Evolutionary Dynamics of Structural Features

Dissertation
(kumulativ)

Zur Erlangung des akademischen Grades doctor philosophiae
(Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von M.A. Nataliia Hübler
geboren am 04. Oktober 1991 in Nischyn, Ukraine

Gutachter/Gutachterin:

1. Prof. Dr. Volker Gast
Friedrich-Schiller-Universität Jena, DE
Institut für Anglistik/Amerikanistik
Sprache und Kommunikation
Ernst-Abbe-Platz 8
D-07743 Jena

2. Prof. Michael Dunn
Universität Uppsala, SE
Engelska parken
Thunbergsvägen 3H

Tag der Verteidigung: 06.06.2023

Contents

1. Introduction

Historical linguistics tries to answer questions about language change, describe the relationships between modern languages and reconstruct their prehistory. However, there is a certain “time barrier” around 6000–10000 BP (Nichols 1992, Greenhill et al. 2017), beyond which classical historical linguistics cannot make inferences about the history of languages. The most frequent approach in historical linguistics uses basic vocabulary to establish relationships between languages and a tree model to explain the relationships in more detail and put language diversification onto a timeline. While this approach is sufficient to uncover relationships *within* language families, it does not unequivocally solve debates around the relationships *between* language families, as e.g. the Altaic/Transeurasian hypothesis, because of the higher time depth. One possibility to extend the time depth is to use grammatical structures. However, the stability of structural features is a debated question, which needs to be clarified before they can be used to investigate deep past. To test the stability of structural features, I use a language sample covering languages spoken all over Eurasia and known under the cover terms (Macro-)Altaic and Transeurasian. The relationships between these languages are well understood and many of them have been in contact with each other and with the neighbouring languages throughout the history. If we are nevertheless able to show that structural features (or at least a set thereof) coded for these languages carry a genealogical signal, then we can use these features to test hypotheses about deep relationships between language families. The procedure used to evaluate the stability of structural features in the current thesis is thoroughly documented and publicly available, so that linguists working on other language families can further advance the research on the evolutionary dynamics of structural features by applying the methods presented here to their own data.

1.1 Structural features

A structural feature is an abstract construct, which describes

- the way something is expressed in a language, e.g. predicative possession: does it use a *habeo*-verb (‘I have a cat.’) or a particular marker, such as locative (lit. ‘The cat is at me.’), dative (lit. ‘The cat is to me.’), comitative (lit. ‘The cat is with me.’)? Is it expressed as an adnominal possession (lit. “My cat exists.”)?
- an availability of a certain marker in the language, e.g. is there a past/present/future tense marker in the language?
- a particular contrast a language makes, e.g. between voiced and voiceless consonants,
- the position of a particular marker or a word in relation to other words, e.g. order of subject and verb, order of adjective and noun, and other grammatical phenomena.

Here, it is formulated as a question about the presence or absence of a particular structure or a marker in the language (for more details on the data and the features, see Section 1.4). Given the sentence from the language grammar in (1), we have already some information on several features. For example, we can answer the question “Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?” with a “yes” and code the feature as “1”: here, we see the marker *-n* 3SG on the verb, which refers to the A argument of the clause, *akin* ‘father’. We also see no plural marking on the noun *ɲami* ‘female deer’, therefore, we have a good indication of a “no”/“0” for the feature “Do cardinal numerals require agreement on noun phrases?”. There is a genitive marker on the noun *akin* ‘father’, *-mi:*, which leads to a “yes”/“1” for the feature “Can adnominal possession be marked by a suffix on the possessed noun?”. The features on case marking “Are there morphological cases for non-pronominal core arguments (i.e. S/A/P)?” and “Are there morphological cases for oblique independent personal pronouns (i.e. not S/A/P)?” can also be coded as present, given that there is an accusative marker on the P argument, the noun *ɲami* ‘female deer’ (*ɲami-βa* female.deer-ACC), and a dative marker on the recipient, the first person singular pronoun (*min-du:* 1SG-DAT), respectively. Ideally, there are more examples and a short description in the grammar to support the decision on feature coding.

