Journal of Language Evolution, 2023, **XX**, 1–18
https://doi.org/10.1093/jole/lzad002
Advance access publication 29 March 2023
Research article

OXFORD

Modelling admixture across language levels to evaluate deep history claims

Nataliia Hübler[1,*] and Simon J. Greenhill[1,2]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[*]Corresponding author: nataliia_huebler@eva.mpg.de

The so-called 'Altaic' languages have been subject of debate for over 200 years. An array of different data sets have been used to investigate the genealogical relationships between them, but the controversy persists. The new data with a high potential for such cases in historical linguistics are structural features, which are sometimes declared to be prone to borrowing and discarded from the very beginning and at other times considered to have an especially precise historical signal reaching further back in time than other types of linguistic data. We investigate the performance of typological features across different domains of language by using an admixture model from genetics. As implemented in the software STRUCTURE, this model allows us to account for both a genealogical and an areal signal in the data. Our analysis shows that morphological features have the strongest genealogical signal and syntactic features diffuse most easily. When using only morphological structural data, the model is able to correctly identify three language families: Turkic, Mongolic, and Tungusic, whereas Japonic and Koreanic languages are assigned the same ancestry.

Keywords: language evolution; typology; linguistics; admixture model.

1. Introduction

To establish language relationships, the 'gold-standard approach' in linguistics applies the comparative method (Durie and Ross 1996) to lexical data to identify homologous traits that diagnose language subgroupings, e.g. phonological innovations and form-meaning pairs of morphemes. Linguists have carefully applied this approach to the world's languages and identified more than 290 primary language families (Greenhill 2015; Hammarström et al. 2020). However, research that aims to identify relationships above the family level—that is, macrofamilies—is often highly controversial (e.g. see Pagel et al. 2013 vs. Mahowald and Gibson 2013; Heggarty 2013). First, the rate of language change is so rapid that any deep signal (such as that needed to prove a macrofamily connection) is likely to be lost after 6,000–10,000 years, and it becomes impossible to disentangle true historical relatedness from borrowing between languages and chance similarity (Ringe 1995, 1999; Nichols 1992). Second, proponents of deeper relationships have often not applied the most rigorous standards (or have been unable to because of the loss of signal) and have been accused of biased selection of data if not outright cherry-picking (Matisoff 1990; Tian et al. 2022).

A third issue comes from the combinatoric explosion in number of comparisons with big language families: how do researchers determine just which families are to be compared first (Ross 1996)? For example, if we wished to identify the relationships between five different language families, there are 105 possible ways of connecting these trees (Felsenstein 1978). In a proposed family like Trans New Guinea, which may have around 40 sub-families (Pawley 2012), there are therefore 10^7 possibilities, which makes evaluating all permutations impossible for even the most dedicated linguist.

The most common approach to represent genealogical relationships between species or languages is a tree of descent. The tree model was popularised in linguistics by Schleicher in 1853 (Schleicher 1853; List et al. 2016; Jacques and List 2019). Recently, it has experienced a new increase in popularity in historical linguistics, especially in combination with Bayesian statistics (Gray et al. 2009; Grollemund et al. 2015; Kolipakam et al. 2018; Koile et al. 2022). However, we can only interpret a tree in an appropriate way if

the relatedness of languages in question has been well established. This criterion means that, while one could build a tree between macrofamilies (e.g. Pagel et al. 2013; Robbeets et al. 2021), it is unclear whether the deeper branches of the tree between families represent the historical *phylogenetic* relationships, or the *borrowing* relationships between the languages (Reesink et al. 2009), or even just chance similarity (Greenhill et al. 2017).

A case in point concerns the deeper relationships between the Turkic, Tungusic, Mongolic, Japonic, and Koreanic language families. One proposal, *Altaic*, links Tungusic, Mongolic, and Turkic languages (Poppe 1960, 1965, 1975). Another proposal, *Transeurasian*, connects these three families to Japonic and Koreanic (Ramstedt 1924; Miller 1971; Starostin et al. 2003; Johanson and Robbeets 2010; Robbeets 2020a). While there are obvious structural/typological similarities between these languages, there is no consensus on the source of these similarities: some linguists attribute them to borrowing between languages (Vovin 2005; Georg 2007; Vovin 2010; Vajda 2020), while others argue for phylogenetic inheritance from a common ancestor (Starostin et al. 2003; Robbeets 2020b). A recent high-profile paper (Robbeets et al. 2021) has proposed lexical cognates for these languages and has reconstructed the putative history of Transeurasian; however, this work has been criticised on a number of grounds with a major point of contention being that the cognates are erroneous (Tian et al. 2022).

In an attempt to provide a principled way forward to these issues, we consider an alternative model that can account for both inheritance and borrowing: the Bayesian clustering algorithm STRUCTURE (Pritchard et al. 2000). STRUCTURE uses an iterative Bayesian approach to model the distribution of samples amongst populations by clustering these samples based on their shared patterns of variation (Porras-Hurtado et al. 2013). As with any clustering algorithm belonging to unsupervised machine learning, STRUCTURE tries to find homogeneous groups within the data. STRUCTURE probabilistically assigns each sample—languages in our case—to these groups, or 'populations'. For example, one language might be assigned with a proportion of 90% in group one, 9% in group two, and 1% in group three. Therefore, each language can comprise a range of group memberships allowing us to quantify the relative ancestry components from each population. As STRUCTURE does not distinguish between vertical inheritance and borrowing between languages, we use the term 'ancestry component' as an agnostic term meaning either or both of these alternatives. In addition to the proportion of each ancestry for each language, the output of the STRUCTURE algorithm also provides the frequency of each feature in each of the ancestry clusters, so we can evaluate, which features are linked to which groups.

In linguistics, Reesink et al. (2009) pioneered the application of STRUCTURE to language data and used it to investigate the relationships between languages of Australia, New Guinea, and surrounding islands. Some of the identified ancestries align well with expected phylogenetic groupings, e.g. the Oceanic (Austronesian), Trans New Guinea, and Australian languages. Other groupings, however, had not been proposed before suggesting convergence between Austronesian and some Papuan non-Trans-New-Guinea speaking groups. In their study, STRUCTURE was able to correctly determine the genealogical relationships between languages despite their geographical separation on many occasions. Since STRUCTURE was able to recognise known language families, Reesink et al. (2009) proposed that the other clusters suggested by STRUCTURE might represent undiscovered genealogical groupings. They warn that the order in which populations are detected, should not be associated with chronology, i.e. the increase in K values and the emerging groupings cannot be interpreted as consequent splits on a timeline, as would be the case with a tree.

Several studies applying STRUCTURE to linguistic data followed the work by Reesink et al. (2009). Bowern (2012) applied STRUCTURE (among other methods) to vocabulary data of Tasmanian languages to estimate the degree of source mixture within them and used the results to reject the previously suggested relatedness of some language groups and rather attribute the similarities to mixing. Syrjänen et al. (2016) tested the performance of STRUCTURE for studying intralingual variation on the example of Finnish dialects. The division of dialects into groups achieved by STRUCTURE corresponds to the traditional views. Norvik et al. (2022) applied Fast-STRUCTURE to Uralic languages spoken predominantly in Northern Europe and Northwestern Asia. STRUCTURE correctly identified Uralic subgroups as well as distinct areas of historical interaction between Uralic languages, demonstrating that typological data can be used for diachronic studies.

In this study, we apply STRUCTURE to a large data set of grammatical structures (Hübler 2021; Hübler 2022) for the five language families that arguably comprise Altaic and Transeurasian: Japonic, Koreanic, Mongolic, Tungusic, and Turkic. Our aim is to evaluate the potential for this approach to provide a way forward for evaluating macro-family proposals in general. As it has been shown that structural features differ in terms of phylogenetic signal they contain and the rate, at which they evolve, we expect some structural features to perform better than others in attributing languages

to language families. To specifically find where these differences are, we split our data set into three samples, covering phonology, morphology, and syntax. Can we identify, first, the accepted language family groupings and, second, any deeper links between these groups? How do these potential groupings play out across phonology, morphology, and syntax? And, finally, can we identify languages that have potentially high amounts of admixture in their histories?

2. Data and Methods

The language sample covers a vast area in Eurasia and includes 60 languages from 5 language families: Turkic, Mongolic, Tungusic, Koreanic, and Japonic (Hübler 2021, 2022). The sample was based on languages with good grammatical descriptions and samples these families reasonably with 21/44 Turkic, 14/17 Mongolic, 11/13 Tungusic, 2/2 Koreanic, and 12/15 Japonic languages represented in Glottolog (Hammarström et al. 2020). These languages were coded for 224 features. We based 189 features on the coding in the Grambank database (Skigård et al. in press), including 6 binarised versions of Grambank features on word order (from 'What is the order of X and Y?' to 'Can X precede Y?' and 'Can Y precede X?'). We extended this with 35 features on phonology and other grammatical markers (8 of these features are proposed by Robbeets 2017). Each feature was coded in binary manner such that '1' encoded trait presence, '0' encoded the absence of this trait in the particular language, and '?' meant that there was not enough information in the grammar or the information was ambiguous for the particular trait. Out of the 224 features, 53 features had identical values for all languages in the sample and were removed from further analysis, leaving a sample of 171 features. More than half of the languages could be coded for 95% of features (162 features), and around two third of the languages could be coded for more than 78% of features (134 features).

