, various hypotheses have been suggested on the basis of either the classical historical-comparative method or lexicostatistic methods (Vladimircov 1929: 44–47; Street 1962: 95; Poppe 1965: 147; Miller 1971: 44; Baskakov 1981: 14; Tekin 1994: 82; Starostin et al. 2003: 236; Blažek and Schwarz 2014; Robbeets 2015: 506).

All classifications proposed so far agree that, first, if a Japonic branch is postulated, Koreanic and Japonic are more closely related to one another than to any of the other branches concerned and, second, that Mongolic forms a binary unity with either Turkic or Tungusic, distinct from the Japano-Koreanic branch. The main difference in the proposals so far has to do with the position of Tungusic vis-à-vis the other branches: Does Tungusic represent a first-order split, which separated simultaneously with Japano-Koreanic and Mongolo-Turkic? Does Tungusic cluster with Japano-Koreanic or does it rather belong with Mongolic and Turkic? And, if the latter is the case, does Tungusic stand in a binary unity with Mongolic or not?

Given these issues, the first set of proposals concerns a polytopology (Fig. 2; Hypothesis A in Fig. 6) whereby Tungusic separated simultaneously from Japano-Koreanic and Mongolo-Turkic. The second set of representations involves a binary topology in which Tungusic clusters with the Japano-Koreanic branch, separately from the Mongolic and Turkic branches (Fig. 3; Hypothesis B in Fig. 6). The third set of conceptions also reflects a binary topology, but here Tungusic clusters with the Mongolic and Turkic branches to form a separate 'Altaic' unity, bifurcated from the Japano-Koreanic unity (Fig. 4; Hypothesis C in Fig. 6). In this view, Tungusic stands in a binary unity with Mongolic and Turkic is the first to branch off from the Altaic unity. Finally, completing the set of logically possible hypotheses with regard to the position of Tungusic, we hypothesised a fourth possible scenario, whereby Turkic stands in a binary unity with Mongolic and Tungusic is the first to branch off (Fig. 5; Hypothesis D in Fig. 6). To enhance comparability, most figures represent simplified versions of the original classifications suggested by the respective scholars. They are adapted from the original, for instance, by leaving out the sub-branching of the individual proto-families, omitting designations for intermediate stages or turning the trees in the horizontal direction.

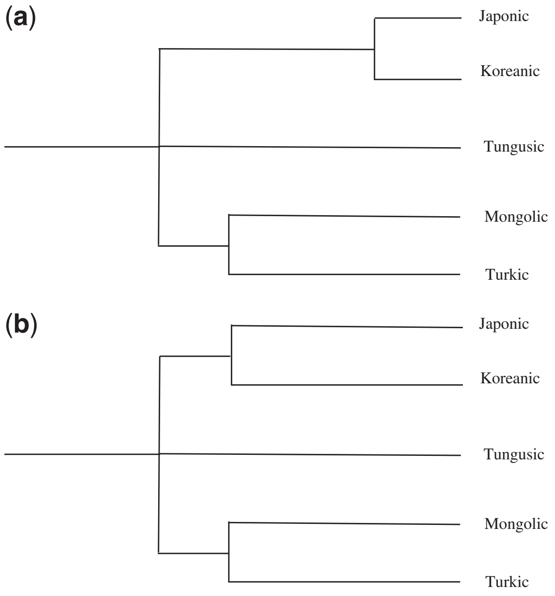

Figure 2. Previous classifications suggesting a polytopology for the Transeurasian family (Hypothesis A in Fig. 6) (Baskakov 1981: 14; Starostin et al. 2003: 236).

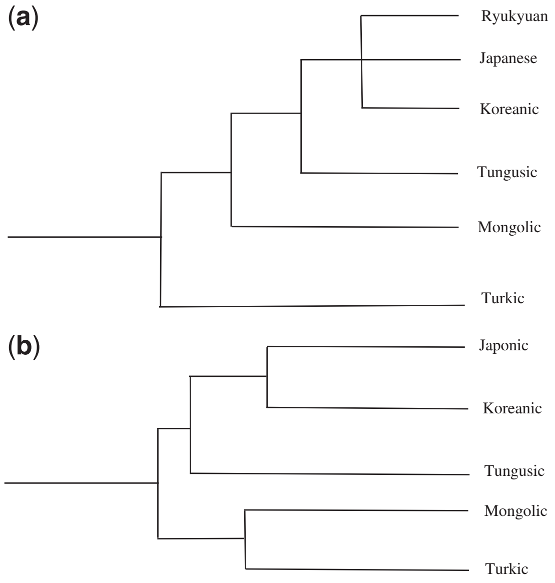

Figure 3. Previous classifications suggesting a binary topology for the Transeurasian family, whereby Tungusic clusters with the Japano-Koreanic unity (Hypothesis B in Fig. 6) (Miller 1971: 44; Blažek & Schwarz 2014: 90).

The polytypology (Fig. 2; HA) is the classification supported by the so-called 'Moscow School'. It was first proposed by Vladimircov (1929: 44–47) and lived on in the view of Baskakov (1981: 14), both scholars using the classical historical-comparative method. More recently, however, this hypothesis was confirmed on the basis of lexicostatistic methods by Starostin and his colleagues. It conceives of the Transeurasian family as

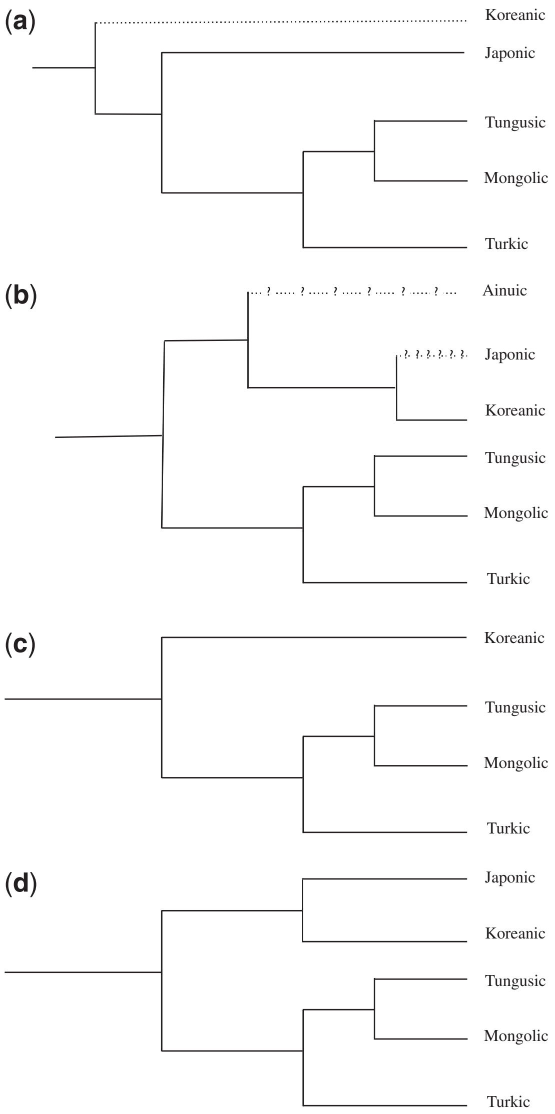

Figure 4. Previous classifications suggesting a binary topology for the Transeurasian family, whereby Tungusic clusters with the Altaic unity and Turkic branches off first (Hypothesis C in Fig. 6) (Street 1962: 95; Poppe 1965: 147; Tekin 1994: 82; Robbeets 2015: 506).

consisting of three principal groups: Turko-Mongolic, Tungusic, and Japano-Koreanic. However, contrary to the classical conception, in Startostin's view, Turko-Mongolic and Japano-Koreanic separated around the same time, in the fourth millennium BCE.

A binary topology whereby Tungusic clusters with the Japano-Koreanic unity is favored by some Transeurasian linguists in the West, especially by specialists in Japanic and Koreanic languages (Fig. 3; HB in

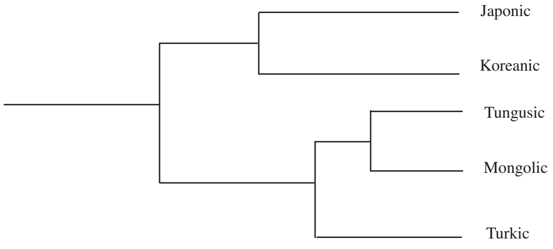

Figure 5. Hypothesized classification suggesting a binary topology for the Transeurasian family, whereby Tungusic clusters with the Altaic unity and branches off first (Hypothesis D in Fig. 6).

Fig. 6). Miller (1971: 44) proposed a unity between Tungusic, Koreanic and Japonic, which recalls the suggestion made by Unger and the Altaic panel (1990: 481) to limit the Transeurasian reconstructions to a 'Macro-Tungusic' perspective, consisting of Tungusic, Koreanic, and Japonic languages only. However, unlike Unger's proposal, Miller conceives of the position occupied by the Ryukyuan languages as independent from Mainland Japanese. The view of a separate unity between Tungusic, Koreanic, and Japonic, which was initially reached by using the classical historical-comparative method, was recently supported by Blažek and Schwarz (2014: 90) application of lexicostatistic methods. However, they conceive of Mongolo-Turkic as the second unity making up the Transeurasian family.