(1) Evenki (Tungusic, Bulatova and Grenoble 1999: 8)

akin-mi:	*min-du:*	*tunŋa-ßa*	*ɲami-ßa*	*ani:-ra-n*
father-POSS.1SG	1SG-DAT	five-ACC	female.deer-ACC	give-AOR-3SG

‘My father gave me five female deer.’

1.2 Stability of structural features

The utility of structural features for historical linguistics is an unsettled question. While Nichols (1992) suggests that structural typology can push back the time limits of historical linguistics back to the dawn of the Neolithic and Dunn et al. (2005) assume that a faint structural signal may reach further back in time than a lexical signal, other scholars deny any phylogenetic signal in structural features altogether (Wichmann and Holman 2009), and still others emphasize the especially attenuated historical signal in them, which can arise through borrowing and inheritance (Reesink et al. 2009). Macklin-Cordes et al. (2021: 212) point out the problems that come along with historical linguistics limiting itself to only lexical data, such as bottle-neck effects and inherent limitations associated with it. The main problem lies in the debate around the stability of structural features: do they evolve slow enough to enable inferences about deeper past and is inheritance form a common ancestor the main source of the structural similarities across languages?

Greenberg (1978: 76) introduces the notion of genealogical stability of a structural feature as follows: “If a particular property rarely arises but is highly stable when it occurs, it should be fairly frequent on a global basis but be largely confined to a few linguistic stocks”. Greenberg thus acknowledges the frequent criticism of structural features for their having a limited number of states and incorporates it in the definition of stability. We can translate this definition of a stable feature into quantitative terms used in the thesis as: a feature is stable if it has a low gain rate (= rarely arises) and a low loss rate (= is highly stable when it occurs).

Nichols (1993) contrasts the “basic grammar” and its unknown rate of change and the basic vocabulary and its known rate of change. Already back then she suggested that using a set of stable structural features complemented with the comparative method would enable us to find higher-level connections *between* well-established language families. She defines *stable* as “minimally prone to be borrowed and maximally prone to be inherited”, and states that stable features are “usable as indicators of probable genetic relatedness a step or two beyond the levels the standard comparative method can now reach” (Nichols 1993: 339). Stable features have to be both persistent in their language families (high likelihood of being inherited) and have a low probability of borrowing (Nichols 1993: 354). Features that dominate cross-linguistically, e.g. SOV order and causatives, cannot be taken as a proof of areal or genetic stability. On the contrary, cross-linguistically rare features, if present consistently in a family or in an area, can indeed be called stable. She comes to a conclusion that head/dependent marking, alignment, inclusive/exclusive oppositions, genders, number oppositions in the noun and detransitivization processes are genetically stable features, whereas causatives (and similar transitivizers) and clause word order are areal features. Numeral classifiers and tones take in an intermediate position (Nichols 1993: 353).

Ten years later, Nichols (2003: 284) recaps the definition of a stable feature as "more resistant to change, loss, or borrowing (than other elements of language)". She notes that a feature can be stable within a language family, within an area or both in a family and in an area. For example, we can speak of first person pronoun stem suppletion as a stable feature in Indo-European, but not cross-linguistically. While some features, which are well-spread in a language family, are likely to be inherited and are rarely lost, there are also "recessive" features, which are scarcely spread in a language family (not always inherited), but are also unlikely to be borrowed by other languages. One such feature is ergative alignment in Indo-European. It is a prominent example illustrating that the probability of inheritance and probability of acquisition are independent. In the current thesis, the possibility of detection of recessive features motivates separate treatment of the rate of gain and rate of loss (Chapter 3). There are also some features known to be typical of and stable in Northern Eurasia, e.g. personal pronouns *mi-Ti* pattern in at least oblique forms (Uralic, Turkic, Mongolic, Tungusic, Yukaghir), SOV word order (although this order is also found elsewhere in the world), vowel harmony etc. The third group of stable features are stable both in areas and in language families, the only drawback being that these features also easily diffuse into other languages.