There are suggestions that structural borrowing happens at different frequencies across linguistic domains in a hierarchical manner. According to Thomason and Kaufman (1988: 38), phonology is borrowed first, and, as the intensity of contact increases, syntax, and morphology follow. We therefore separated our data into these three broad categories to see if this helps tease apart the different ancestries across different linguistic levels. We categorise the features based on the language level they target: phonological shape, word, and clause. The first category is 'phonological' (14 features) and comprises traits tracking aspects of vowel harmony (4 features), phonotactic constraints (3 features), voicing/aspiration distinctions in consonants (4 features), and distinctions between l/r.

The second category is 'morphological' (71 features), which targets words and aspects of morphology encoded by a bound marker. The most prominent functional categories belonging to this domain include morphological tense-aspect-mood-evidentiality marking (12 features), quantification (11 features), deixis (9 features), valency marking (9 features), flagging and indexing (10 features), derivation (5 features), possession (5 features). We defined 'morphological' here as having the word as the scope, so this category also includes features like numeral classifiers and ideophones, which are not directly morphological, but we did not want to add more categories with small numbers of features and 'morphology' was the closest category. We note that often the morphological features that usefully define language groups are cognate features derived from morphological paradigms. We do not include these types of features here as they are often predicated on a particular subgrouping hypothesis and we did not want to prejudge our results and build in support for the hypotheses we are testing.

The third category is 'syntactic' (82 features). 'Syntactic' features comprise features that have the whole clause or the nominal phrase as their scope. Most features here concern phonologically free marking. In some features there is variation, e.g. the feature on negation marking appearing clause-finally vs. clause-initially: in many languages in the sample negation is marked by a suffix on the verb, and, due to SOV word order, it appears to be clause-final, although the negation marker is bound—the focus of the feature is on the position and not on the phonological boundness. Some of the functional categories included are word order (13), TAME+ (9 features), interrogation (8 features), negation (6 features), and possession (6 features).

If the scope of the feature covers both syntax and morphology according to its definition, but in the languages in question there is only phonologically bound marking, i.e. is relevant for the feature, then the feature is assigned to the category 'morphological'. For example, the feature GB105 'Can the recipient in a ditransitive construction be marked like the monotransitive patient?' asks both about marking with an adposition and an affix (as well as indexing on the verb, if no flagging is available), but in the languages in question recipients and monotransitive patients are marked by suffixes, therefore this feature is assigned to the 'morphological' category.

To infer the underlying population structure that describes our data we applied the STRUCTURE algorithm (Pritchard et al. 2000). While originally developed for genetic data, it has been successfully applied to linguistic data (Reesink et al. 2009; Bowern 2012; Syrjänen et al. 2016; Norvik et al. 2022).

To make the method more applicable to language data, we did not use the linkage model and set ploidy to 1. We ran the algorithm multiple times: each time with a different number of assumed populations K from 2 to 10 and repeating the process 50 times for each K. We set the starting value of α, the Dirichlet parameter for degree of admixture, to 1 and allowed it to be inferred. Allele/trait frequencies were allowed to correlate among populations.

The STRUCTURE output provides several estimates, which can be used to select the optimal number of clusters. First, there is mean log likelihood (Fig. 1, first column). Second, there is a posterior probability of data for a given K (A metric in Fig. 1). Further, there are three more metrics described in Evanno et al. (2005), calculated as follows:

- rate of change of the likelihood function with respect to K (B metric in Fig. 1), $L'(K) = L(K) - L(K-1)$,
- difference between successive values of $L'(K)$ (C metric in Fig. 1), $|L''(K)| = |L'(K+1) - L'(K)|$ and
- ΔK, the modal value of the distribution of which indicates true K (D metric in Fig. 1), $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|)/s[L(K)]$

Pritchard et al. (2010: 15–17) recommend an *ad hoc* procedure to choose the best K, namely to inspect the distribution of $L(K)$ across runs and K's. There are three basic components of this procedure: 1. a jump in probability before the optimal K value, 2. high variation between runs after the optimal K value, 3. 'plateauing' of probability starting from the optimal K value. First, the quality of the model improves rapidly as the number of clusters increases, but then the improvement slows down and an increase in K does not lead to a significant increase in probability. The point after which the significant increase in probability stops should be taken as the true K value, i.e. the number of clusters inherent to the data (the so-called 'plateauing').

We split the data (Hübler 2021) into three sets, according to the language level assigned to the feature, and ran STRUCTURE 50 times on each of the data sets based on language levels for K from 2 to 10. Out of the 50 runs for each language level, we selected the

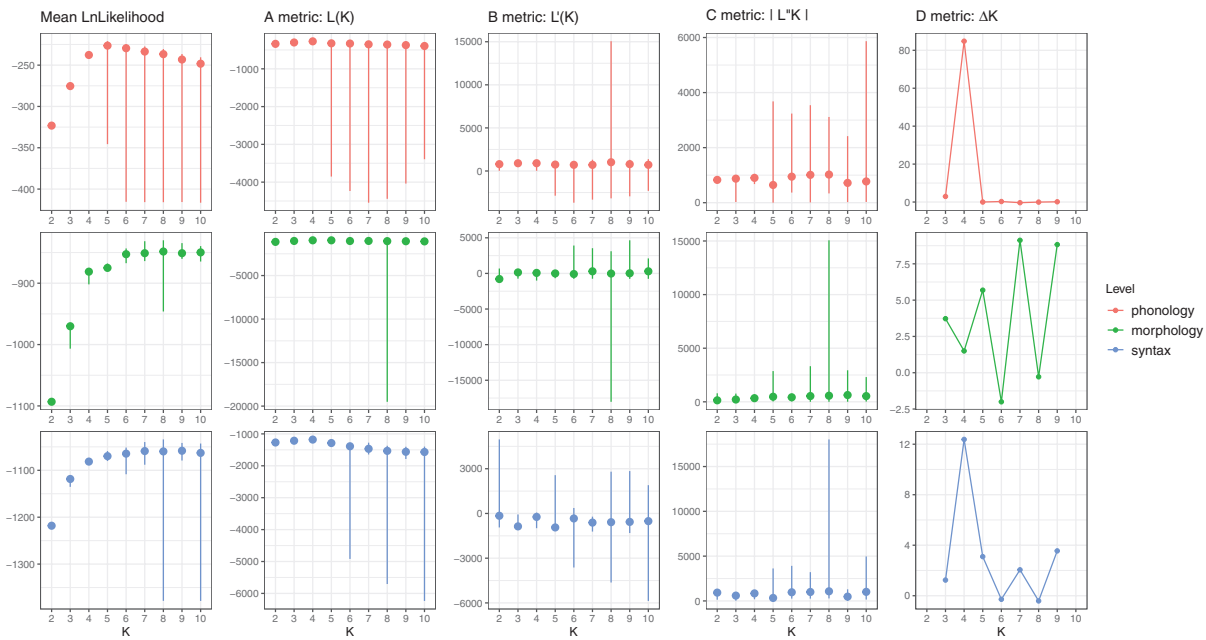

Figure 1 Variation in the log likelihood of K from 2 to 10 across 50 runs and three language levels. The bars indicate the whole range of values (from minimum to maximum value) and the points indicate the median value. Each row represents a language level: phonology, morphology and syntax. Each column represents a different metric, which indicates the most probable number of assumed populations (K) (Evanno et al. 2005). The first two metrics are the mean log likelihood ('LnLikelihood') and the posterior probability of data for a given K, $L(K)$. The third metric is the rate of change of the likelihood function with respect to K. The fourth metric is the difference between successive values of $L(K)$. The fifth metric is ΔK, calculated according to the formula $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|)/s[L(K)]$. The first and the last metrics provide most informative results and indicate that $K = 4$ is, on average, the most plausible number of clusters in the data (plateauing after $K = 4$ in mean log likelihood and the highest modal value at $K = 4$ in ΔK).

admixture proportions from the run with the highest log probability of data for further analysis and visualisation of the results.

3. Results

Following the *ad-hoc* procedure described in Pritchard et al. (2010: 15–17), we take into account plateauing to choose the best K value, which is the one directly at the beginning of the plateauing. We can see it distinctly in Fig. 1 for mean log-likelihood: the mean log likelihood continues to increase substantially until $K = 4$ for phonology and for morphology but starts plateauing at $K = 5$. It is less obvious in syntax, where there is a substantial jump in likelihood from $K = 2$ to $K = 3$, but a smaller one from $K = 3$ to $K = 4$, after which we definitely observe plateauing. Another indication of a true K value is an increase in variation between runs: we observe it starting with $K = 5$ for phonology, but less so for morphology and syntax. In the case of phonology, we see an increase in log-likelihood up to $K = 5$ (argument in favour of $K = 5$) and an increase in variation starting from $K = 5$ (argument against $K = 5$ and in favour of $K = 4$). In case of syntax, we see a high jump in log-likelihood from $K = 2$ to $K = 3$, but the log-likelihood keeps growing after $K = 3$, until it reaches $K = 7$, the variation is higher starting from $K = 6$.

Following the ΔK method (Evanno et al. 2005) to determine the true number of populations in the data (Fig. 1, D metric), $K = 4$ is the best assumed number of populations for phonology and syntax, but it does not have a clear modal value for morphology and shows two peaks: at $K = 7$ and $K = 9$. For the sake of comparability of the results and following the interpretation of the distribution of the mean log-likelihood in Fig. 1, we chose $K = 4$ for morphology as well. For admixture profiles at other assumed K's, see Supplementary Figs. S1–S3. For the admixture profile based on the whole data set, without a split based on language level, see Supplementary Fig. S4.

As we have strong prior expectations that the language clusters will mostly correspond to (larger) language families, we can label each recovered group with the language family that its members are derived from. For example, in Fig. 2, the orange ancestry component is primarily linked to the Turkic languages, violet to Mongolic, green to Tungusic, and pink to the Japonic languages. The Koreanic languages share their ancestry either with Mongolic or with Japonic languages, depending on the linguistic level.