Most western scholars involved in the classification of the Transeurasian languages using the classical historical-comparative method agree on a binary topology for the Transeurasian family, whereby Tungusic is classified in a unity with Turkic and Mongolic (Fig. 4; HC in Fig. 6). Poppe (1960, 1965: 147) included Korean as a separate branch of Altaic but later he remarked that 'Korean is a language only partly belonging to the field of Altaic studies' (Poppe 1975: 172), referring to the possibility that Korean could be a non-Transeurasian language imposed on a Transeurasian substratum. This possibility is indicated with a dotted line in Fig. 4. In his review of Poppe (1960), Street (1962: 95) suggested a different configuration for the Japanese and Korean branches, speculating that the Japano-Koreanic branch could eventually cluster with Ainu. The dotted line with the question marks in Fig. 4 represents Street's uncertainty about the inclusion of Japanese and Ainu. Tekin (1994: 82) included Koreanic in the classification, assuming that proto-Koreanic was first to branch off from the Transeurasian unity, but he did not accept the inclusion of Japonic into the family. Robbeets' (2015) tree confirms the basic classification of this 'western' school.

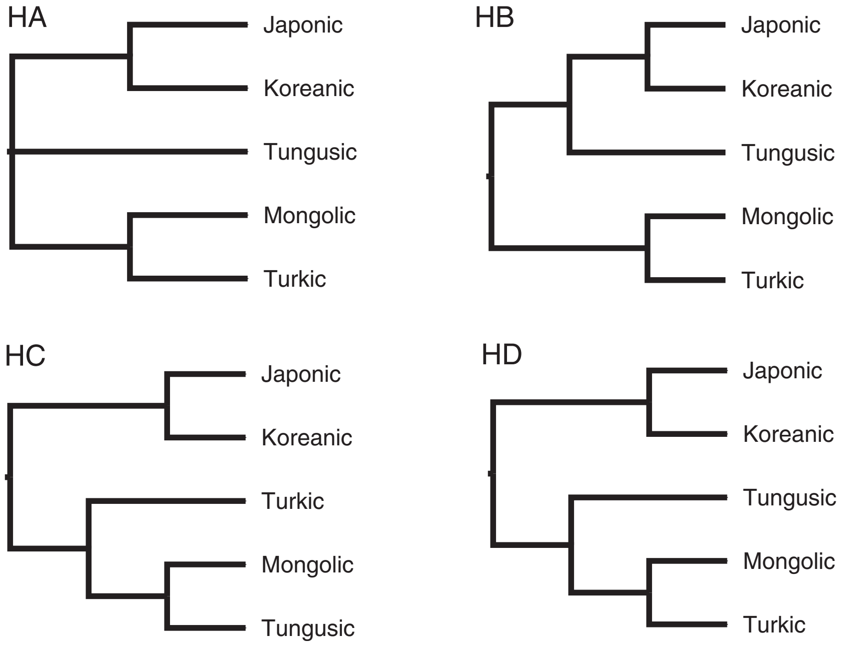

Figure 6. Summary of the four overarching hypotheses that we tested.

Our research starts from the three most recently proposed hypotheses that are representative for each set of proposals, given in Figs 2–4, and we add Fig. 5 (HD in Fig. 6) to complete the set of logical possibilities with regard to the position of the Tungusic branch: While Tungusic stands alone in HA, it clusters with Japonic and Koreanic in HB and with Altaic in the remaining hypotheses, whereby in HC it stands in a binary unity with Mongolic and in HD it does not.

In addition, there are some extralinguistic factors, related to the genetic and archaeological past, that motivate the inclusion of hypothesis D in the set of alternatives. Recent paleogenetic studies (Wang et al. 2016; Siska et al. 2017; Jeong and Wang in press, 2019) indicate that contemporary speakers of Tungusic languages are genetically continuous with ancient individuals of the Neolithic Boisman culture (4825–2470 BC) in the Southern Primorye.[1] Contemporary Turkic and Mongolic speakers share this ancestral eastern lineage, albeit with an increasing western admixture from the Bronze Age onwards. In addition, archaeologists show that the basic subsistence strategy in the Southern Primorye during the Neolithic and Bronze Age was agriculture supplemented by fishing, hunting, and gathering, while pastoralism was gradually adopted in the homelands of Turkic and Mongolic speakers West of the Liao River and on the Eastern steppes (Taylor 2016; Taylor et al. 2017; Jeong and Wang in press, 2019). Thus, genetics and archaeology converge on a scenario whereby the linguistic ancestors of the Tungusic speakers separated at an early time from the linguistic ancestors of Turkic and Mongolic speakers, who from the Bronze Age onwards started to share not only an increasing degree of genetic admixture with Western steppe herders but also adopted a common pastoralist subsistence strategy.

For the purposes of this article, it is legitimate to group the above nine existing classifications into four distinct hypotheses because some variations involve mere timing or uncertainty about the inclusion of a certain subgroup, but do not affect the essential structure of the family. The variation in Fig. 2 is due to a difference in branch length and thus absolute time depth, the variation in Fig. 3 has to do with whether the separation of Turkic vis-à-vis Mongolic follows a breakaway or a binary split model, and the variation in Fig. 4 is the result of uncertainty about the inclusion of Japonic or Koreanic as a subgroup of the family. However, none of these differences involves basic family structure. Hence, we aim to test the four overarching hypotheses summarized in Fig. 6.

[1] The use of the term ‘continuous’ in this context implies that the ancient genome continued as a virtually unbroken line into the contemporary genome and that it would be plotted on top of the contemporary genome in a principal component analysis.

Table 1. Conservative vs. alternative approach to the coding of basic vocabulary item 80. WOOD.

Conservative coding of WOOD				
Japanse	Korean	Evenki	Khalka	Turkish
ki	*namu*	*moo*	*mod(on)*	*tahta*
‘tree, wood’	‘tree, wood’	‘tree, wood’	‘tree, wood’	‘wood, plank’
0	0	1	1	0
Alternative coding of WOOD				
Japanse	Korean	Evenki	Khalka	Turkish
mori	*mey*	*moo*	*mod(on)*	—
‘woods’	‘hill’	‘tree, wood’	‘tree, wood’	—
1	1	1	1	0

3. Alternative coding principle

Classical Bayesian approaches to language classification usually start from a basic vocabulary list and code 1 for words with identical basic meaning that display a cognate form and 0 for such words lacking a cognate form. Even if this coding principle, illustrated for the basic item WOOD in Table 1, would yield reliable results for the internal structure of the individual daughter languages, it is expected to leave only a weak historical signal at a deeper Transeurasian level.

This is due to the considerable time depth of the Transeurasian family combined with a relatively high extinction rate of the daughter languages. By way of comparison, the Austronesian and Indo-European language families respectively count about 1,200 and 445 living languages, whereas hardly 50 Transeurasian languages have survived to the present. Thus, for language families in general, with the elapse of time, meanings will have diverged in such a way that at best only a few cognate sets with identical basic meanings are left. However, in the Transeurasian case, many languages that may have reflected these identical meanings will have died out. Hence, the effect of cognate attrition for the Transeurasian languages is expected to be much stronger than that for the Austronesian languages because first, the common ancestor of the former is estimated around 4700 BCE, while the latter is only around 3500 BCE, and second, because much fewer Transeurasian languages have survived.[2] Therefore, restricting our coding to cognate sets with identical meaning would fail to produce a meaningful Transeurasian tree.

Given this challenge, we have chosen to apply an alternative coding principle, illustrated in Table 1, which starts from the Transeurasian reconstructions displaying a basic meaning, e.g. pTEA **mɔrɔ* ′tree, wood′. Here we code 1 for presence of a cognate in a daughter language and 0 for absence of a cognate, irrespective of whether the meanings are identical or not. In our example, we code 1 for Japanese and Korean in addition to Evenki and Khalkha. This coding principle yields a stronger historical signal because it triggers a higher number of positive values than the classical principle.

Applying this coding principle to our data, we reach the list of codings given in the Supplementary Data (SI 4). In order to illustrate our expectation that the traditional coding would produce less historical signal than the alternative one, we added a further specification to the coding of our basic vocabulary data (Supplementary Data, SI 5). Here, we code 0 for the absence of a cognate, 1 for the attestation of a semantically equivalent cognate and 2 for the attestation of cross-semantic cognate. On a total number of 2,735 cognates, we count 1,621 cognates with equivalent meaning and 1,114 cross-semantic cognates. In the conventional cognate coding, the total number of forms yielding code 1 would be equal to or lower than 1,621. The expectation of a number lower than the number of semantically equivalent cognates is motivated by the convention that if there is competition between two forms with equivalent meaning in the traditional approach, the one with the most frequent use is selected. When using the cross-semantic coding, the number of cognates in the dataset is notably larger (2,735) compared to the number of semantically strictly equivalent cognates (≤1,621). By consequence, the traditional coding is expected to produce less positive values, i.e. less 1 codings and more 0 codings.