Wichmann and Holman (2009) define stability as "the probability that a given language remains unchanged with respect to the feature during 1000 years, that is, the feature undergoes neither internal change nor diffusion during the interval." They introduce metrics to measure stability and compare their performance on a simulated data set. Their results show that the best metric is the one that assesses the similarity of related languages with respect to the feature compared to unrelated languages. This metric is reminiscent of the definition of the phylogenetic signal and the Fritz and Purvis' D (used in Chapter 3). They divide the features into four categories based on the relative stability: very unstable, unstable, stable, very stable. Features that have a strong pragmatic motivation tend to be unstable or very unstable. Such features include stress and rhythm, plural marking, definite and indefinite articles, politeness distinctions in pronouns, imperative-hortative systems, epistemic or evidentiality distinctions, negation and distance contrasts in demonstratives. Stable features comprise gender, affix and constituent ordering and case marking.

The question of the stability of structural features has been gaining more attention in linguistic scholarship recently, and the latest research suggests that stable structural features are well comparable to core basic vocabulary in their usability for deep historical reconstruction (Dediu and Levinson 2012) and, although most grammatical features change faster than basic vocabulary, there is a set of structural features that evolve at a slow rate (Greenhill et al. 2017).

A recent study uses phylogenetic comparative methods (see Section 1.5) to reconstruct the grammar of Proto-Indo-European (Carling and Cathcart 2021). Along with the ancestral state reconstruction, they measure the rate, at which features evolve, and categorize the features into four groups based on this rate. First, there are features that are lost and gained at a high rate. In Indo-European, such features are presence of case on adjectives, clitics, distinctions between dative and genitive marking, absence of case on nouns, and alignment systems in the simple past. Second, there is a group of features of high stability, to which languages are frequently attracted. This group includes presence of case on nouns, case difference between A and O for nouns and pronouns, masculine/feminine distinction, noun-relative word order, possessor-noun word order, present progressive by auxiliary and absence of neuter gender and vocative case. Third, there are "recessive" features, which are more rarely inherited and are likely to be lost even if they get inherited. Such features are presence of future tense by participle, future tense by particle, more than seven cases, more than five genders, tripartite alignment, ergative alignment in pronouns, active-stative alignment, double oblique alignment, V2 word order and VSO word order. Lastly, there is a group of highly stable features, which rarely arise and are rarely lost. These features are presence of adjective-noun word order, agglutination for case, agreement on prepositions, case on the last member of an NP, definite articles, definite suffixes on adjectives, definite suffixes on nouns, neuter gender, a noun class for animates and a synthetic future tense. Carling and Cathcart (2021: 25) find the following relationships between the evolutionary rate and reconstructability of features: features that are frequently gained and almost never lost are reconstructed with high probability, whereas features that are frequently lost and almost never gained are reconstructed with a low probability. There is more variation in reconstructability in other two groups, so that it is difficult to postulate an equivocal relationship between the rate and the reconstructability of the features.

Despite the obviousness of the necessity of “basic grammar” and the almost thirty years that have passed by since Johanna Nichols suggested it, we still do not have a set of universally stable features. The investigation of the stability of structural features has the potential to advance the field of historical linguistics in several ways. First, it would enable historical linguists to reconstruct the grammars of ancient languages and give us an idea of what the languages spoken 10 000 years ago looked like. Second, it would enable historical linguists to search for links between language families not previously established as related and thus make inferences about the origin of language families and language isolates. Connecting this linguistic evidence with genetic and archaeological evidence would let us learn more about early population movements and human prehistory in general.

While we might not be able to construct a set of cross-linguistically stable features, we can approach it by investigating the differences in stability of structural features and their sources. Further development of the field of stability in language would bring us nearer to the investigation of the deep language history as well as or even at a higher resolution than we can currently achieve with basic vocabulary.

This dissertation addresses the stability of structural features from a quantitative perspective by measuring the phylogenetic signal and the evolutionary rate of change and tests the performance of different sets of structural features in uncovering genealogical relationships between languages.

1.3 Historical linguistics and the Altaic question

In order to establish a relationship between languages, historical linguists must show that languages fulfill a number of conditions. They collect lists of morphemes and vocabulary items with identical or nearly identical meaning among the languages, which are potentially related, and try to establish regular sound correspondences (Hale 2003, Weiss 2015: 128). These form-meaning correspondences are crucial for the investigation of genealogical relationships between languages and provide ground for one of the main criticisms of structural features, which necessarily lack the “form” part per definition.