To summarise the inferred admixture proportions, we calculated the mean level of admixture for each language family. We summed all admixture proportions, which do not belong to the population with the highest proportion in most languages in that particular family (see Table 1). We see the lowest admixture at the level of morphology (on average, 6.6%, SD = 3%), followed by phonology (19.6%, SD = 16%) and syntax (29.8%, SD = 11%). Among the language families, Japonic languages have the lowest average level of admixture (13.3%, SD = 14%), followed by Koreanic (13.7%, SD = 1.7%), Turkic (14%, SD = 8%), Tungusic (22.3%, SD = 14%), and Mongolic (30%, SD = 20.3%) languages (see Supplementary Tables S1–S3).

The Turkic languages stand out as a cluster with the same dominant ancestry, apart from several exceptions, on all language levels. All Turkic languages, except for Chuvash (30% of 'Mongolic' and 18% of 'Tungusic' ancestry), show the lowest levels of admixture at the morphological level. At the phonological level, several Turkic languages show the highest proportions of 'Mongolo-Koreanic' ancestry among all Turkic languages (in descending order: Chagatai 51%, Northern Uzbek 43%, Chuvash 29%, Tuvan 27%, etc.). At the syntactic level, Northern Siberian languages, Dolgan and Yakut, and a South Siberian language, Tuvan, are the languages with the highest admixture levels (more than 65%). In particular, Dolgan and Yakut have a high proportion of 'Mongolic' (47% and 49%, respectively) and 'Tungusic' (12% and 24% respectively) ancestries, Tuvan has a high proportion of 'Mongolic' (29%), 'Tungusic' (16%), and 'Japono-Koreanic' (28%) ancestries.

The Mongolic languages have an internal split at the phonological level: the first group, comprising Eastern Mongolic languages (apart from Khalkha and, to a lesser extent, Ordos), Moghol and Middle Mongol, shares its ancestry with Turkic languages and the second group, comprising Southern Periphery[1] Mongolic languages, Khalkha, and, to a lesser extent, Ordos and Dagur, shares its ancestry with Koreanic languages. There is no such split at the morphological and syntactic levels: Mongolic languages stand out as a rather homogeneous group at the morphological level, apart from Buriat (52% 'Mongolic' and 46% 'Turkic'), and show a high level of admixture at the syntactic level ('Mongolic' ancestry in Ordos comprises 37%, in Mangghuer 34%, in Moghol 24%).

The Tungusic languages stand out as a separate group at the phonological and morphological levels, but not at the syntactic level, where they show a high level of admixture (especially Central-Western Tungusic[2] languages). Manchu shows considerable proportions of 'Turkic' (21%) and 'Mongolo-Koreanic' (33%) ancestries at the phonological and 'Japono-Koreanic' (19%) and 'Mongolic' (34%) at the morphological level. It has the highest 'Tungusic' component at the syntactic level (80%, compared to 43% at the phonological and 46% at the morphological level).

Figure 2 Population structure at $K = 4$. Each row corresponds to a language and each column corresponds to a language level. Languages appear in the order of 1) language families, divided by a black horizontal line: Japonic (from Ura to Eastern Old Japanese), Koreanic (Middle Korean and Korean), Tungusic (from Manchu to Evenki), Mongolic (from Middle Mongol to Mangghuer), Turkic (from Old Turkic to Uighur), 2) branches according to the Glottolog (Hammarström et al. 2020) classification, wherever possible. For each language, the coloring of the bar represents the proportion of each ancestry in the language. The following ancestries roughly correspond to each of the language families: 'pink'—Japonic/Koreanic, 'violet'—Mongolic/Koreanic, 'green'—Tungusic, 'orange'—Turkic. We see the lowest admixture in morphology and the highest in syntax. Japonic languages appear as the most homogeneous group and Mongolic languages as the most heterogeneous group on average. Abbreviations: Tk = Turkic, Tg = Tungusic, M = Mongolic, K = Koreanic, J = Japonic.

The Koreanic languages share the highest proportion of their ancestry with Japonic languages at the morphological and syntactic levels and with Mongolic languages at the phonological level. They are most homogeneous at the phonological and morphological level and are admixed to around 1/3 at the syntactic level (Middle Korean 36%, Korean 32% of 'non-Japono-Koreanic' ancestry).

Table 1 Admixture proportion across language families and language levels.

Language family	Phonology	Morphology	Syntax
Japonic	0.21 (±0.26)	0.04 (±0.05)	0.15 (±0.11)
Koreanic	0.04 (±0)	0.03 (±0)	0.34 (±0.05)
Mongolic	0.47 (±0.39)	0.11 (±0.01)	0.32 (±0.21)
Tungusic	0.13 (±0.09)	0.09 (±0.05)	0.45 (±0.28)
Turkic	0.13 (±0.15)	0.06 (±0.01)	0.23 (±0.08)

Apart from Eastern Old Japanese (61% of 'Tungusic' and 29% of 'Mongolo-Koreanic' ancestry) and Yonaguni (92% of 'Mongolo-Koreanic' ancestry), Japonic languages stand out as a group separate from Koreanic languages at the level of phonology but form a cluster with Koreanic languages at the other two levels.

Do the results from STRUCTURE in terms of the linguistic traits and their patterning into language families also appear plausible in terms of traditional historical linguistic approaches? To evaluate this we can ask if the contribution of each feature into a cluster found by STRUCTURE matches the reconstructability of that feature in the respective proto-language for the cluster. In a recent study (Hübler 2022), phylogenetic methods were used to reconstruct the probability of each of these traits of being 'present' in an ancestral language, e.g. feature X had a high probability of being part of the ancestral proto-language Y with a probability of Z. We took these probabilities and compared them to the contribution of features to different ancestries in the current study. Here we see clear overlaps between features being reconstructed as present or absent and the contribution of features to ancestry clusters (see Fig. 3). Most of the features are concentrated in two corners of the plots (upper right, i.e. 'double' present, and bottom left, i.e. 'double' absent) indicating that either features are both present in the respective ancestry and can be reconstructed as present in the proto-language, or they are absent in the respective ancestry and can be reconstructed as absent in the proto-language. This finding indicates that the ancestry component identified by STRUCTURE for each feature is remarkably consistent with the features found in the reconstructed proto-languages.

4. Discussion

4.1 Linguistic groups

Overall, our results indicate that the best-fitting model to describe these data has four distinct clusters across all three language levels (Fig. 1). The predominant ancestries roughly correspond to the language families the languages belong to and to an extent that we can allow ourselves to name them after the language families: 'Turkic', 'Mongolic', 'Tungusic', 'Koreanic', 'Japonic' (Fig. 2, for the population structure without a split into subsets based on language level, see Supplementary Fig. S4). In some cases, all languages of a particular language family share their dominant ancestry with the languages of another language family, so it is helpful to name that ancestry after both of these families, as in case of Koreanic and Mongolic at the phonological level at $K = 4$, where the best tentative name for this ancestry appears to be 'Mongolo-Koreanic', and in case of Japonic and Koreanic languages at the morphological and syntactic levels at $K = 4$ and the resulting name 'Japono-Koreanic'.

The clusters found at the morphological and syntactic levels are very similar. These two levels strongly distinguish the Tungusic, Turkic, and Mongolic languages and cluster Japonic and Koreanic together. At the morphological level, there is very little admixture between these clusters, while the syntactic level is less distinct with the Tungusic languages in particular showing some similarities to the other families. The phonological level shows broadly similar groupings to the other levels but tends to cluster Mongolic and Koreanic together, leaving Japonic as its own distinct group. The phonology also breaks apart the Mongolic languages, placing some with Koreanic and others with Turkic.

Although we cannot say that the division into more clusters (K's) matches the branches of hypothetical trees of these language families, we do observe some similarities in ancestries between more closely related languages (see Supplementary Figs. S1–S3), e.g. Northern Uzbek and Chagatai starting from $K = 2$ in Phonology, Southern Periphery Mongolic languages (Mangghuer, Mongghul, Shira Yughur, Dongxiang, Bonan) starting from $K = 2$ in Phonology, North Siberian languages (Dolgan and Yakut) starting from $K = 5$ in Morphology and from $K = 3$ in Syntax, Central Western Tungusic languages (Ulch, Orok, Nanai) at $K = 2$ and $K = 4$ in Syntax, Even dialects (Moma Even and Beryozovka Even) at $K = 4$ in Syntax.