4. Dataset

In our study, we use lexical data from fifty contemporary and seven historical varieties of the Transeurasian languages, including twenty-three Turkic languages in addition to Old Turkic, ten Mongolic languages in

[2] Lexicostatistic dating methods such as Blažek and Schwarz (2009) reach 4750 BC for the timedepth of Transeurasian, a date, which converges with our own Bayesian estimation of the root age of Transeurasian as 4700 BC. As the latter timedepth does not substantially differ from that of our prior, we cannot draw firm conclusions on the dating or proto-Transeurasian on the basis of Bayesian inference alone, but taken together with the lexicostatistic result, it may serve as an indication. There is a relative consensus as far as the timedepth of Austronesian is concerned; see Blust (2013: 24–29).

addition to Written Mongolian and Middle Mongolian, ten Tungusic languages in addition to Manchu and Jurchen, Korean in addition to Middle Korean and Japanese and five Ryukyuan languages in addition to Old Japanese. We collected these data by consulting dictionaries in combination with written sources against the background of previous etymological proposals evaluated in Robbeets (2005). Historical varieties are only used for the purpose of reconstruction and not integrated as data-points in the tree. This is because historical varieties were reconstructed using the comparative method due to which they possess a certain level of uncertainty which we do not prefer to include in our analyses. By restricting ourselves to a limited number of contemporary languages, we make abstraction from many different local varieties, such as for instance the various dialects spoken in neighboring villages on the Ryukyuan Islands that are in variation with the five local languages of Amami, Okinawa, Miyako, Yaeyama, and Yonaguni.

Judging from the number of languages recognized in Glottolog, our set of languages is amply sampled; for Turkic twenty-three languages out of twenty-seven are represented, for Japonic six out of fifteen, Koreanic one out of two, Mongolic ten out of seventeen, and Tungusic ten out of thirteen, so there is no potential that our sampling strategy causes the problems suggested in the biological literature (Hillis et al. 2003).

The lexical data we collected correspond regularly in form and function to such an extent that they yield proto-Transeurasian reconstructions for items that occur on the Leipzig-Jakarta 200 basic vocabulary list (Tadmor et al. 2010). For the regular vowel and consonant correspondences underlying our reconstructions, we refer to the Supplementary Data (SI 1). We reached 150 etymologies spread over 107 distinct items of the basic vocabulary list.

The cognate coding allows for a certain degree of semantic development in the daughter branches—in other words, inexact meaning correspondence between lexical items coded as cognates—but only to the extent that the proposed semantic development between proto-Transeurasian and the daughter branches answers to cross-linguistically observed polysemies, semantic associations, or grammaticalizations. The penultimate column of Supplementary Table 3 (SI 2) indicates the cross-linguistic availability—and hence, the regularity or acceptability—of a proposed semantic correspondence. To this end, we consulted the database of cross-linguistic colexifications published by List et al. (2014), which brings together instances where two or more meanings are simultaneously covered by the same lexical item in a certain language. The number in the penultimate column corresponds to the number of the relevant semantic community in List et al. (2014), while the number behind the slash is the number of nodes in that community. Whereas the number of the community is an arbitrary number for identification, the number of nodes (n) represents how many meanings can be joined together in the same network. This means that a given set of meanings compared in Supplementary Data (SI 2) represents an acceptable semantic correspondence that ranks among a total of n permissible semantic associations.

The tag ‘Gram’ means that the development is a cross-linguistically well-attested grammaticalization process. This is for instance the case for the development of the verb (25) ‘to do, make’ in an iconic pro-verb following sound symbollic expressions (Heine and Kuteva 2002: 112–13), the development (39) ‘this’ from a demonstrative into a personal pronoun, (Heine and Kuteva 2002: 119–20) and the development of the interrogative pronoun (50) ‘what?’ into an interrogative particle. Finally, the tag ‘Poly’ marks a semantic development that is supported by a polysemy, which is not reported in List et al. (2014), although it is found across the languages of the world and/or in one or more Transeurasian lexemes of the dataset. For instance, the development (12) ‘breast’ into ‘heart’ is acceptable, even if it is lacking in List’s database. That is because it is attested in a number of languages in Australia, such as Arabana and Wangkanguru, in which the term for ‘heart’ is a reduplication of the term for ‘chest’ (Wilkins 1996: 289). Moreover, it is also found on a distinct, unrelated etymon in a Transeurasian language, notably Yonaguni *ccimu* ‘heart, liver’, which is derived as Yo. *ccimuti* ‘breast’. Using cross-linguistically observed patterns of colexification, semantic association, and grammaticalization in this way, we can provide an empirical base for the degree of semantic latitude permitted in our Transeurasian comparisons.

We appended an overview of the basic vocabulary shared between the Transeurasian languages (Supplementary Data, SI 2) in addition to a detailed documentation of the underlying etymologies (Supplementary Data, SI 3). This dataset is new in the sense that it expands, revises, and updates the etymologies evaluated and proposed in Robbeets (2005, 2015). We expanded the previous dataset with etymologies for basic items that do not have a Japanese cognate, we consistently added cognates from the Ryukyuan languages, we carried out a detailed morphological analysis in order to delimit the roots more precisely and to identify petrified suffixes more accurately and, we answered to criticism in reviews of earlier etymologies.

The response to this criticism is included in the discussion of individual etyma in Supplementary Data (SI 3). Among the ten reviews of the lexical evidence in Robbeets (2005), six were very (Gözaydin 2006; Rozycki 2006; Büyükmavi 2007; Décsy 2007) to mildly positive (Kara 2007; Dybo 2016), while four were negative (Knüppel 2006; Georg 2007; Miller 2007 and Vovin 2009). Of the 359 lexical etymologies proposed as core evidence in Robbeets (2005), seventy-three were criticized in the reviews. Among these, twenty-three etymologies for basic vocabulary present in the Leipzig-Jakarta list in Supplementary Data (SI 2/3) are met with objection, i.e. 1 fire, 3 to go, 5 a/b mouth, 7 blood, 8 bone, 12 breast, 30a tooth, 32c big, 39a this, 46 bite, 54 new, 55 burn, 56 not, 63 soil, 68 skin, 80a wood, 84 ash, 92 shade, 94 salt, 96 wide, 97 star, and 99 hard. In some cases, where we considered the criticism legitimate, we left out the problematic part of the etymology under discussion. In other cases, we answered to the cricitism and motivated our decision to leave the etymology unchanged.

It is highly unlikely that all similarities between the basic items in our dataset are the result of contact instead of genealogical relationship. Traditionally, the strength of basic vocabulary lies in the fact that words with basic meanings tend to resist borrowing more successfully than random lexical items. The very fact that we find 150 Transeurasian etymologies covering 107 distinct basic vocabulary concepts thus is a strong argument against borrowing by itself. In addition, we can advance other arguments against borrowing, such as (1) the misfit with the expected borrowing hierarchy; (2) the misfit with the expected typology of verbal borrowing; (3) the regularity and complexity of sound correspondence; (4) the occurrence of broken contact chains; (5) the multiple setting; and (6) the well-spread distribution of the cognates; see also Robbeets (in press, 2019).

First, among the concepts of the Leipzig-Jakarta list, we find fifty-nine actions, thirty-two property words, twenty-three deictic or grammatical items and eighty-six nominal concepts. Out of ninty-one concepts for actions and property words, we find fifty-nine Transeurasian verbal etymologies, which means that as much as 65% of the basic verbal concepts on the Leipzig-Jakarta list are etymologized. Out of twenty-three concepts for deictic and grammatical items, we find thirteen etymologies, which implies that 57% of the basic deictic and grammatical items on the list are etymologized. Out of eighty-six nominal concepts, we find thirty-seven etymologies, indicating that only 43% of the nominal concepts are covered by a Transeurasian etymology. Empirically, it is observed that languages tend to borrow lexical items more easily than grammatical ones and nouns more easily than verbs (a.o. Wohlgemuth 2009; Matras 2009; Tadmor et al. 2010). In contrast to this tendency, there are more correlations for verbs (65%) and deictic and grammatical items (57%) in the Transeurasian basic vocabulary than for nouns (43%). This observation indicates that it is unlikely that the comparative sets can be explained by borrowing, as borrowing would be expected to yield more correspondences in nouns than in verbs and grammatical markers.

Second, as far as the mechanisms of loan verb accomodation are concerned, most recipient languages can be categorized into two distinct groups: borrowed verbs either arrive as verbs, needing no formal accommodation, or, they arrive as nonverbs and need formal accommodation. Most Transeurasian languages can be assigned to the second group because they display a clear preference for the nonverbal strategy (Wohlgemuth 2009: 159, 161). If the thirty Transeurasian verbal etymologies in our basic vocabulary list would be the result of borrowing, they would represent instances of the verbal strategy. This would run against the observable preference of the Transeurasian languages to apply the nonverbal strategy to loan verbs.