After establishing a genealogical relationship between languages, historical linguistics can proceed with reconstructing ancestral languages with the help of the Comparative Method. This method allows to reconstruct the phonological system, some vocabulary and parts of grammar of languages spoken thousands of years ago, for which no written attestations are available. Such reconstructed languages are called "proto-languages".

The Comparative Method resorts on the family tree as a model of diversification. The tree of descent was used to describe the development of the Indo-European language family as early as 1853 by August Schleicher. While the tree model is useful in many cases and allows us to make inferences about the proto-languages and their age if combined with Bayesian modelling like in BEAST (Bouckaert et al. 2014), it does not take into account other processes that shape language evolution (François 2015).

One of the restrictions of the tree model is that it can only be used to investigate the relationships between languages that have a proven common ancestor. While it is well applicable to the established language families, it cannot be used to describe relationships above the language family level. According to Glottolog (Hammarström et al. 2020), there are 245 language families and 182 isolates (i.e. languages, for which no genealogical relationship with other languages is attested)[1]. There have been attempts to group some of the numerous language families together, but they rarely received a broad recognition among historical linguists. One such attempt is the Altaic (or Transeurasian) hypothesis. It suggests that the languages of five language families (Turkic, Mongolic, Tungusic, Koreanic, Japonic) spoken in Northern Eurasia are related. The term *Altaic* most commonly includes Turkic, Tungusic and Mongolic languages, more rarely also Japonic and Koreanic. There are also variations of the term, such as "Core Altaic", or "Micro-Altaic", which includes Turkic, Tungusic and Mongolic languages, and "Macro-Altaic", or "Transeurasian", which includes all five language families, Turkic, Tungusic, Mongolic, Japonic, Koreanic. While the term "Atlaic" originated in the 19th century (suggested by Matthias Alexander Castrén), the term "Transeurasian" is relatively new and was coined by Johanson and Robbeets (2010).

The typological similarities between the Altaic languages were noticed as early as in the 17th century and provoked deeper investigation of these languages, in particular the comparison of the lexical items. While it is not debated that the languages are similar typologically, the source of these similarities remains unclear. There are four main explanations for structural similarities between languages: common ancestor (or inheritance, vertical transmission, following the tree metaphor), language contact (or borrowing, horizontal transmission), chance similarity and language universals. We will

[1] According to Campbell (2013), there are around 420 distinct language families including language isolates. These counts can differ depending on what is considered to be a language or a dialect.

discuss each of these sources of similarities each in its turn.

Many structural features common in the Transeurasian languages, such as subject-object-verb word order, clause chaining, agglutination, suffixing morphology, vowel harmony, can be ascribed to language universals (Greenberg 1966, Nichols 1992): all these grammatical phenomena are found in other languages of the world (even if more commonly in Northern Eurasia) and are not exclusive of these five families. According to Nichols (1992), the only features connecting the Altaic languages left after taking language universals into account are personal pronouns. Pronouns in Altaic (or “Micro-Altaic”, i.e. Turkic, Mongolic, Tungusic) have the so-called M-T system, e.g. *men* 1SG - *sen* 2SG in Crimean Tatar, Turkmen, Kara-Kalpak (for further information, see Schwarz et al. 2020), but this pattern is also common in other languages spoken in Eurasia, e.g. in Uralic (*minä-sinä* in Finnish).

Not only universals explain the structural similarities between the languages in question, but also parallel historical changes, or *homoplasy.* The number of states a structural feature can take is limited, e.g. there are only two possibilities for the order of subject and verb, it is therefore difficult to exclude chance resemblances. Even in vocabulary, where there are more possibilities to combine sounds into words, chance plays a significant role and needs to be excluded before a deep family relationship is proposed. Random pairs of seemingly related words (because of the equal form and meaning) are not sufficient to prove the relatedness of the languages (Guy 1995, Ringe 1999, Campbell 2003).

The most debated sources of similarities are inheritance and borrowing. It is often the case (and Altaic/Transeurasian languages are no exception) that related languages are also spoken in the same area, which makes it even more difficult to distinguish between an areal spread of a structural feature (horizontal transfer) and inheritance from a common ancestor (vertical transfer).