Koreanic and Japonic languages appear to be very similar morpho-syntactically and share the same ancestry clusters at the two levels. The origin of these similarities remains a highly debated topic: one hypothesis suggests that Japonic and Koreanic have a common ancestor (Martin 1966; Whitman 2012), another one attributes the similarities to prolonged contact (Vovin 2017). Most scholars seem to nevertheless agree on the origin of Japonic languages on the Korean peninsula, where Koreanic and Japonic languages co-existed for a prolonged time span (Vovin 2017), and on the subsequent spread of the Japonic-speaking

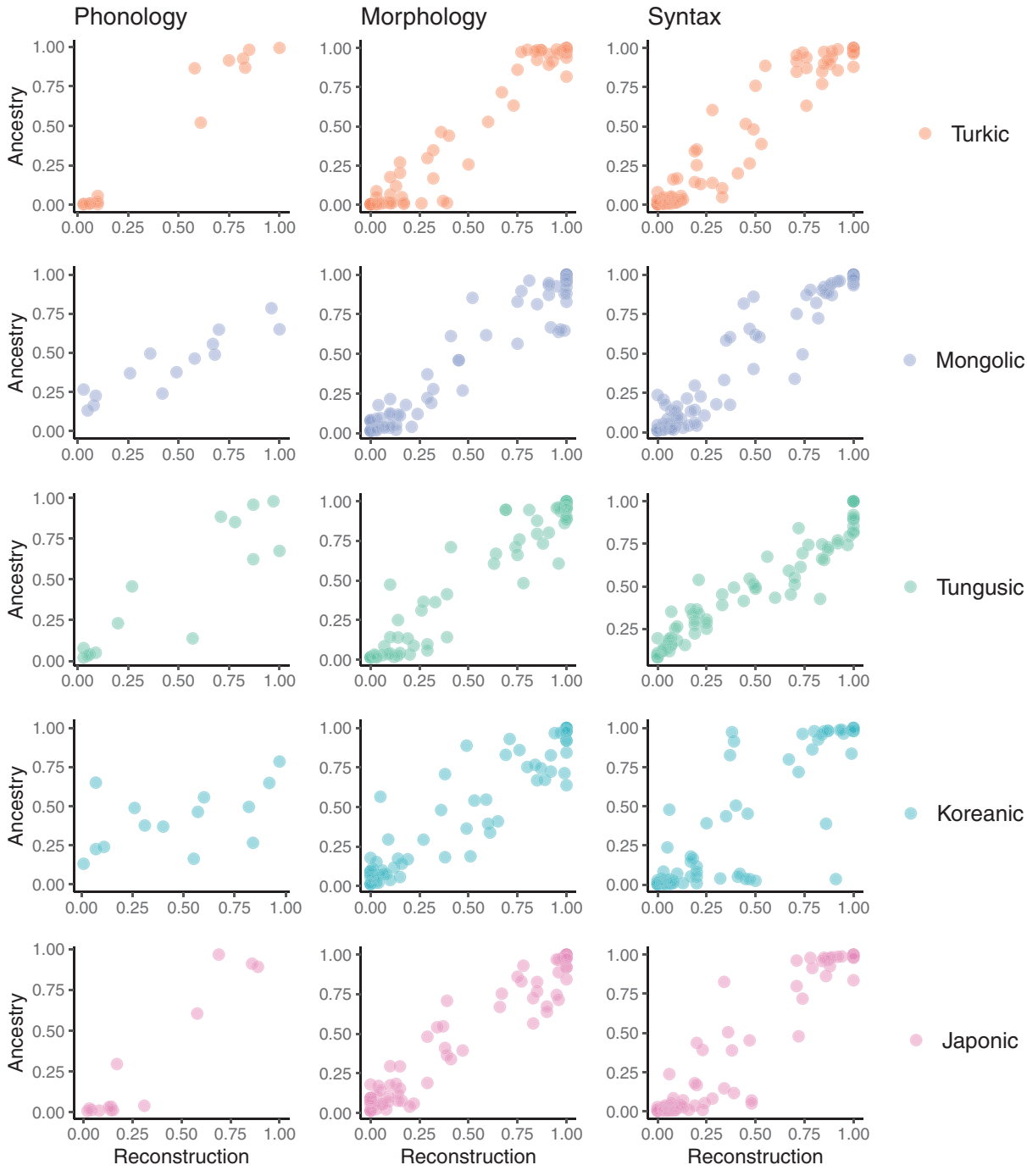

Figure 3 Estimated feature frequencies at $K = 4$ shown as proportions. Abbreviations: Tk = Turkic, Tg = Tungusic, M = Mongolic, K = Koreanic, J = Japonic. The horizontal bar corresponding to each feature consists of feature frequencies (presence) in each of the four assumed ancestries. Each frequency lies within a range between 0 and 1. The range of each bar thus has a cumulative frequency between 0 and 4, i.e. max. 1 for each ancestry.

population (the Yayoi culture) to the Kyūshū island with wet-rice farming (Whitman 2011). Despite the close contact between Koreanic and Japonic languages, there is an only weakly attested transfer of morphemes between Old Korean and Old Japanese (Francis-Ratte and Unger 2020). Since the language groups are so similar morpho-syntactically and this similarity cannot be easily explained by borrowing, a genealogical relationship appears as the most plausible explanation for this similarity.

Our results, while agnostic as to whether this relationship between the two families is due to inheritance or diffusion, are consistent with this debate, and indicate that the STRUCTURE approach does identify these potential deeper groupings with the added benefit of pinpointing, which linguistic traits should be investigated further by traditional approaches. However, simulation studies into the efficiency of the STRUCTURE approach (Hubisz et al. 2009) have suggested that there is a tendency to over-cluster small populations with few members (like Koreanic in our sample). Therefore, the Koreanic and Japonic cluster might be partly due to STRUCTURE's attempt to account for languages by assuming minimal admixture. However, this effect is mitigated as data sets increase in size, and our data set has more loci than the problematic ranges identified in Hubisz et al. (2009), suggesting that this result is not an artifact of the small sample size. To explicitly test whether Koreanic and Japonic would still cluster together if language families were sampled equally, we ran the analysis 10 times on samples with two languages per language family. The result shows that if STRUCTURE does find structure in the data (i.e. if languages do not have an equal proportion of each ancestry), then Japonic and Koreanic are reliably clustered together even in small data sets (for more detailed information on this subsampling, see Supplementary Fig. S5).

4.2 Feature frequencies

While no particular feature can be taken as responsible for the population structure at hand, the STRUCTURE software provides information on the frequency of each feature in each ancestry. By using this frequency we can construct a structural 'profile' of each ancestry (Fig. 4). The formulations of the morphological and syntactic features that were used for the current study predominantly originate from the Grambank database, which itself is based on the feature set initially developed to capture the linguistic diversity of the languages of Sahul and Melanesia (Dunn et al. 2005). Many of the features relevant to that region are absent in Northeast Asia and were coded as '0' accordingly. Other features have such low frequencies in the languages of the sample (1–2 languages out of 60) that their contribution to the respective ancestries is minimal (the lower tail of the Morphology and Syntax graphs in Fig. 4). If we had a similar database of phonological features at our disposal, we would have a similar picture for phonological features, too. Since such a database does not exist yet, we compiled a set of features that captured the variation in the region well (some of them were mentioned in Robbeets 2017). While it is true that the features with low density in the area also contribute to the assignment of languages to ancestries, their effect is nevertheless low, as is the effect of the features present in almost all the languages and contributing equally to all ancestries.

This distribution comes about due to a high typological homogeneity of the sample: around one-fourth of morphological features and syntactic features are equally present in all language families and therefore have equal proportions in all ancestries. Other features stand out as present only in one or two ancestries. For example, SV word order (SV), postpositions (Postp), possessor-possessed order (PossPossessed), demonstrative- (DemN) and adjective-noun (AdjNoun, all in Syntax) order are common in all languages in the sample and are thus present in all ancestries to the same extent. These are likely to be widespread and common linguistic features, providing little diagnostic value for subgrouping.

What we are most interested in are the features in the middle of Fig. 4: these features have unequal proportions in the four ancestries and are decisive in attributing languages to ancestries. Some features contribute equally to two or three ancestries, while others are confined to one particular ancestry. For example, marking of S and A arguments on the verb by a suffix (features ASuffVerb and SSuffVerb in Morphology) is typical of Turkic, Tungusic, and some Mongolic languages, but not of Japonic and Koreanic languages. An inclusive/exclusive distinction (InclExcl in Morphology) is typical of some Tungusic and Mongolic, but not of Turkic, Japonic and Koreanic languages—and this is reflected in the frequency of this feature in the corresponding ancestries. A three-way contrast in demonstratives (Dist3Dem in Morphology) is a feature connecting Turkic, Koreanic, and Japonic languages, whereas the presence of ideophones (Ideophones in Morphology) is shared by Tungusic, Koreanic, and Japonic languages. Turkic and Mongolic languages use bare verb roots to form the imperative of the second person singular (VRoot2SGImper in Morphology), while Japonic, Koreanic, and Tungusic resort to a dedicated morpheme. The distribution of alienable/inalienable possession (AlienPoss in Morphology) contributes to the 'Tungusic' ancestry, which corresponds well to what we know about Tungusic languages (Tsumagari 1997). Adjectives that receive verbal marking are typical of Japonic and