Third, the comparative sets for basic vocabulary display regular correspondences for each consonant of the verb root and for each but the root-final vowel, conform to the requirements in Supplementary Data (SI 1). Even if extensive contact can result in systematic sound correspondences, it is unlikely that this is the case here because some sound correspondences reflect divergence to such an extent that they cannot be attributed to the mere imitation of model sounds in a process of borrowing. This is, for instance, the case for the homoganic and heteroganic cluster correspondences 5, 6, 11, 12, 17, and 18 in Supplementary Table 1 (SI 1).

Fourth, gaps in the attestation of members of an etymology, whereby a cognate is absent in one or more intermediate contact branches are indicative of borrowing. The absence of a Korean member in the etymology (12) 'breast', for instance, makes a borrowing scenario whereby the word got borrowed directly from Tungusic into Japonic rather unlikely.

Fifth, most examples of borrowing have a binary setting in common: they typically go from a model language into a recipient language. Especially for verbs and grammatical markers, examples of the same item progressing into a third or fourth language are relatively rare. However, the Transeurasian unity consists of five families and it is especially grammatical markers and verbal concepts that is well distributed over all branches. Since repetitive borrowing is particularly rare for verbs and grammatical items, this observation argues against borrowing.

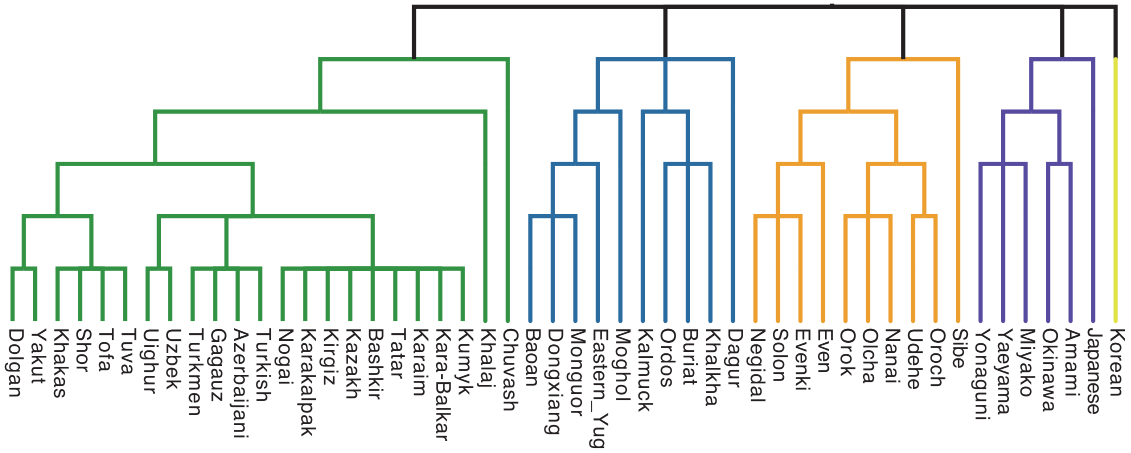

Figure 7. Monophyletic constraints imposed on the tree based on classification in Glottolog. Families are colored according to their family membership as in Fig 1.

Finally, the distribution of a certain basic item to a single language or to only few languages of a certain subgroup could serve as an indication of borrowing. However, such cases do not occur among our basic vocabulary etymologies. Based on his hypothesis that West Old Japanese or its immediate predecessor absorbed a large number of loanwords from Old Korean, Vovin (2010) proposed to reject all cognates that are missing in the Ryukyuan languages as probable loanwords. However, out of 101 proto-Japonic forms in our basic vocabulary list, 82 are supported by reflexes in the Ryukyuan languages, corresponding to 81% of all proto-Japonic reconstructions. The solid distribution of Japonic cognates in the Ryukyuan languages reduces the probability of borrowing from Old Korean.

5. Calibrations

Since the amount of data is relatively small, we introduce as much of the prior information that we have into the analysis as we can. By using rooted time trees in our analysis instead of undirected trees, we capture the fact that these languages evolved through time, which helps restrict possible histories. We know that languages evolved through time, so we should use that information in the model. We add two kinds of prior information to the tree: monophyletic constraints restricting possible combinations of language groupings and timing information restricting the age of internal nodes in the tree. Figure 7 shows the set of monophyletic constraints within language family phylogenies from Glottolog (Hammarström et al. 2017), which are considered non-controversial. Note that no constraints are implied above the family level.

As far as timing restrictions is concerned, we added rather generously sized time intervals for the most recent common ancestors of the language families (nodes coinciding with proto-Turkic, proto-Mongolic, proto-Tungusic, and proto-Japonic), but not for Korean because it is the sole living descendant of proto-Koreanic. We supported our estimates of the upper and lower time limit for each node (Table 2) mainly with information from linguistic and historical sources.

Linguists and archeologists associate proto-Japonic with the beginnings of Yayoi-culture (900 BCE–300 CE) on the Japanese Islands (Hudson 2002; Robbeets 2005; Unger 2009; Whitman 2011; Miyamoto 2016). Proto-Japonic, the ancestor of Mainland Japanese and the Ryukyuan languages, is thought to have separated when Yayoi culture started to spread northeastwards over the Japanese Islands in the early centuries BCE. This chronological estimation is in line with Lee and Hasegawa's (2011) Bayesian phylogenetic analysis, dating Proto-Japonic divergence at 182 BCE. Lexicostatistic approaches such as Hattori's (1976: 43) estimated the breakup between 0 and 500 CE, while the Automated Similarity Judgement Program (ASJP) yielded 436 CE, be it with a margin of error of 29% (Holman et al. 2011). The ancestor of the languages now spoken in the Ryukyuan Islands is thought to have remained in northeastern Kyushu until around 900 CE, when full-scale agriculture was introduced to the Ryukyus. The derivation of Ryukyuan from an early Kyushu dialect is consistent with the distribution of the main accent types over Japan and across the Ryukyu Islands and may reflect different waves of founder populations to different islands in the Ryukyu chain at

Table 2. Upper and lower time limits for the most recent common ancestors of the language families on the Transeurasian tree. From these limits, normal distributions with mean μ and variance σ (last column) were fitted such that the 95% HPD coincides with end and start times. The normal distributions were used as calibrations in the Bayesian analysis.

Constraints	Lower bound	Upper bound	Calibration
Proto-Japonic	500 CE	200 BCE	Normal ($\mu\mu$=150 BCE, sσ=175)
Proto-Tungusic	500 CE	600 BCE	Normal ($\mu\mu$=50 BCE, $\sigma\sigma$=275)
Proto-Mongolic	1300 CE	1000 CE	Normal ($\mu\mu$=1150 CE, sσ=75)
Proto-Turkic	100 CE	500 BCE	Normal ($\mu\mu$=200 BCE, $\sigma\sigma$=200)

slightly different times (Unger 2009: 105–6, 211, de Boer in press, 2019). In sum, linguists estimate the breakup of proto-Japonic between 200 BCE and 500 CE.

Chronologically, Proto-Tungusic is a relatively shallow entity. Applying Starostin's lexicostatistic methods, Korovina (2011) dated Proto-Tungusic to the sixth century BCE, but other computational methods such as Bayesian inference and the AJSP yielded much younger dates, notably 200 CE (Oskolskaya et al. unpublished manuscript) and 681 CE (Holman et al. 2011: 854), respectively. Referring to the name change in Chinese dynastic chronicles of the Tungusic ethnonym 'Yilou' to 'Wuji', Robbeets (2015: 16–18) situated the breakup of Proto-Tungusic at the end of the Han period (206 BCE–220 CE). Pevnov (2012: 32) estimated that Proto-Tungusic could not be younger than two thousand years on the basis of a rough measure of mutual intelligibility. A chronological interval roughly between 500 BCE and 500 CE was supported by Janhunen (2012: 8), who placed the breakup of Proto-Tungusic in the Iron Age, in line with the diversification of other language families in the area. In sum, linguists estimate the time depth of proto-Tungusic between 600 BCE and 500 CE.

Proto-Mongolic is nearly equivalent with the language spoken by the historical Mongols around the time of the Mongol Empire (1206–1368), which is documented in historical sources, written in several different scripts and collectively termed Middle Mongol. Some of the written varieties of Middle Mongol contain features that reflect dialectal forms slightly different from, or even earlier than the stage of Proto-Mongolic. Therefore, the depth of the Mongolic family, as measured on the basis of both written documents and living languages included in our dataset is no more than 700 to 1000 years (Rybatzki 2003), although the AJSP yields a date as early as 267 BCE (Holman et al. 2011: 854). Nevertheless, given the close similarity with Middle Mongolian, it is reasonable to estimate the time depth of proto-Mongolic between 1000 and 1300 CE. However, there were historical languages related to Mongolic spoken also in southwestern Manchuria, and perhaps even further south. The most unambiguous evidence comes from Khitan, the dynastic language of the Liao Empire (907–1125). The Khitan lineage may be traced back to several historical and prehistorical ethnopolitical groups in the region, including the Tabghach of the Northern Wei (386–534), the Xianbei or 'Serbi' (208 BC–235 AD) and the Donghu (the first millennium BC). The written traces of these groups are too fragmentary to be included in our dataset, but they seem to represent an extinct branch parallel to the Proto-Mongolic lineage. Although the information on Khitan and its presumed ancestors is still very scarce, we may date the breakup of the Macro-Mongolic to a time level preceding Proto-Mongolic by at least several centuries, most probably to the very period of the Xianbei.