The Altaic scholarship has numerous supporters on both sides of the debate. Some of the proponents of a genealogical explanation for the similarities between the Altaic languages were/are Nicolas Poppe, Gustav John Ramstedt, Roy A. Miller, Samuel Martin, Sergej and George Starostin and Martine Robbeets. The most enthusiastic opponents were/are Gerhard Doerfer, Gerard Clauson, Alexander Vovin and Stefan Georg. The most important milestones in the recent phase of the Altaic debate are the publication of the Etymological dictionary of the Altaic languages (Starostin et al. 2003, followed by a response by Vovin 2005, which was in its turn followed by a response from the editors of the dictionary, Dybo and Starostin 2008), the manuscript on the relatedness of Japonic and Koreanic with the Altaic languages (Robbeets 2005) and the connection of the descent of the hypothesised Transeurasian speakers with the spread of agriculture in Northeast China (Robbeets et al. 2021, followed, almost by tradition, by an opposing paper, Tian et al. 2022).

As of now, "Transeurasian" as a genealogical grouping is not widely accepted in the linguistic community. This can have to do with the nature of the languages (too scarce paradigmatic morphology, which is highly valued in testing genealogical relationships, based on the experience on Indo-European – scholars tend to apply the same methods as used for Indo-European to other language families and often fail to do so), history of languages (too few members survived, e.g. Tungusic, Koreanic, too shallow in terms of time depth, e.g. Mongolic), or other factors. If we consider structural features as an additional source of information, we face the scarcity of data: the grammars of many Transeurasian languages are not properly described, some of the languages became dormant even before scholars had a chance to document them (this is especially a case for Tungusic languages, some of which are either severely endangered or already dormant). Thus, we cannot adduce grammatical structure in its full potential as evidence for or against the relatedness of these languages, even if we decided to give structural features a chance.

Not only the genealogical relationship among the five language families is disputable – there is no consensus on the internal structure of the Transeurasian unity among the scholars, who support the hypothesis of a common ancestor of the Transeurasian languages. Scholars working on these languages have applied different methods and came to different conclusions on the internal groupings (see Robbeets 2020 for a detailed discussion of the possible classifications of the Transeurasian unity). Baskakov (1981) and Starostin et al. (2003) suggest a three-branch structure, grouping Mongolic with Turkic and Japonic with Koreanic, with Tungusic as a separate branch. Miller (1971) suggests a tree with consequent splits, where Turkic branches off first, followed by Mongolic and later Tungusic language families; Japanese, Ryukyuan and Koreanic constitute a separate three-way branch. The main uncertainty concentrates around the position of the Tungusic family: there are suggestions to group it either with Mongolic and Turkic (more often with Mongolic alone) or with Koreanic and Japonic.

Throughout the thesis, I refer to the geographically adjacent languages of the five language families spoken predominantly in Northern Asia by the term "Transeurasian", without any further implications of a genealogical relationship between these languages. The Transeurasian languages build a perfect basis for the investigation of the stability of structural features: the relationships between the languages inside the 5 language families are well understood in most cases and the languages have been in contact throughout history despite the vast spread all over Eurasia. Therefore, if (part of) structural features can be shown to have a phylogenetic signal and evolve at a slow rate, this would be *despite* language contact. The question of the relatedness of the five language families plays only a marginal role in the current thesis (see Chapter 2 for a tree of these languages based on structural features). Therefore, there will be no deep discourse into the history of the debate, but a description of a typological profile of the languages (Chapter 2), pointing out the synchronic similarities and divergences in the structures across these languages.

1.4 Language sample and data availability

The language sample comprises 60 languages belonging to 5 language families: 12 Japonic, 2 Koreanic, 14 Mongolic, 11 Tungusic, 21 Turkic languages (see Table 1.1 for the genealogical classification and Chapter 3, Figure 1 for the geographical distribution of the languages). For these languages, an extensive language grammar or a grammar sketch were available.

Each language was coded for 224 features, out of which 189 are from the Grambank database (Hammarström et al. 2017), 6 non-binary Grambank features were binarised and 35 features were added for their relevance and variability in the region (8 out of these 35 features are based on the feature set from Robbeets 2017). Out of 224 features, 53 features were absent in all languages, leaving the sample of 171 features with some variation across the languages of the sample. More than half of the languages could be coded for 95% of, or 162, features, around two-thirds of the languages could be coded for more than 78% of, or 134, features. The data was converted to a standardised format, the cross-linguistic data format (Forkel et al. 2018), and has been made public (*https://doi.org/10.5281/zenodo.7135936*). The cross-linguistic data format allows linguists to share data more easily and answer questions, among others, on linguistic diversity at a more global scale using extensive data sets, all in the same format.