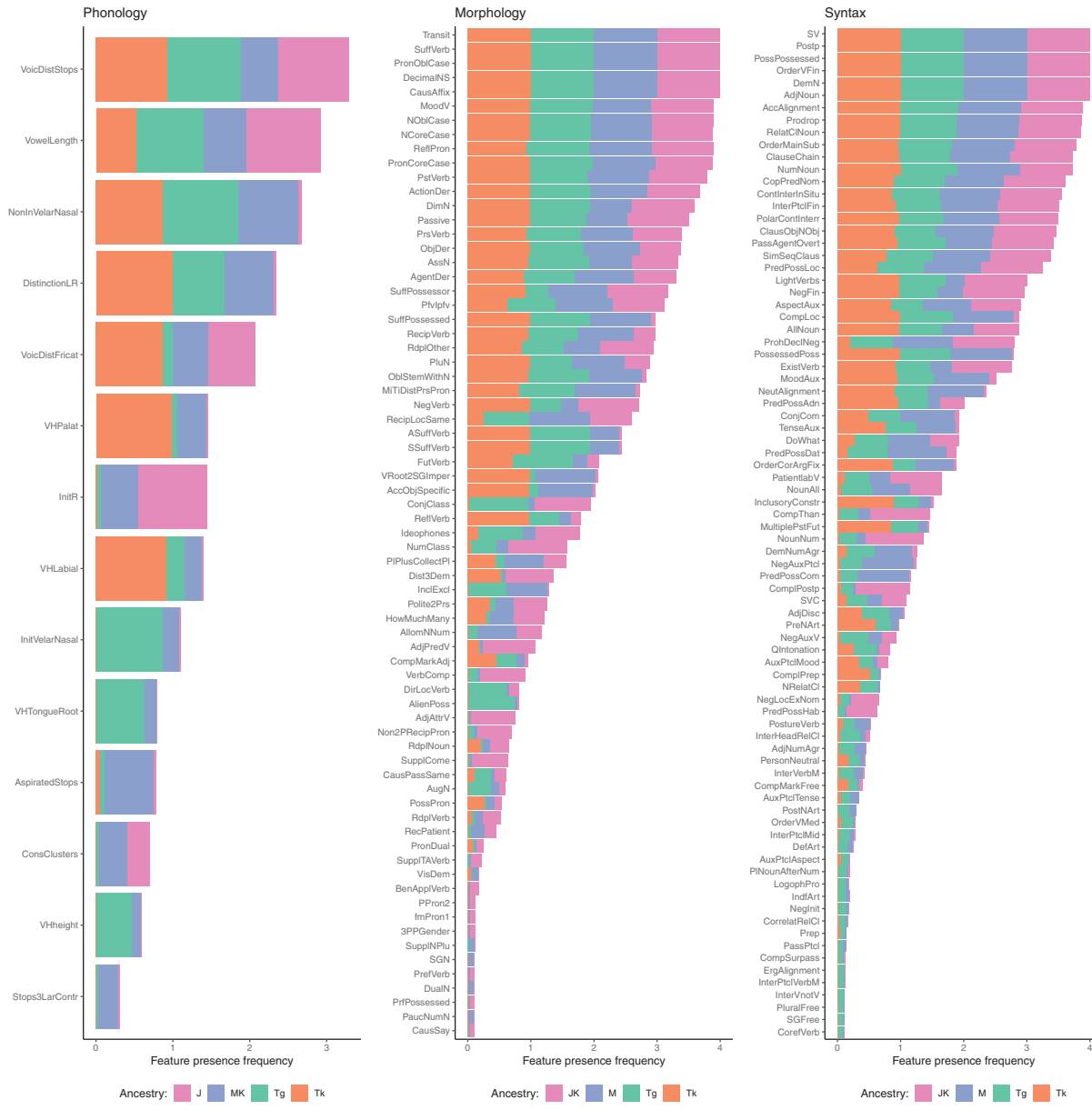

Figure 4 Reconstructability of features as being present/absent in the proto-language, following the results of ancestral state reconstruction by Hübler (2022), vs. contribution of features to the ancestries. The reconstruction of features corresponds to their contribution to respective ancestries: if a feature is reconstructed as absent in the proto-language, it is unlikely to contribute to the respective ancestry and vice versa.

Koreanic languages, and this is reflected in the contribution of these features to the 'Japono-Koreanic' ancestry (AdjAttrV and AdjPredV in Morphology). On the other hand, marking of the possessed by a suffix (SuffPossessed), oblique stems of personal pronouns ending in a nasal consonant (OblStemWithN), and a *mi/ti* distinction in personal pronouns (MiTiDistPrsPron, all in Morphology) are not typical of Japonic and Koreanic languages, but connect the three Micro-Altaic language families, Turkic, Mongolic, and Tungusic. However, the latter feature is widespread across all Eurasia and might be of areal rather than genealogical origin.

Among syntactic features, the features that distinguish Japonic and Koreanic languages are a

comparative construction with a marker that has neither a locational nor a 'surpass/succeed' meaning (CompThan), noun-numeral word order (NounNum), a postposed complementizer in the verbs of thinking/knowing (ComplPostp), predicative possession construction with a 'habeo'-verb (PredPossHab). There are only few typical Turkic syntactic features—in most cases, Turkic ancestry is defined as the absence of particular features, which are present in Tungusic and Mongolic, e.g. agreement between the demonstrative and the noun in number (DemNumAgr), negation marked by an auxiliary particle (NegAuxPtcl), predicative possession with a dative argument (PredPossDat), the order Noun-'all'(NounAll), a difference between the prohibitive and the declarative negation marking (ProhDeclNeg, all in Syntax and virtually absent in Turkic). Some of the few typical Turkic features are a preposed complementizer (ComplPrep in Syntax) in verbs of thinking/knowing (note that this is in contrast to Japonic and Koreanic languages, which use a postposed complementizer), a prenominal article (PreNArt, though not obligatory), an inclusory construction (InclusoryConstr) and multiple future/past tenses (MultiplePstFut).

We do not see a clear differentiation between Tungusic and Mongolic features in Syntax—the feature presence mostly appears in a symmetrical fashion, and this lack of clear-cut differences corresponds to the mixed ancestry profiles of these languages in Fig. 2. Auxiliary verbs used to mark negation (NegAuxPtcl), interrogation marked by intonation only (QIntonation), internally-headed relative clauses (InterHeadRelCl) have higher proportions of Tungusic ancestry than of any other ancestry. However, these features are at the lower end of the figure and have only marginal influence on the constitution of ancestries.

While the feature on the marking of predicative possession with a comitative argument contributes considerably to Tungusic and Mongolic ancestry (Fig. 4), it is present in such Turkic languages as Yakut and Dolgan, which also exhibit high proportions of Mongolic ancestry (Fig. 2). In Yakut, the proprietive suffix *-LA:X* is used to mark the possessed in a predicative possession construction (Pakendorf and Stapert 2020: 443). There is no agreement upon the origin of this suffix: the comitative suffix *-lUx* is already present in Middle Mongol and is reconstructed for Proto-Mongolic (Janhunen 2003). On the other hand, it might have a Turkic origin and has been used for possessive adjectival nouns (Schönig 2003). If it was borrowed from Turkic into Mongolic, then already at a much earlier time, probably Pre-Proto-Mongolic.

We can tentatively explain the clustering of Mongolic and Koreanic languages at the phonological level by the set of phonological features these two language families share: the most striking features present only in Mongolic and Koreanic languages are three laryngeal contrasts in stops and aspiration in stops. These features separate Koreanic languages from Japonic and some Mongolic languages from Turkic: Mongolic languages with an aspiration distinction in stops and/or three laryngeal contrasts in stops tend to share most/some of their ancestry with Koreanic languages and those without any of these features with Turkic. In Koreanic, the aspirated consonants arose from consonant clusters—Proto-Koreanic did not have a laryngeal contrast among consonants (Whitman 2012: 28) and it must have developed later in Middle Korean (Sohn 2015). In contrast, reconstructions of Proto-Mongolic show both strong and weak consonants (Janhunen 2003: 5) (aspiration is often one of the features of strong consonants), and the contrast between aspirated/unaspirated consonants is found in many Mongolic languages. Given the shared ancestry in phonology, a hypothesis that Mongolic and Koreanic languages converged in the course of their history is tempting. While sources on Koreanic mostly emphasize language contact with Chinese (Sohn 2020), sources on Mongolic mention the century-long Mongolic rule over Korea (starting from 1231, Rozycki 1990: 148). We cannot say with certainty whether aspirated stops in particular developed independently in Koreanic and Mongolic languages, but if horizontal transfer did happen, then the direction was most likely from Mongolic to Koreanic.

4.3 Correlation between features

The features in our data set are logically independent, i.e. given the value for one feature we cannot directly predict the value of another feature. However, there are known relationships between features in all language domains. The positively correlated features will have symmetrical ancestry proportions. Examples of such feature pairs are the order of the possessor and possessed (PossPossessed) vs. the presence of postpositions (Postp) and subject-verb order in intransitive clauses (SV) vs. verb-final word order in transitive clauses (OrderVFin, all in Syntax): both features in these pairs have a symmetrical distribution and, additionally, a low information value in the division of languages into ancestries. Another example is the distribution of voicing in sets of consonants: if a language has a voicing distinction in fricatives, it will most likely also have a distinction for stops. A similar implication can be assumed for the position of velar nasals: if a velar nasal is allowed in word-initial position, it is likely to be allowed in word-medial or word-final position as well.

The negatively correlated features have a complementary distribution: if one language has feature X,

it will most probably not have feature Y. The features of vowel harmony are not mutually exclusive, but it is rather unlikely that a language will have two types of vowel harmony. However, there is often both labial and palatal vowel harmony in Turkic languages; the typical complementary distribution is between tongue root and palatal vowel harmony. Another complementary pair is pointed out by Tsumagari (1997), namely the presence of a genitive case and of the alienability suffix: in Tungusic languages spoken in China the possessor in the attributive possession construction is marked by a suffix (in our sample, these are Solon and Manchu), but there is no alienability marking. This preference is characteristic of the Manchu-Mongolian complex. It also goes in line with the absence of S/A marking on verbs. These features in Tungusic languages are due to the interference with Manchu and Mongolian (both with an official status and impact), whereas Manchu itself was influenced by Mongolian and, more intensively, Chinese (Tsumagari 1997: 181–183).

If we excluded one of the features in these pairs, we would lose valuable information that might help distinguish the ancestries from each other. While we see velar nasals with a phonemic status across multiple ancestries, they can take in a word-initial position predominantly in Tungusic languages. Here we have two relevant features that are interlinked—the presence of a velar nasal in a word-medial or word-final position and the presence of a velar nasal in the word-initial position. While we could exclude some of these features, this would be an *a posteriori* decision, which risks 'cherry-picking' features that fit particular hypotheses. We therefore decided to leave these features in the analysis, but caution that future work should more carefully investigate their data sets to balance the risk of over-counting support for a particular grouping against artificially building in support for a grouping into the analysis.