Based on evidence from contact linguistics, the earliest split between the two principal branches of Turkic, i.e. Bulgharic and Common Turkic, is usually dated to between the middle of the first millennium BCE and the turn of the eras (Janhunen 2009). Turkic phylogenies relying on quantitative methods basically support the lower estimate. The following dates are obtained by lexicostatistic calculations: the third century BCE (Tenishev et al. 2006), around 120 BCE (Mudrak 2009), the beginning of the first century CE (Dybo 2007). A preliminary Bayesian analysis of the Turkic family Savelyev (in press, 2019) dates the split of Proto-Turkic into Common Turkic and Bulgharic branches approximately to 200 BCE, with a highest posterior density interval between 2000 BCE and 400 CE. As this time depth coincides with the beginning of the Xiongnu empire (209 BCE–100 CE), the association of Xiongnu with Proto-Bulgharic does not seem unreasonable. However, given the relatively large credible interval involved in the Bayesian dating, the breakup of proto-Turkic may also be connected with the first disintegration of the Xiongnu confederation under influence of the military successes of the Chinese in 127–119 BCE (Mudrak 2009). In sum, the time depth of the breakup of Proto-Turkic can be estimated between 500 BCE and 100 CE.

6. Methods

Bayesian phylogenetic inference has never been applied to the Transeurasian family before. Previously, methods used for the classification of the Transeurasian family were restricted to the classical comparative method and the lexicostatistic method. The lexicostatistic method is a distance-based method, which estimates the relationship between two languages by measuring the amount of difference in shared cognate proportion between them. In contrast, the classical comparative method makes use of a character-based method in order to generate trees. More specifically, it relies 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). Here, we apply a Bayesian method using BEAST (Bouckaert et al. 2014), which is also character-based, and 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.

Cognate data was ascertained in such a way that some patterns never occur in our data: each site in the alignment contains a one in at least two families. Therefore, sites with one or more ones in only one family never occur, neither do sites with only zeros. The likelihood of the data P_{data} is affected by this sampling strategy; instead of calculating the P_{data}, we actually only calculate the likelihood of the data under condition it does not form a pattern with only ones occurring in a single family. To compensate for this, we use ascertainment correction (Felsenstein 2004), which calculates the ascertained likelihood $P_{ascertained}$ as $P_{data}/(1{-}P_{forbidden})$ where $P_{forbidden}$ is the probablity of all patterns we ruled out. We calculate $P_{forbidden}$ by calculating for each family (except Korean) the probability of patterns with question marks for each of the languages in the family and zeros for all other languages. For Korean, we calculate the probability of all languages being zero, but Korean being 1. The sum of these probabilities contains four times the probability that each language is zero, so we subtract three times the probability of each language being zero to obtain the probability of forbidden patterns $P_{\mathrm{forbidden}}$. From this we calculate the ascertained likelihood from the likelihood of the data P_{data} as follows $P_{ascertained}{=}P_{data}/(1{-}P_{forbidden})$.

A birth/death sampling process is most appropriate for tree prior when analyzing languages, and the Yule model (Gernhard 2008) is the simplest such model, which assumes a pure birth process governed by a single parameter: the birth rate. We found that more complex models resulted in a worse fit with our data (details below).

There are various models of evolution of characters along branches of a tree that may be suitable for our data. We compared three frequently used models; the binary GTR (with and without gamma rate heterogeneity over sites), covarion (Tuffley and Steel 1998), and stochastic Dollo (Nicholls and Gray 2008) models. The GTR is a variant of the general reversible model for nucleotides (Rodriguez et al. 1990) adapted for two states. The model allows cognates to become present and absent multiple times along branches, and offers a simple model of evolution. The gamma rate heterogeneity over sites model (Yang 1996) caters for different rates of evolution for different cognates and works well with the GTR model. The covarion model (Tuffley and Steel 1998), like the GTR model, allows cognates to become absent and present multiple times along a branch, but furthermore allows a character to be in a fast or a slow state. When in a fast state, changes between present and absent are very frequent, and when in a slow state changes occur much less often. More parameters are involved but the covarion model has a remarkable property that it is not very sensitive to borrowing. The stochastic Dollo model (Nicholls and Gray 2008) is based on the Dollo principle that a cognate can only appear once, but can be lost multiple times. This closely fits the definition of cognate, but is also sensitive to borrowing events. For a more detailed description of these models see, for example, Bouckaert and Robbeets (2017) or Bouckaert et al. (2012) Supplementary Data for details of these models. Furthermore, we considered the pseudo Dollo and pseudo Dollo covarion models (Bouckaert and Robbeets 2017). When tracking the history from the root of the tree to a leaf, these models allow at most one birth followed by at most one death. However, unlike the stochastic Dollo model, multiple births are allowed, for example at siblings in a tree, thus catering for borrowing events.

We assume a molecular clock implying that over time there is a certain rate of change along branches of the tree that is constant on average. However, in practice rates on branches are not constant, and a model where rates on branches are independent of other branches, but are drawn from a log normal distribution (with its mean and standard deviation estimated) fits much better and is flexible enough to capture most rate variation. This clock model is known as the uncorrelated relaxed clock with log normal distribution (Drummond et al. 2006).

The covarion model with uncorrelated relaxed clock with log normal distribution tends to perform well with cognate data (e.g. Bouckaert et al. 2012; Bowern and

Table 3. Comparing various evolutionary models. Marginal likelihood and standard deviation was estimated using nested sampling. Bayes factor calculated with respect to the best fitting pseudo Dollo covarion model with relaxed clock. A Bayes factor greater than 100 is taken as decisive support against the covarion model (Kass and Raftery 1995). Pseudo Dollo covarion with relaxed clock (bolded in table) is the best fit for our data with pseudo Dollo with gamma rate heterogeneity and relaxed clock second with a Bayes factor of 13.4. All other models are decisively outperformed ($BF \gg 100$) by the best fitting model.

Substitution model	Log marginal likelihood	Standard deviation	BF PDCov RC vs. others
Covarion, relaxed clock	−1952.0	1.03	2.62E + 62
Covarion, strict clock	−2018.0	0.88	1.24E + 91
Binary GTR, relaxed clock	−2003.8	1.53	8.26E + 84
Binary GTR, 4 gamma, relaxed clock	−1999.8	1.21	1.47E + 83
Stochastic Dollo, relaxed clock	−2472.9	1.75	4.41E + 288
Pseudo Dollo, relaxed clock	−1819.5	1.07	7.27E + 04
Pseudo Dollo, 4 gamma, relaxed clock	−1810.9	1.01	1.34E + 01
Pseudo Dollo covarion, relaxed clock	**−1808.3**	**1.06**	**1.00E + 00**
Pseudo Dollo covarion, strict clock	−1877.6	0.84	1.29E + 10

Atkinson 2012; Gray et al. 2009). To determine the most suitable model for our data, we use this as our base model and compare other models against it by estimating the marginal likelihood using nested sampling (Maturana et al. 2017) with 100 particles yielding a standard deviation of around 1. From these estimates, we calculated Bayes factors (Kass and Raftery 1995) to decide which model best fits our data. Table 3 shows the results; changing from relaxed to strict clock results in a sixty-eight point drop of the marginal likelihood with respect to our base model. Likewise, changing the substitutio model to binary GTR (with or without gamma rate heterogeneity) or stochastic Dollo model results in large drops in ML estimates. However, changing to a pseudo Dollo model (with or without gamma) increases ML estimates, but the pseudo Dollo covarion gives an even better estimates. In fact, the pseudo Dollo covarion model with relaxed clock has an overwhelmingly better fit than any other model ($BF \gg 100$ with respect to any other model).

We consider four main hypotheses that cover most of the variation in proposed groupings of the Transeurasian language families. Hypothesis A is the polytopology (HA in Fig. 6), hypothesis B represents the binary topology in which Tungusic clusters with the Japano-Koreanic branch (HB in Fig. 6), Hypothesis C the binary topology in which Tungusic clusters with Mongolic and Turkic, with Turkic branching off first (HC in Fig. 6), Hypothesis D the binary topology in which Tungusic clusters with Mongolic and Turkic with Tungusic branching of off first (HD in Fig. 6).