The language sample is motivated both by genealogical representation and by convenience (i.e. availability of materials). Some language families were sampled more exhaustively than others, since there were no more materials on the languages than already included. From the already small Koreanic language family, only Modern Korean and Middle Korean were coded, due to unavailability of resources on other languages, in particular Jejueo, at the stage of data collection. The state of documentation and availability of materials is similar in Tungusic languages, where the data is scarce (often, there is only a word list available) and the grammars short. For most

Table 1.1: Language sample: Classification from Glottolog (Hammarström et al. 2020), according to Johanson (2020) (Turkic); Whaley and Oskolskaya (2020) (Tungusic); Heinrich et al. (2015) (Japonic). For space reasons, the branch names "Common Turkic" and "Ryukyan" were dropped and names for subbranches used directly instead

Family	Branch/Subbranch	Name
Japonic	Japanesic	Eastern Old Japanese, Japanese
	Northern Ryukyu	Okinoerabu, Shuri, Tsuken, Ura, Yuwan
	Southern Ryukyu	Hateruma, Ikema, Ogami, Tarama, Yonaguni
Koreanic		Korean, Middle Korean
Mongolic	Eastern Mongolic	Buriat, Kalmyk, Khalkha, Oirat, Ordos
		Khamnigan Mongol
	Middle Mongol	Middle Mongol
	Moghol	Moghol
	Southern Periphery Mongolic	Bonan (=Bao'an), Dongxiang
		Mangghuer, Mongghul
		Shira Yughur
Tungusic	Central Tungusic	Nanai, Oroch, Orok, Udihe, Ulch
	Manchu-Jurchen	Manchu
	Northern Tungusic	Beryozovka Even, Moma Even, Evenki, Negidal, Solon
Turkic	Bolgar	Chuvash
	Kipchak	Bashkir, Crimean Tatar, Kara-Kalpak, Kazakh, Nogai, Tatar
	North Siberian Turkic	Dolgan, Yakut (=Sakha)
	Oghuz	Azerbaijani, Gagauz, Turkish, Turkmen
	Khalaj	Khalaj
	South Siberian Turkic	Khakas, Shor, Tuvan
	Turkestan Turkic	Chagatai, Northern Uzbek, Old Turkic, Uighur

Japonic languages, there was only a grammar sketch and a short text available (mostly from Shimoji and Pellard 2010 and Heinrich et al. 2015). The Turkic language family has most languages and the materials available on them are the most abundant, both qualitatively and quantitatively; therefore, not all Turkic languages with available materials were coded, but only selected ones, which already make Turkic languages look "over-represented" in the language sample. The basis for the Mongolic languages constitutes the collection of grammar sketches in Janhunen (2003).

1.5 Phylogenetic comparative methods

Similarities between the evolution of species and the evolution of languages were noticed long ago (Darwin 1871), and a real break-through in their utilisation is noticeable in the last decades (Atkinson and Gray 2005, Bowern and Evans 2015). New quantitative methods have been applied to various language data with increasing frequency to investigate the relationships within and between language families (Gray et al. 2009, Grollemund et al. 2015, Kolipakam et al. 2018, Robbeets and Bouckaert 2018) and to test the evolutionary dynamics of linguistic data (Verkerk 2014, Greenhill et al. 2017, Carling and Cathcart 2021, Phillips and Bowern 2022). The main advantage of these methods lies in the quantification of abstract concepts of change and stability (Bowern and Evans 2015: 8).

While linguists most commonly resort to the comparative method to reconstruct the ancestral state of languages, biologists have come to favour computational methods to reconstruct the ancestral state of species. Ancestral state reconstruction (ASR) is often used in evolutionary biology as a way of inferring a character state of organisms that lived thousands or even millions years ago. In linguistics, phylogenetic comparative methods have been used for reconstructing higher numerals (Calude and Verkerk 2016), the morphosyntax of Proto-Indo-European (Carling and Cathcart 2021) and the origin of ergative alignment in Pama-Nyungan (Phillips and Bowern 2022).