4.4 Stability of structural features

There is an ongoing debate about the long-term stability of structural features and their use to identify language relationships (Nichols 1992; Dunn et al. 2008; Greenhill et al. 2017; Cathcart et al. 2018; Macklin-Cordes et al. 2021). However, it cannot be excluded that some structural features, like some parts of the lexicon, e.g. basic vocabulary, are useful in establishing genealogical relationships between languages (Nichols 1992). It has been argued that phonology and (inflectional) morphology provide better clues about linguistic descent than lexical data (Ringe et al. 2002: 65). Macklin-Cordes et al. (2021) measured phylogenetic signals for phonotactic data in 112 Pama-Nyungan languages and found a phylogenetic signal in binary (presence/absence of a biphone), segment-based and sound-class-based data sets. In particular, 39% of the total data set shows evidence of a phylogenetic signal and only 4% of characters are consistent with a phylogenetically random distribution. They describe their results as surprising, as previously it was assumed that Australian phonotactic restrictions are homogeneous and do not contain much historical information. Cathcart et al. (2018) use phylogenetic and spatial models of linguistic evolution to investigate the evolutionary dynamics of typological features. Their aim is to tease apart different forces that cause change, such as areal pressure, chance, and universal tendencies. Among other conclusions, they suggest that the development of particular word orders in Indo-European languages and the loss of verb agreement in several North Germanic languages are more likely to have been influenced by language contact than to have emerged due to other reasons. One of their results is that different word orders have different sources of loss and gain: V2 loss is highly areal, whereas V2 gain is not. A study on the stability of structural features based on the language sample of Transeurasian languages (Hübler 2021) suggests that levels of language grammar differ in their stability. Phonological and morphological features appear to be most stable (they change at a slower rate and have a higher phylogenetic signal), whereas features on the clause and nominal phrase level change at a faster rate and have a lower phylogenetic signal (Hübler 2022). Recent research on the evolution of Indo-European grammar compares morphological and syntactic features and concludes that morphological features (i.e. features that target phonologically bound elements) have a lower evolutionary rate (Carling and Cathcart 2021)—a finding our current results also support.

Our result that morphological features are especially stable (and thus better for reconstructing genealogical relationships) goes in line with the previous findings on the stability of structural features in Austronesian languages (Greenhill et al. 2017). Such features as inclusive vs. exclusive distinctions and gender distinctions fall into the slow-evolving category, and these are the features that belong to the morphological level, which shows here high precision in attributing languages to language families. The features on the relative order of elements (order of numeral and noun, order of subject and verb) are reported to be rather unstable (in the medium and fast rate categories), and such features belong to the syntactic level in this study, which shows highest levels of admixture. Our results are also consistent with suggestions that morphological features are the last to be borrowed in language contact situations (Thomason and Kaufman 1988); morphological

features show the lowest levels of admixture across all language families and recover language families with the least amount of false attributions.

Since we see that morphological features have the highest potential to carry a historical signal that might be due to descent and not areal dispersal, we suggest that the potential connection between Japonic and Koreanic may well be a historical clustering, i.e. is not just due to borrowing. Therefore, it is difficult to ascribe the morphological similarity to horizontal transfer. In terms of deeper relationships between the five families, however, we find little evidence for relationships above the family level beyond Japonic and Koreanic. While we do not wish to formally evaluate the evidence for or against Altaic and Transeurasian in this paper, we find little evidence for any deeper connections in these data. Instead, we find that the most likely clustering of these data is into the constituent language families, with a potential connection between Japonic and Koreanic. Perhaps this failure to identify deeper links between the putative Altaic family groups indicates a shortcoming in this approach, however we note that Reesink et al. (2009)'s analysis of Melanesian languages did find previously deeper connections suggesting that STRUCTURE can find these clusters in principle if they are present. We need more studies like ours and Reesink et al.'s on a wider range of languages and linguistic data to evaluate the potential of this approach for testing deep language relationships more fully.

4.5 Differences in admixture across languages and families

Language families differ in the level of admixture they exhibit. Japonic and Turkic languages appear as more or less homogeneous clusters across all tested *K*'s and language levels. The level of admixture across language domains varies most in Mongolic languages—these languages also show the highest admixture on average. This may be due to the fact that Mongolic languages diverged relatively recently (since the 13th century) and experienced a dialect chain break-up-like development.

Manchu stands out among other Tungusic languages at all levels and shows high levels of admixture. This can be explained by its known grammatical peculiarity (Gorelova 2002: 5–6): it forms its own branch among Tungusic languages, with only one more language belonging to it, Xibe, for which not enough material is available to consider it in the current study. Specifically, it is the most analytical language among all the Tungusic languages. Since there are no other strongly analytical languages in the sample, it cannot be assigned any particular ancestry, but rather shares almost equal proportions of three out of four ancestries (different combinations at different levels). It is hypothesized that analytical structures in Manchu are the predecessors of synthetic structures present in other Tungusic languages, and therefore Manchu can be viewed as more archaic than other Tungusic languages. In addition to this, Manchu stood under the constant influence of the Chinese language (Gorelova 2002: 27), but since there is no Sinitic language in the sample, we cannot see any 'Sinitic' ancestry in Manchu—it is rather reflected in a mix of other ancestries.

4.6 Sociolinguistic situation and language contact

One intriguing possibility is that admixture occurs at different levels depending on different types of sociolinguistic and language contact situations. For example, Thomason and Kaufman (1988: 37–38) state that, depending on the duration of cultural pressure from the source-language speakers, all language material can be borrowed, but that features of inflectional morphology would be the last to be borrowed, following phonological, phonetic, and syntactic elements. While lexical borrowing can occur even when there is casual contact, intensive long-term bilingualism is necessary for structural features to get borrowed. Thomason and Kaufman (1988) show on the example of contact and subsequent influence of Russian on Eskimo and English on Japanese that phonological features are the first to be incorporated into the language. Where some phonological borrowing has happened, syntactic borrowing is to be expected next.

Most often, we see parallel admixture profiles in phonology and syntax, suggesting that language contact was rather intensive. At other times, the amount of contact is highest in the syntactic domain, e.g. in North Siberian Turkic languages Yakut and Dolgan.[3] These languages have been in intensive contact both with Mongolic- and Tungusic-speaking groups (Even and Evenki in particular). On the one hand, Yakut speakers shift to Russian, on the other hand, other minority groups, like Even and Evenki, shift to Yakut (Pakendorf 2007). In such a situation, we would expect to find Tungusic features in Yakut and Dolgan—note the proportion of Tungusic ancestry in these languages in Fig. 2. The influence of these linguistic groups upon each other is not limited by phonological and syntactic borrowing, although this type of borrowing is prevalent. Among morphological borrowing, we see derivational and inflectional morphemes (and sometimes even paradigms) borrowed from Yakut into Evenki and Even, from Evenki into Yakut, from Mongolic into Yakut and Evenki, etc. (Anderson 2020). Some phonological differences between the closely related languages Yakut and Dolgan can be ascribed to the stronger Tungusic influence on Dolgan (Anderson 2020; Stapert 2013). The Mongolic influence (rather traces of Middle Mongol/

Written Mongolian or of several Mongolic dialects) upon Yakut was so strong that early investigators of the language could not unanimously decide upon its affiliation and suggested that it was mongolicized and then turkicized in the course of its history (Pakendorf 2007)—note the high proportion of Mongolic ancestry in its ancestry profile in Fig. 2 in Syntax.

It is rarely documented, which structural features were borrowed from which language at which stage. However, we have grounds to assume that in situations of prolonged intensive contact also structural borrowing took place. Turkic and Mongolic languages have been in constant contact throughout their history. In prehistoric times, Mongolic languages underwent the influence of Turkic languages. Bulgharic words were borrowed in Mongolic until the 4th century AD and Common Turkic loanwords are found in Middle Mongol. Among Turkic languages, Chagatai had the strongest impact on Middle Mongol. Starting from the 13th to 14th centuries, the direction of borrowing changed, and Turkic languages borrowed lexical and morphosyntactic material from Mongolic languages. Especially prominent is the influence of Middle Mongol on the Chagatai phonetics (Schönig 2003)—note that Chagatai exhibits around 50% of Mongolic ancestry at the level of phonology. Other Turkic languages, such as Yakut, Tuvan, and Khakas, underwent Mongolic influence after the Middle Mongol period. Tuvan stayed in contact with Mongolic languages, such as Khalkha, Oirat, and Buriat, also afterwards (Schönig 2003). Until 1900, the Tuvan language was not written, and the only literate speakers could read and write Mongolian (Krueger 1997: 87). There are Mongolic traces in the phonology and syntax of Tuvan, which can be ascribed to the prolonged contact with Mongolic languages. In particular, the long vowels are not originally Turkic, but most probably a Mongolic loan (Krueger 1997: 96–97). This contact history is consistent with its admixture profile in Fig. 2, which shows around 30% of Mongolic ancestry at the level of phonology and syntax. The admixture profiles of Turkic and Mongolic languages support the general assumption that phonological and syntactical borrowing precedes morphological borrowing: we see a high amount of admixture between Turkic and Mongolic languages especially at the phonological level, which corresponds to the first stage of structural interference according to Thomason and Kaufman (1988).

While Tungusic languages on both sides of the Chinese-Russian border have been influenced by Chinese, Mongolian, and Manchu (though Tungusic, Manchu is very different from other Tungusic languages), the Tungusic languages spoken in East Siberia show features that would be expected from intensive contact with Russian (Tsumagari 1997), such as agreement in case and number between a modifier and a noun. Most often, the influence goes in the direction of Chinese, Mongolian, Manchu, Yakut, and Russian into Tungusic languages. Tsumagari (1997: 183) come thus to a conclusion that 'the linguistic diversity within Tungusic reflects past contacts with the prestige languages' in each of the areas, where these languages are spoken (Manchuria, Lower Amur, East Siberia). While we see high Mongolic and/or Turkic ancestry proportions in these languages (Fig. 2, Syntax), we cannot see Russian or Chinese impact, because they are not included in the sample and their influence might be masked as some other ancestry.