We test which of the four hypotheses is best supported by the data by restricting the tree space to conform to the hypotheses and estimating the marginal likelihood for each of the hypotheses. The Bayes factor is then calculated as the ratio of marginal likelihoods (in practice, we exponentiate the difference between the log marginal likelihoods). We constrained the analysis for HB, HC, and HD by simply adding monophyletic constraints in addition to the ones depicted in Fig. 7. Since we are working with binary trees, HA requires an extra condition on top of the same monophyletic constraints of HB, namely that the branch above the Tungusic clade has length zero, and the branch above Japano-Koreanic has length zero as well.

7. Results

So far, we used the simple Yule model as tree prior. The birth–death skyline (BDSKY) model (Stadler et al. 2013) is a flexible alternative that allows us to explore more complex tree priors than the one parameter Yule model. We considered splitting the time range into two epochs with one birth rate for each epoch giving a two parameter model. However, comparing the 1 epoch BDSKY model with the two epoch BDSKY model (since the BDSKY is slightly differently parameteried than the Yule model in BEAST) gives a Bayes factor of 4.9 in favour of the 1 epoch model, so we conclude that increasing the number of parameters in the model cannot be justified. The results presented below are based on the Yule prior.

Table 4 gives the priors and posteriors based on an MCMC run, and Bayes factor comparing the different hypotheses in the rows to those in the columns when using the best fitting model as shown in Table 3.

Table 4. Prior and posteriors based on an MCMC run and Bayes factors calculated from marginal likelihoods comparing the different hypotheses (HB, HC, and HD shown in Fig. 6).

Topology	Prior	Posterior	HB	HC	HD
HB	9.51%	6.27%		**291.01**	0.04
HC	8.82%	0.02%	0.00		0.00
HD	5.42%	88.58%	**24.79**	**7214.15**	

A Bayes factor between 6 and 10 is considered as strong evidence for the competing hypothesis in the row, while a factor over 10 (bold in Table 4) is considered to be very strong evidence. Therefore, Table 4 indicates that the hypothesis with the highest credibility is HD, the binary topology in which Tungusic clusters with Mongolic and Turkic with Tungusic branching off first (Fig. 6). Since HD has a BF > 20 wrt HB we conclude that HA is not plausible since if HA was the correct model we would not be able to distinguish HB from HD. The results presented in Table 4 are based on the Yule prior. Results with the next best fitting model from Table 3 shows even stronger support for HD (BF = 87.56 wrt HB). Likewise, model comparison based on marginal likelihoods with both best and second best fitting models give similar but stronger BFs than in Table 4.

Unfortunately, as far as timing is concerned, the root age of the inferred phylogeny has large uncertainty when using a Yule prior (95% HPD Interval [2.55kya, 7.54kya] a priori [4.33kya, 9.45kya] a posteriori), so we cannot draw firm conclusions on the timing aspects of our results. This is not surprising given the small amount of data compared to many other cognate-based analyses: there are just 240 informative sites in our data while it is not unusual to have over ten thousand sites (Gray et al. 2009; Bouckaert et al. 2012; Bowern and Atkinson 2012). However, we can make substantial statements about the topology.

We further captured our results in a DensiTree (Fig. 8). The internal structure of the family is close to the hypotheses proposed by Transeurasian scholarship in the past (HA, HB, or HC in Fig. 6), but does not coincide with any of them. The DensiTree classification seems to favor the classification proposed by the Russian school (Fig. 2) in that it views the Turkic branch as most closely related to Mongolic. However, similar to the view in the West (Fig. 3), it pictures Tungusic, Mongolic, and Turkic as a separate unity. We get a strong historical signal with 98.3% support for Japano-Koreanic, 90.3% for Tungusic-Mongolic-Turkic, and 100% for Mongolic-Turkic. There is 3.4% support for constructing a Tungusic branch outside the rest as proposed in Fig. 2 and 6.31% support for establishing a separate Tungusic-Japano-Koreanic unity as in Fig. 4. There is no support (0%) for the Mongolic-Tungusic unity proposed in Fig. 3.

8. Discussion

The implications of our research are two-fold. First, there are methodological implications for Bayesian phylogenetic inference applied to proposed language phyla at relatively deep time depths and with relatively sparse sets of surviving daughter languages. Second, there are factual implications for the existing theories of Transeurasian relationships.

From a methodological point of view, we applied an unconventional coding method and improved the modeling by way of a pseudo Dollo covarion model. Our choice of coding strategy allows us for the first time to infer relationships among different language families based on cognate data. The classical method of cognate judgment results in far fewer cognates being found to be shared among language families. Consequently, it does not allow inference of such relationships because it has a much more stringent requirement regarding not allowing small semantic changes in cognates. Our new coding strategy allows us to use analytical approaches that give us insight into the relationships between the various Transeurasian families. Besides, the new strategy allows to integrate classical-comparative linguistic research more effectively into Bayesian approaches as we use etymologies established by expert linguists as our dataset.

Moreover, our recently introduced pseudo Dollo covarion model with relaxed clock served as a new model for capturing evolution in our cognate sets. We showed that for our dataset it had a better fit than the currently existing models.

From a factual perspective, our results can be situated in the broader context of the scholarly debate on Transeurasian, which centers around two issues: the distinction between the effects of inheritance and borrowing and the internal classification of the Transeurasian family. Our research addresses the major objections raised against Transeurasian affiliation by showing that the Transeurasian languages share a substantial proportion of basic vocabulary, refining the evidence, answering the criticisms raised against particular basic etymologies, and arguing that not all correlations in the basic vocabulary can be explained as the result of language contact.

Moreover, our results solve a long-standing question about the exact shape of the Transeurasian tree by providing a quantitative basis to test various competing

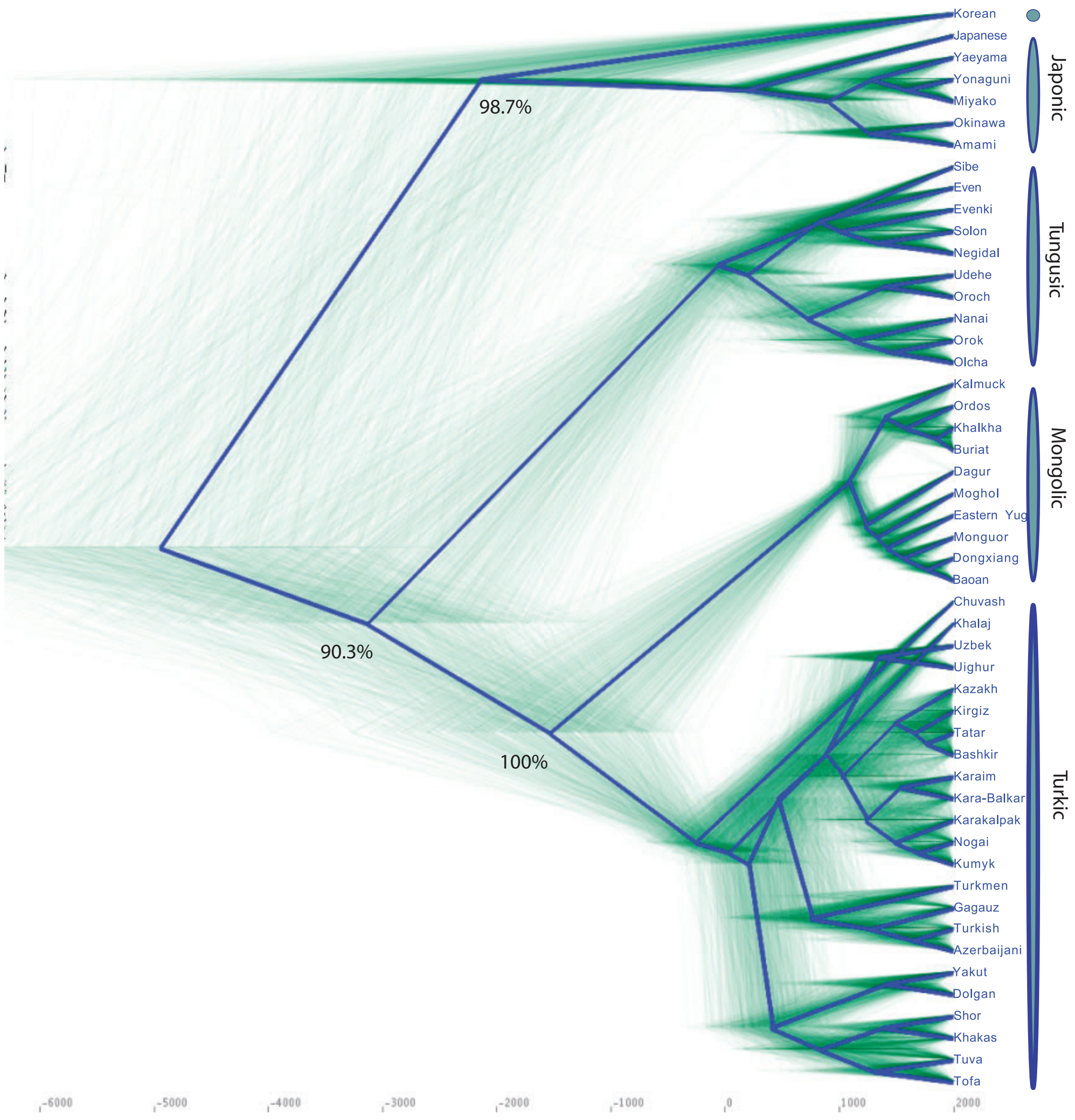

Figure 8. A DensiTree representing the posterior distribution of phylogenies. The blue tree is the maximum clade credibility tree; the green trees are individual trees from the posterior sample. Numbers in the tree indicate posterior clade support for the families. Mongolic and Turkic have 100% support, Korean and Japonic 98.7% and Tungusic with Mongolic, and Turkic has 90.3% support.

hypotheses with regard to the internal structure of the Transeurasian family.