It might be tempting to assume that phylogenetic comparative methods are redundant for ASR, because the most frequent feature value on a branch seems most likely to be the ancestral value for the particular node. An advantage of using a phylogenetic model to reconstruct ancestral states lies in differentiated weighting of probabilities and limiting the influence of the majority-rules principle. For example, if a feature is present in many modern languages, but is absent in an older language, it is not reconstructed as "present" at the proto-language level only because of its high frequency. On the other hand, a feature absent in the archaic languages can be reconstructed as "present" in the proto-language based on its "behavior" on other branches of the tree, e.g., frequent loss on some other branches (Carling and Cathcart 2021: 24).

ASR is a standard tool used to estimate the values of a trait (in our case "feature") for an internal node of a tree (in our case language subgroupings). ASR allows us to describe the past and the evolution of features: what did a language in question most probably look like? Did it have a particular feature or rather not? Since the Comparative Method cannot be applied to structural data and since structural data is coded in a way comparable to data in genetics, we can apply these methods to language data without significant adaptations.

ASR requires the feature values for each of the languages, a topology, branch lengths (optional) and a model of feature evolution as input (Pagel 1999, Ronquist 2004, Litsios and Salamin 2012). The quality of a reconstruction relies on the quality of the underlying phylogenetic tree. An increasingly common solution to this problem is evaluating the reconstruction along multiple trees that arise from Bayesian tree building. This approach allows us to estimate not only the ancestral state, but also the uncertainty around its reconstruction.

Since there is no consensus on the relatedness of the languages in the sample, I take individual topologies of the five language families to reconstruct ancestral states of the features for each proto-language. To get these topologies (a data-free "pseudo-posterior"), I use BEAST (Bouckaert et al. 2014) and a classification from glottolog to fix the groupings of languages (see Chapter 3 for a detailed description of the method and the results). This way, I can not only reconstruct an ancestral state for a node, but also measure the (un)certainty in these reconstructions.

A reconstruction always takes a value between 0 (meaning the feature was absent in the proto-language) and 1 (meaning the feature was present in the proto-language). These results are interesting from the perspective of Altaic/Transeurasian linguistics: the scholars interested in the Altaic/ Transeurasian question can investigate the reconstructions more closely and compare them across the five language families to see if they coincide in several families. In Chapter 3 of this thesis, I do not take the comparisons beyond pairwise matches between language families.

Ancestral state reconstruction goes hand in hand with evolutionary rate and phylogenetic signal: the slower a feature evolves and the higher the phylogenetic signal, the more accurate will the ASR be for that feature. The evolutionary rate measures the tempo, with which features evolve over time. The slower the feature evolves, the more stable it is and thus the more "useful" for the inference of deep linguistic past. Since there are no strong grounds to believe that features are lost and gained at the same rate (or rather there are strong grounds to believe the opposite, see Section 1.2), I measure both the rate of gain and the rate of loss and compare these separately with the phylogenetic signal.

Phylogenetic signal describes how much a feature value of one language depends on the feature value of another language due to the relatedness of these languages (Revell et al. 2008). In Chapter 3 of this thesis, I use metric D (Fritz and Purvis 2010) to measure the phylogenetic signal in structural features. It takes the sum of sister-clade differences in values across the tree. If related languages share the same feature value, the sum of the differences (the D value) will be low and the feature will thus have a high phylogenetic signal. A high D value thus corresponds to a low phylogenetic signal and a low D value to a high phylogenetic signal.

1.6 Trees, waves and admixture models

The relationships between languages can be represented in different ways: as a simple binary tree, a dated Bayesian tree, a Network, a NeighbourNet, an admixture plot, with the list to be extended. The representation has implications for our understanding of the relationships between the languages. Different types of phylogenetic analysis can produce different outputs and thus lead to different hypotheses about the diversification of languages (Heggarty et al. 2010). The two models used most frequently to describe divergence within a language family are 1) a splits model, i.e. a branching family tree structure, and 2) a wave model (Schmidt 1872, Wolfram and Schilling-Estes 2003), i.e. a dialect continuum. These two models correspond to two scenarios: 1) a group of language speakers splits into two, and these groups do not maintain contact, 2) a group of language speakers expands over a territory, and the speakers remain in some kind of contact. Even though the first scenario is very rare and almost not realistic, the tree model, which captures it well, is sometimes a useful simplification of language history, which allows to classify languages and show the degree of relatedness between languages. However, in most cases, it is neither sufficient nor accurate in describing the history of a language family. Most often, neither of the two models alone can describe the relationships between languages properly, because both horizontal and vertical processes are inherent to language evolution.