An effective predictor of the category of the features to be borrowed is the typological distance between the languages in contact. Since Mongolic and Turkic languages are very close typologically, verb stems could be easily borrowed and equipped with the native suffixes (Schönig 2003).

Taking all this reasoning into account, we would suggest that the intensity of contact accounts for the most diversity between the interference patterns among language levels: where the contact was rather shallow, we see more phonological borrowing. With the intensification of contact syntactic borrowing joins in. Only prolonged intensive contact leads to borrowing of morphological features.

4.7 Limitations

One potential limitation of the approach we have applied here is that the method can only identify admixture between languages sampled in the data set, which can impact the interpretations (Lawson et al. 2018). One prominent example of this limitation here is Chuvash, a Turkic language belonging to the Bulgaric branch and its sole surviving representative. While all Turkic languages are very similar in terms of grammar, Chuvash differs significantly from the Turkic profile. Some of this differentiation looks to be caused by random innovations ascribed to its early divergence from the Turkic lineage (around 2000 years from other languages, Savelyev and Robbeets 2020), other differences result from language contact, especially with the Uralic languages. The isolation of Chuvash is reflected in its admixture profile: it has different amounts of 'Mongolic' and 'Tungusic' ancestries at different levels. What we cannot see in its admixture profile, is Uralic ancestry, because no Uralic languages were included in the study. This is a general limitation to the interpretation of the results: while STRUCTURE is generally a helpful resource, it can only provide feedback on the data it was given as input. If we do not include Uralic languages in the study, but their influence is relevant to the region, 'Uralic' ancestry will be masked as some

other ancestry derived from the given data. Similarly, the converse is true, if we were to include a completely unrelated language family—Mayan, perhaps—then the admixture profile would indicate shared similarities between Mayan and these languages. These limitations can be avoided, however, by inspecting the features that STRUCTURE allocates to each ancestry component. For example, if all languages are admixed to a similar degree, then this would mean there was no inherent structure in the data such as we would expect when comparing unrelated groups like Mayan and Turkic, and any shared features should be linguistically trivial (i.e. very common features showing chance similarity).

Despite these limitations, the approach used in this study helps us correctly identify three out of the five language families (while two families are too similar structurally and share the ancestry). This means, on the one hand, that the information stored in structural features is sufficient to attribute languages to language families, and, on the other hand, that a method accounting for both inheritance and borrowing provides valid results in terms of genealogical relationships between languages. The grouping of language families with each other differs, depending on the language level: we find Turkic, Mongolic, Tungusic, and Japono-Koreanic at morphological and syntactic levels, but Turkic, Tungusic, Japonic, and Mongolo-Koreanic at the phonological level.

5. Conclusions

One of the critiques of structural features is that they diffuse easily and that it is difficult to trace and consequently exclude borrowings. STRUCTURE offers an elegant solution to this problem: it is compatible with an interpretation in terms of vertical descent and horizontal transmission and provides information on the level of admixture between individuals—in our case languages. Therefore, there is no need to determine and eliminate borrowed features in advance: their presence is visible in the results, their sources can be more easily interpreted, they do not impact the conclusions in a negative way and do not invalidate them. Nevertheless, the results should be treated with caution: ancestries of languages not present in the sample can be masked as 'false' ancestries, and language families with only a few members tend to cluster with language families with more members.

Our analysis shows that morphological features have the strongest genealogical signal and syntactic features diffuse most easily. When using only morphological structural data, the model is able to correctly identify three language families: Turkic, Mongolic, and Tungusic, whereas there are not enough structural dissimilarities between Japonic and Koreanic languages to assign them to different ancestries. Even a small number of phonological features can help put preliminary language family boundaries: with only 16 phonological features we are able to postulate Turkic, Tungusic and Japonic language families, whereas 82 syntactic features are not enough to find clear boundaries of the Tungusic language family. Now that the results here show that morphological structural features have an especially precise historical signal, one can use them to establish relations between other language families, for which no relatives are known because of the time limitations of the comparative method.

The approach we have applied here provides a powerful way forward for debates about macro-family relationships. First, language structures can readily be evaluated and identified, even on a global scale (Skirgård et al. in press), without having to postulate controversial proto-forms. Second, the STRUCTURE analysis is agnostic as to whether the groupings reflect shared ancestry or admixture between languages meaning that researchers can include a range of data and then evaluate the reasons for the clusters on a per-feature basis later. Third, the clustering approach here provides a computationally feasible solution to the problem of combinatoric explosion of comparisons in larger data sets. We suggest that this approach will help move these long-standing—and acrimonious—debates onto a more solid quantitative footing that will enable us to carefully and robustly identify language relationships at a deeper level.

Supplementary Data

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

Author contributions

N.H. designed research; N.H. performed research; S.J.G., N.H. analyzed data; N.H., S.J.G. wrote the paper.

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 and from the Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology. We thank Brigitte Pakendorf for the idea of applying STRUCTURE to this data and Ron Hübler for his help in writing the code collecting the results. We also thank two anonymous reviewers for helpful discussion.

Data availability

The data used for the analysis in the manuscript can be found at: https://zenodo.org/record/5720838\#.YmJz8y0RppQ. The code, detailed results, plots and other materials can be found at https://zenodo.org/record/7188422\#.Y0aADS8Rr0o.

Notes

1. Shirongol languages and Shira Yughur, spoken mostly in Gansu Province, China.
2. Nanai, Orok, Ulch.
3. Dolgan is substantially different from Yakut in terms of lexicon and phonetics. Structurally, however, these two languages are very similar.

References

Anderson, Gregory D. S. (2020) 'Form and Pattern Borrowing Across Siberian Turkic, Mongolic, and Tungusic Languages'. In: Robbeets, M., and A. Savelyev (eds.) *The Oxford Guide to the Transeurasian Languages*, pp. 715–725. Oxford: Oxford University Press.

Bowern, Claire (2012) 'The Riddle of Tasmanian Languages', *Proceedings of the Royal Society B: Biological Sciences*, 279: 4590–4595.

Carling, Gerd, and Chundra Cathcart (2021) 'Reconstructing the Evolution of Indo-European Grammar', *Language*, 97(3):561–598.

Cathcart, Chundra, et al. (2018) 'Areal Pressure in Grammatical Evolution', *Diachronica*, 35: 1–34.

Dunn, Michael, et al. (2008) 'Structural Phylogeny in Historical Linguistics: Methodological Explorations Applied in Island Melanesia', *Language*, 84: 710–59.

Dunn, Michael J., et al. (2005) 'Structural Phylogenetics and the Reconstruction of Ancient Language History', *Science*, 309: 2072–5.

Durie, Mark, and Malcolm Ross (1996) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. New York & Oxford: Oxford University Press.

Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet. (2005) 'Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: A Simulation Study', *Molecular Ecology*, 14: 2611–2620.

Felsenstein, Joseph. (1978) 'The Number of Evolutionary Trees', *Systematic Biology*, 27: 27–33.

Francis-Ratte, Alexander T., and Marshall Unger (2020) 'Contact Between Genealogically Related Languages: The Case of Old Korean and Old Japanese', In: Robbeets, Martine and Alexander Savelyev (ed.) *The Oxford Guide to the Transeurasian Languages*, pp. 705–14. Oxford: Oxford University Press.

Georg, Stefan. (2007) 'Review of Martine Robbeets: Is Japanese related to Korean?', *Turcologica*, 64: 259–91.

Gorelova, Liliya M. (2002) *Manchu Grammar*. Brill Academic Publishers.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill (2009) 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323: 479–83.

Greenhill, Simon. (2015) 'Demographic Correlates of Language Diversity', In: *The Routledge Handbook of Historical Linguistics*, pp. 557–578. Abingdon & New York: Routledge.

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, et al. (2017) 'Evolutionary Dynamics of Language Systems', *Proceedings of the National Academy of Sciences*, 114: E8822–29.

Grollemund, Rebecca, et al. (2015) 'Bantu Expansion Shows that Habitat Alters the Route and Pace of Human Dispersals', *Proceedings of the National Academy of Sciences*, 112: 13296–13301.

Hammarström, Harald, et al. (2020) Glottolog 4.2.1. Max Planck Institute for the Science of Human History. Accessed 03 June, 2020. https://glottolog.org/accessed2020-06-03.

Heggarty, Paul. (2013) 'Ultraconserved Words and Eurasiatic? The 'Faces in the Fire' of Language Prehistory', *Proceedings of the National Academy of Sciences*, 110: E3254.

Hubisz, Melissa J., et al. (2009) 'Inferring Weak Population Structure with the Assistance of Sample Group Information', *Molecular Ecology Resources*, 9: 1322–32.

Hübler, Nataliia. (2021) hueblerstability. <https://doi.org/10.5281/zenodo.5720838>.

Hübler, Nataliia. (2022) 'Phylogenetic Signal and Rate of Evolutionary Change in Language Structures', *Royal Society Open Science*, 9: 211252.

Jacques, Guillaume, and Johann-Mattis List (2019) 'Save the Trees: Why We Need Tree Models in Linguistic Reconstruction (and When We Should Apply Them)', *Journal of Historical Linguistics*, 9: 128–167.