Our application of Bayesian phylolinguistics to infer a phylogeny of the Transeurasian languages is unprecedented. Previous classifications were based on either classical historical-comparative linguistics or lexicostatistics. The difference between prior approaches and the current Bayesian approach lies in both the methodology and the dataset. Whereas the lexicostatistic method is a distance-based method, measuring the proportion of cognates between two languages, the classical comparative and the Bayesian methods are character-based methods. The former is a parsimony method, inferring trees on the basis of shared innovations, while the latter generate trees by quantifying how likely it is that the observed data have been produced by a certain model of the evolution of cognates along that tree.

In addition, the datasets used in previous approaches differ from ours. Whereas classical historical linguistic approaches started from shared innovations in phonology, lexicon and morphosyntax, previous lexicostatistic approaches, just like our current Bayesian approach, were based on basic vocabulary. However, our collection of Transeurasian basic vocabulary is innovative because it is based on the recently developed Leipzig-Jakarta list, sifts and revises questionable etymological proposals, answers previous criticisms and expands the dataset with material from understudied languages such as the Ryukyuan languages.

Because of our innovative morphological approach and our fine-tuning of the dataset, we were able to obtain new results that allow us to determine the most competitive hypothesis with regard to the internal structure of the Transeurasian family. The hypothesis with the highest credibility is a binary topology in which Tungusic clusters with Mongolic and Turkic, with Tungusic branching off first (HD in Fig. 6). This solves a controversy between the Russian school, proposing a binary Mongolo-Turkic unity (Fig. 2) and the western school, supporting a Tunguso-Mongolic unity (Fig. 3), in strong favor of the Russian school as there is no support (0%) for the proposed Tunguso-Mongolic unity. Moreover, it solves a second dispute about the placement of Tungusic either in a cluster with Japonic and Koreanic, or in a cluster with Turkic and Mongolic, in favor of the second view with a credibility of 90.3%. Overall, our results move the field forward in that they provide a quantitative basis to test various competing hypotheses with regard to the internal structure of the Transeurasian family.

9. Conclusion

For the first time in the history of linguistics, we integrated classical historical-comparative linguistics and computational Bayesian phylolinguistics to infer a phylogeny of the Transeurasian languages. For this purpose, we introduced a new dataset, a new coding principle and applied a newly introduced evolutionary model for capturing the historical behavior of the cognate sets. Our dataset consisted of 150 comparative sets of Japanic, Koreanic, Tungusic, Mongolic, and Turkic cognates, which yield proto-Transeurasian reconstructions for 107 Leipzig-Jakarta basic vocabulary items. We applied an unconventional coding principle, permitting a certain degree of semantic freedom in the development of the cognates. In comparison with the conservative approach, which is restricted to semantic equivalents only, our new approach enabled us to capture a stronger historical signal at a more remote time depth. We further applied our recently introduced pseudo Dollo covarion model with relaxed clock as a new model for capturing evolution in our cognate sets. We showed that it had a better fit than the currently existing models for our dataset.

Applying the alternative coding and improved modeling to our renewed dataset in a Bayesian setting, we provided a quantitative basis to test various competing hypotheses with regard to the internal structure of the Transeurasian phylum. We found that a hypothesized binary topology for the Transeurasian family, whereby the Tungusic subgroups clusters with the Altaic unity and branches off first, represented the best supported Transeurasian phylogeny.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

Acknowledgements

The authors would like to thank Wayne Lawrence, Sander Savelyev, and Sonya Oskolskaya for their contributions to the dataset. The research leading to these results has received funding from the European Research Council under the Horizon 2020 Program/ERC Grant Agreement no. 646612 granted to M. R. This research was further funded by Marsden grant (UOA1308), granted to R. B. (http://www.royalsociety.org.nz/programmes/funds/marsden/awards/2013-awards/).

References

Baskakov, N. A. (1981) *Altaiskaja Sem'ja Jazykov i Ee Izučenie*. Moscow: Nauk.

Blažek, V. and Schwarz, M. (2009) 'Koguryo and Altaic. On the Role of Koguryo and Other Old Korean Idioms in the [Sic.] Altaic Etymology', *Journal of Philology: Ural-Altaic Studies*, 1/1: 13–25.

—— and —— (2014) 'Jmennádeklinace v Altajských Jazycích', *Linguistica Brunensia*, 62/1: 89–98.

Blust, R. (2013). *The Austronesian Languages*. Canberra: Asia-Pacific Linguistics.

Bouckaert, R. et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337/6097: 957–60.

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

—— and Robbeets, M. (2017) 'Pseudo Dollo Models for the Evolution of Binary Characters along a Tree', *BioRxiv*.

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

Büyükmavi, M., (2007) 'Rezension Von Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?', *Oriens Extremus*, 46: 306–10.

Clauson, G. (1956) 'The Case against the Altaic Theory', *Central Asiatic Journal*, 2: 181–7.

de Boer, E. (in press, 2019) 'The Classification of the Japonic Languages'. In: Robbeets, M., Neshcheret, N. and Savelyev, A. (eds) *The Oxford Guide to the Transeurasian Languages*. Oxford: Oxford University Press.

Décsy, G. (2007) 'Review of Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?', *Eurasian Studies Yearbook*, 79: 157.

Doerfer, G. (1963–1975) *Türkische und Mongolische Elemente im Neupersischen, unter besonderer Berücksichtigung älterer neupersischer Geschichtsquellen, vor allem der Mongolen- und Timuridenzeit*, vols 1–4. Wiesbaden: Franz Steiner.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4/5: e88.

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

Dybo, A. (2007) *Lingvističeskie kontakty rannih tjurkov. Leksičeskij fond. Pratjurkskij period [Linguistic contacts of the early Turks. Lexical stock. Proto-Turkic period]*. Moskva: Vostočnaja Literatura.

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

Felsenstein, J. (2004) *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Georg, S. (2007) 'Review of Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?', *Korean Studies*, 12: 247–78.

Gernhard, T. (2008) 'The Conditioned Reconstructed Process', *Journal of Theoretical Biology*, 253/4: 769–78.

Gözaydin, N. (2006) 'Dergi ve kitap dünyasïndan', *Türk Dili*, 653: 467–71.

Gray, R. D., Drummond, A. J. and Greenhill, S. J. (2009) 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323/5913: 479–83.

Hammarström, H., Forkel, R. and Haspelmath, M. (2017) 'Glottolog 3.0. Jena: Max Planck Institute for the Science of Human History' <http://glottolog.org> accessed 9 Sep 2017.

Hattori, S. (1976) Ryūkyū Hōgen to Hondo Hōgen [the Dialects of the Japanese Main Islands and the Ryukyuan Dialects]. *Ifa Fuyū Seitan Hyakunen Kinenkai Okinawagaku No Reimei [the Dawn of Okinawan Studies]*, pp. 7–55. Tokyo: Okinawa Bunka Kyokai.

Heine, B. and Kuteva, T. (2002) *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hillis, D. M. et al. (2003) 'Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? ', *Systematic Biology*, 52/1: 124.

Holman, E. et al. (2011) 'Automated Dating of the World's Language Families Based on Lexical Similarity', *Current Anthropology*, 52/6: 841–75.

Hudson, M. (2002) 'Agriculture and Language Change in the Japanese Islands'. In: Bellwood, P. and Renfrew C. (eds) *Examining the Farming/Language Dispersal Hypothesis*, pp. 311–8. Cambridge: McDonald Institute for Archaeological Research.

Janhunen, J. (2009) 'Reconstructing the Language Map of Prehistorical Northeast Asia'. In: Karttunen, K. (ed.) *Anantam Śāstram: Indological and Linguistic Studies in Honour of Bertil Tikkanen*, pp. 283–305. Helsinki: Suomalais Ugrilainen Seura.

—— (2012) 'The Expansion of Tungusic as an Ethnic and Linguistic Process'. In: Malchukov, A. and Whaley, L. (eds) *Recent Advances in Tungusic Linguistics*, pp. 5–16. Wiesbaden: Harrassowitz.