Janhunen, Juha. (2003) 'Proto-Mongolic', In: Janhunen, Juha (ed.) *The Mongolic Languages*, pp. 1–29. London and New York: Routledge.

Johanson, Lars, and Martine Irma Robbeets (2010) *Transeurasian Verbal Morphology in a Comparative Perspective: Genealogy, Contact, Chance*, Vol. 78. Otto Harrassowitz Verlag.

Koile, Ezequiel, et al. (2022) 'Phylogeographic Analysis of the Bantu Language Expansion Supports a Rainforest Route', *Proceedings of the National Academy of Sciences*, 119: e2112853119.

Kolipakam, Vishnupriya, et al. (2018) 'A Bayesian Phylogenetic Study of the Dravidian Language Family', *The Royal Society Open Science*, 5: 171504.

Krueger, John R. (1997) *Tuvan Manual, Volume 126 of Uralic and Altaic Series*. Bloomington: Indiana University Press.

Lawson, Daniel J, Lucy Van Dorp, and Daniel Falush (2018) 'A Tutorial on How Not to Over-Interpret STRUCTURE and ADMIXTURE Bar Plots', *Nature Communications*, 9: 1–11.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, and Philippe Lopez, Bapteste Eric. (2016) 'Unity and Disunity in Evolutionary Sciences: Process-based Analogies Open Common Research Avenues for Biology and Linguistics', *Biology Direct*, 11: 39.

Macklin-Cordes, Jayden L, Claire Bowern, and Erich R. Round (2021) 'Phylogenetic Signal in Phonotactics', *Diachronica*, 38: 210–258.

Mahowald, Kyle, and Edward Gibson (2013) 'Short, Frequent Words are more Likely to Appear Genetically Related by Chance', *Proceedings of the National Academy of Sciences*, 110:E3253.

Martin, Samuel E. (1966) 'Lexical Evidence Relating Korean to Japanese', *Language*, 42: 185–251.
Matisoff, James A. (1990) 'On Megalocomparison', *Language*, 66: 106–20.
Miller, Roy Andrew. (1971) *Japanese and the Other Altaic Languages*. University of Chicago Press.
Nichols, Johanna. (1992) *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.
Norvik, Miina, et al. (2022) 'Uralic Typology in the Light of a New Comprehensive Dataset', *Journal of Uralic Linguistics*, 1: 4–42.
Pagel, Mark, et al. (2013) 'Ultraconserved Words Point to Deep Language Ancestry Across Eurasia', *Proceedings of the National Academy of Sciences*, 110: 8471–6.
Pakendorf, Brigitte. (2007) *Contact in the Prehistory of the Sakha (Yakuts): Linguistic and Genetic Perspectives*. Doctoral Dissertation, Leiden University.
Pakendorf, Brigitte, and Eugénie Stapert (2020) 'Sakha and Dolgan, the Northern Siberian Turkic Languages', In: Robbeets, Martine and Alexander Savelyev (eds.) *The Oxford Guide to the Transeurasian Languages*, pp. 430–45. Oxford: Oxford University Press.
Pawley, Andrew. (2012) 'How Reconstructible is Proto Trans New Guinea? Problems, Progress, Prospects', In: Hammarström, Harald and Wilco van den Heuvel (ed.) *History, Contact and Classification of Papuan Languages*, Vol. 1, pp. 88–164. Port Moresby: Linguistic Society of Papua New Guinea.
Poppe, Nicholas N. (1960) *Vergleichende Grammatik der altaischen Sprachen [Comparative Grammar of the Altaic Languages], Volume I: Vergleichende Lautlehre [Comparative phonology]*. Wiesbaden: Otto Harrassowitz.
Poppe, Nicholas N. (1965) *Introduction to Altaic Linguistics*. Wiesbaden: Otto Harrassowitz.
Poppe, Nikolaj Nikolaevič. (1975) 'Altaic Linguistics: An Overview', *Gengo no kagaku [Sciences of Language]*, 6: 130–86.
Porras-Hurtado, et al. (2013) 'An Overview of Structure: Applications, Parameter Settings, and Supporting Software', *Frontiers in Genetics*, 4: 98.
Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly (2000) 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics*, 155: 945–959.
Pritchard, Jonathan K, et al. (2010). *Documentation for STRUCTURE software: Version 2.3*. University of Chicago, Chicago, IL, 1-37.
Ramstedt, Gustaf John. (1924) 'A Comparison of the Altaic Languages with Japanese', *Transactions of the Asiatic Society of Japan Second Series*, 7: 41–54.
Reesink, Ger, Ruth Singer, and Michael Dunn (2009) 'Explaining the Linguistic Diversity of Sahul Using Population Models', *PLoS Biology*, 7: e1000241.
Ringe, Don. (1995) '"Nostratic" and the Factor of Chance', *Diachronica*, 12: 55–74.
Ringe, Don. (1999) 'How Hard is it to Match CVC-Roots?', *Transactions of the Philological Society*, 97: 213–244.
Ringe, Don, Tandy Warnow, and Ann Taylor (2002). 'Indo-European and Computational Cladistics', *Transactions of the Philological Society*, 100: 59–129.
Robbeets, Martine. (2017) 'The Transeurasian Languages', In: *The Cambridge Handbook of Areal Linguistics*, pp. 586–626. Cambridge University Press.
Robbeets, Martine. (2020a) 'The Classification of the Transeurasian Languages', In: Robbeets, Martine and Alexander Savelyev (eds.) *The Oxford Guide to the Transeurasian Languages*, pp. 31–39. Oxford: Oxford University Press.
Robbeets, Martine. (2020b) 'The Typological Heritage of the Transeurasian Languages', In: Robbeets, Martine and Alexander Savelyev (eds.) *The Oxford Guide to the Transeurasian Languages*, pp. 127–44. Oxford: Oxford University Press.
Robbeets, Martine, et al. (2021) 'Triangulation Supports Agricultural Spread of the Transeurasian Languages', *Nature*, 599: 616–621.
Ross, Malcolm D. (1996) 'Contact-induced Change and the Comparative Method: Cases from Papua New Guinea', In: Durie, Mark and Malcolm D. Ross (eds.) *The Comparative Method Reviewed*, Vol. 24, pp. 180–218. Oxford: Oxford University Press.
Rozycki, William V. (1990) 'A Korean Loanword in Mongol?', *Mongolian Studies*, 13: 143–151.
Savelyev, Alexander, and Martine Robbeets (2020) 'Bayesian Phylolinguistics Infers the Internal Structure and the Time-depth of the Turkic Language Family', *Journal of Language Evolution*, 5: 39–53.
Schleicher, August. (1853) 'Die Ersten Spaltungen des Indogermanischen Urvolkes', *Allgemeine Monatsschrift für Wissenschaft und Literature*, 3: 786–7.
Schönig, Claus. (2003) 'Turko-Mongolic relations', In: Janhunen, Juha (ed.) *The Mongolic Languages*, pp. 403–19. London and New York: Routledge.
Skirgård, H., H. J. Haynie, D. E. Blasi, et al. (in press) 'Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss'. Science Advances
Sohn, Ho-Min. 2015. Middle Korean and Pre-Modern Korean. In *The handbook of Korean linguistics*, ed. Lucien Brown and Jaehoon Yeon, 439–458. Malden, MA: John Wiley & Sons, Inc.
Sohn, Ho-min. (2020) 'Language Contact in Korean', In: *The Oxford Handbook of Language Contact*, pp. 540–55. Oxford University Press.
Stapert, Eugénie. (2013) *Contact-induced Change in Dolgan: An Investigation into the Role of Linguistic Data for the Reconstruciton of a People's (Pre-)History*. Leiden University.
Starostin, Sergei A, Anna Dybo, Oleg Mudrak, and Ilya Gruntov (2003) *Etymological Dictionary of the Altaic Languages*. Leiden: Brill.
Syrjänen, Kaj, et al. (2016) 'Applying Population Genetic Approaches within Languages: Finnish Dialects as Linguistic Populations', *Language Dynamics and Change*, 6: 235–283.
Thomason, Sarah Grey, and Terrence Kaufman (1988) *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.
Tian, Zheng, et al. (2022) 'Triangulation Fails When Neither Linguistic, Genetic, nor Archaeological Data Support the Transeurasian Narrative', *bioRxiv*, preprint: not peer reviewed <https://www.biorxiv.org/content/early/2022/06/12/2022.06.09.495471>. doi:10.1101/2022.06.09.495471. Stamp date 06 September, 2022.

Tsumagari, Toshiro, (1997) 'Linguistic Diversity and National Borders of Tungusic', *Senri Ethnological Studies*, 44: 175–86.

Vajda, Edward. (2020) 'Transeurasian as a Continuum of Diffusion', In: Robbeets, Martine and Alexander Savelyev (eds.) *The Oxford Guide to the Transeurasian Languages*, pp. 726–34. Oxford: Oxford University Press.

Vovin, Alexander. (2005) 'The End of the Altaic Controversy. In memory of Gerhard Doerfer', *Central Asiatic Journal*, 49: 71–132.

Vovin, Alexander. (2010) *Koreo-Japonica: A Re-evaluation of a Common Genetic Origin*. Honolulu, HA: University of Hawai'i Press.

Vovin, Alexander. (2017) 'Origins of the Japanese Language', In: *Oxford Research Encyclopedia of Linguistics*. Retrieved 17 May 2022, from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-277.

Whitman, John. (2011) 'Northeast Asian Linguistic Ecology and the Advent of Rice Agriculture in Korea and Japan', *Rice*, 4: 149–158.

Whitman, John B. (2012) 'The Relationship Between Japanese and Korean', In: Tranter, Nicholas (ed.) *The Languages of Japan and Korea*. London: Routledge.