Jeong, C., and Wang, C. (in press, 2019) 'Transeurasian unity from a population genetic perspective', in Robbeets, M., Neshcheret, N. and Savelyev, A. (eds) *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

Johanson, L., and Robbeets, M. (2010) 'Introduction'. In: Johanson, L. and Robbeets, M. (eds) *Transeurasian Verbal Morphology in a Comparative Perspective: Genealogy, Contact, Chance*, pp. 1–5 (Turcologica 78). Wiesbaden: Harrassowitz.

Kara, G. (2007) 'Review of Robbeets, Martine 2005. Is Japanese related to Korean, Tungusic, Mongolic and Turkic?', *Anthropological Linguistics*, 49: 95–98.

Kass, R. E., and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association*, 90/430: 773–95.

Knüppel, M. (2006) 'Ein Beitrag zur Japanisch-Koreanisch-Altaischen Hypothese', *Wiener Zeitschrift für die Kunde des Morgenlandes*, 98: 353–64.

Korovina, E. (2011) *Leksika prirodnogo i kul'turnogo okrieniya v Tunguso-Man'ččurskich jazykach (v istoriko-tipologieskom osveščenii). [Tungus-Manchu vocabulary related to natural environment and cultural activities (in historical and typological perspective)]*, BA dissertation. Moscow: Russian State University of Humanities.

Lee, S., and Hasegawa, T. (2011) 'Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic Languages', *Proceedings of the Royal Society B*, 278: 3662–9.

List, M. et al. (2014) *Database of Cross-Linguistic Colexifications*. <http://clics.lingpy.org/main.php> accessed 24 Oct 2017.

Martin, S. (1966) 'Lexical Evidence Relating Korean to Japanese', *Language*, 42: 185–251.

Matras, Y. (2009) *Language Contact*. Cambridge: Cambridge University Press.

Maturana, P. et al. (2017) 'Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling', *arXiv Preprint arXiv: 1703.05471*.

Miller, R. A. (1971) *Japanese and the Other Altaic Languages*. Chicago: The University of Chicago Press.

—— (2007) 'Review of Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?', *Ural-Altaische Jahrbücher Neue Folge*, 21: 274–9.

Miyamoto, K. (2016) 'Archaeological Explanation for the Diffusion Theory of the Japonic and Koreanic Languages', *Japanese Journal of Archaeology*, 4: 53–75.

Mudrak, O. A. (2009) *Klassifikacija tjurkskih jazykov i dialektov s pomoščju metodov glottohronologii 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]*. Moscow: RGGU.

Nicholls, G. K. and Gray, R. D. (2008) 'Dated Ancestral Trees from Binary Trait Data and Their Application to the Diversification of Languages', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70: 545–66.

Oskolskaya, S., Koile, E., and Robbeets, M. (unpublished manuscript) 'A Phylogenetic Approach to the Classification of the Tungusic Languages'.

Pevnov, A. M. (2012) 'The Problem of the Localization of the Manchu-Tungusic Homeland'. In: Malchukov, A. and Whaley, L. (eds) *Recent Advances in Tungusic Linguistics*, pp. 17–40. Wiesbaden: Harrassowitz.

Poppe, N. (1960) *Vergleichende Grammatik Der Altaischen Sprachen. Teil 1, Vergleichende Lautlehre. (Porta Linguarum Orientalium, Neue Serie, 4*). Wiesbaden: Otto Harrassowitz.

—— (1965) *Introduction to Altaic Linguistics*. Wiesbaden: Otto Harrassowitz.

—— (1975) 'Altaic Linguistics: An Overview', *Gengo No Kagaku [Sciences of Language*, 6: 130–86.

Ramstedt, G. J. (1952–1957) *Einführung in Die Altaische Sprachwissenschaft I: Lautlehre; II: Formenlehre*. Helsinki: Suomalai-Ugrilainen Seura.

Robbeets, M. (2005) *Is Japanese Related to Korean, Tungusic, Mongolic and Turkic? (Turcologica 64*). Wiesbaden: Harrassowitz.

—— (2015) *Diachrony of Verb Morphology: Japanese and the Transeurasian Languages. (Trends in Linguistics Studies and Monographs 291*). Berlin: Mouton-De Gruyter.

—— (ed.) (2017) *Transeurasian Linguistics*. Vol. 1. The History of the Debate. London: Routledge.

—— (in press, 2019) 'Basic Vocabulary'. In: Robbeets, M.; Neshcheret, N. and Savelyev, A. (eds) *The Oxford Guide to the Transeurasian Languages*. Oxford: Oxford University Press.

Rodriguez, F. J. et al. (1990) 'The General Stochastic Model of Nucleotide Substitution', *Journal of Theoretical Biology*, 142/4: 485–501.

Rozycki, W. (2006) 'Review of Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?', *Mongolian Studies*, 28: 114–5.

Rybatzki, V. (2003) 'Intra-Mongolic Taxonomy'. In: Janhunen, J. (ed.) *The Mongolic Languages*, pp. 364–90. London: Routledge.

Savelyev, A. (in press, 2019) 'A Bayesian approach to the classification of the Turkic languages'. in: Robbeets, M., Neshcheret, N. and Savelyev, A. (eds) *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

Siska, V. et al. (2017) 'Genome-Wide Data from Two Early Neolithic East Asian Individuals Dating to 7700 Years Ago', *Science Advances*, 3: e1601877.

Stadler, T. et al. (2013) 'Birth–Death Skyline Plot Reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (HCV)', *Proceedings of the National Academy of Sciences*, 110/1: 228–33.

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

Street, J. (1962) 'Review of Nikolaus Poppe, Vergleichende Grammatik Der Altaischen Sprachen, Teil I', *Language*, 38: 92–98.

Tadmor, U., Haspelmath, M. and Taylor, B. (2010) 'Borrowability and the Notion of Basic Vocabulary', *Diachronica*, 27/2: 226–46.

Taylor, W. (2016) 'Horse Demography and Use in Bronze Age Mongolia', *Quaternary International*, 30: 1–13.

—— et al. (2017) 'A Bayesian Chronology for Early Domestic Horse Use in the Eastern Steppe', *Journal of Archaeological Science*, 81: 49–58.

Tekin, T. (1994) 'Altaic Languages'. In: Asher, R. E. (ed.) *The Encyclopedia of Language and Linguistics*, Vol. 1, pp. 82–85. Oxford: Pergamon Press.

Tenishev, E. R. et al. (2006) *Sravnitel'no-Istoričeskaja Grammatika Tjurkskih Jazykov. Pratjurkskij Jazyk-Osnova. Kartina Mira Pratjurkskogo Etnosa Po Dannym Jazyka* [*A Historical-Comparative Grammar of the Turkic Languages. The Proto-Turkic Language. The Worldview of the Proto-Turks Based on Linguistic Evidence]*. Moscow: Nauka.

Tuffley, C. and Steel, M. (1998) 'Modeling the Covarion Hypothesis of Nucleotide Substitution', *Mathematical Biosciences*, 147: 63–91.

Unger, J. M. (1990) 'Summary Report of the Altaic Panel'. In: Baldi, P. (ed.) *Linguistic Change and Reconstruction Methodology (Trends in Linguistics. Studies and Monographs 45)*, pp. 479–82. Berlin: Mouton de Gruyter.

—— (2009) *The Role of Contact in the Origins of the Japanese and Korean Languages*. Honolulu: University of Hawaii Press.

Vladimircov, B. J. (1929) *Sravitel'naja Grammatika Mongol'skogo Pis'mennogo Jazyka i Xalxaskogo Narečija. Vvedenie i Fonetika*. Leningrad: Nauka.

Vovin, A. (2005) 'The End of the Altaic Controversy', *Central Asiatic Journal*, 49: 71–132.

—— (2009) 'Review of Robbeets, Martine 2005. Is Japanese Related to Korean, Tungusic', *Mongolic and Turkic? Central Asiatic Journal*, 53: 105–47.

—— et al. (2010) *Koreo-Japonica: A re-evaluation of a common genetic origin* (Center for Korean Studies Monograph). Honolulu, University of Hawai'i Press.

Wang, C. C. et al. (2016) *Reconstructing Population History in East Asia. 7th International Symposium on Biomolecular Archaeology (ISBA7)*. The Oxford University Museum of Natural History, Oxford.

Whitman, J. B. (2011) 'Northeast Asian Linguistic Ecology and the Advent of Rice Agriculture in Korea and Japan', *Rice*, 4: 149–58.

Wilkins, D. P. (1996) 'Natural Tendencies of Semantic Change and the Search for Cognates'. In: Durie, M. and Ross, M. (eds) *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, 264–306. Oxford: Oxford University Press.

Witsen, N. (1692) *Noord En Oost Tartarye. Bondig Ontwerp Van Eenige Dier Landen En Volken, Welke Voormaels Bekent Zijn Geweest*. Amsterdam: François Halma.

Wohlgemuth, J. (2009) *A Typology of Verbal Borrowings*. Berlin: Mouton de Gruyter.

Yang, Z. (1996) 'Among-Site Rate Variation and Its Impact on Phylogenetic Analyses', *Trends in Ecology & Evolution*, 11/9: 367–72.