Depending on the history of the relevant language family and the surviving languages, the nature of any given language family can be either more tree-like or more wave-like. Network-type methods can be used to measure the “tree-(like)ness” or “net-ness” of a particular dataset (Huson and Bryant

2010). In tree-type methods, consistency indexes and retention indexes can be used as a measure to quantify the tree-likeness of the dataset. In Bayesian tree-building methods, the posterior probability indicates the support of a particular branch (Heggarty et al. 2010, Gray et al. 2010).

The composition of the individual language families comprising the Transeurasian unity differs considerably, which leads to different expectations as to the degree of tree-likeness of each of the families. Turkic languages consist of the two main branches, Bulgharic and Common Turkic languages. The first branch is represented by its only survivor, Chuvash, and the second branch comprises all the other languages, most of which are very similar structurally. Mongolic languages as we know them today developed from a single Mongolic language starting from approx. 13th century, after most Mongolic languages were wiped out by Genghis Khan. Therefore, the boundaries between languages and dialects are rather fluent and the language/dialect status is often unclear – different scholars describe them either as dialects or as languages (see the controversy around Kalmyk-Oirat-Darkhat, represented as a single language on Glottolog). There is no uncontroversial classification of the Mongolic languages – the best proxy remains the geographical classification. The history of Japonic languages is not quite unequivocal[2], but the language family clearly consists of two main branches: Japanese (and its dialects) and the Ryukyuan languages. The status of the Ryukyuan languages is also debatable: some sources describe these as dialects, while others ascribe them the language status. As for Tungusic languages, we have merely several survivors of the family, and most of these remaining languages are already endangered.

These diverse language family histories only reinforce the fact that neither of the two established models alone are satisfactory for an adequate description of the history of the Transeurasian languages. As for the more popular tree model, it is inappropriate and insufficient for studying the history of the Transeurasian languages with the data at hand for several reasons. First, it assumes the relatedness of the languages it is applied to, which, in our case, is highly questionable. Second, it assumes that language communities stop all contacts abruptly and diverge into separate languages from that point on, which is not the case for most Transeurasian languages. Third, the tree model performs especially poorly if applied to structural data because of undetected borrowing and chance similarities.

For these three reasons, and especially because the tree model is recognized as inadequate for the description of diffusion (Wolfram and Schilling-Estes 2003), we need a method that provides valid results if applied to data

[2] There is a hypothesis that they originated on the continent beside Koreanic languages and spread from there to the Japonic archipelago.

containing borrowings. A way forward is an admixture model implemented in the software STRUCTURE. It allows to identify genetically homogeneous groups of individuals (in our case languages) using a Bayesian approach (Pritchard et al. 2000). This method is the most widely used one in population genetics compared to other Bayesian clustering methods (Evanno et al. 2005).

In contrast to a phylogenetic tree, STRUCTURE is a clustering algorithm and does not assume language relatedness, which makes it usable in situations, where the relationships between languages have not been fully clarified. STRUCTURE tries to find homogeneous groups within the data in a range that has been provided by the researcher (for example, from 2 to 15, depending on what is feasible for the particular research question and for the particular amount of data). Each language is assigned with particular probability to one or more of the groups. For example, given 3 groups in the data, a language X can contain 70% of its ancestry from group 1, 10% from group 2 and 20% ancestry from group 3.

There are methods that are usually applied to choose the most probable number of clusters in the data (K), which are described in detail in Chapter 4. It is interesting both to investigate the population structure at the most probable K and at lower or higher K's. For example, if we assumed that there are only 2 clusters in the data (which might not be the most probable number of groups in the data), we might want to see how languages could be divided into these two clusters.

In linguistics, it has been applied to investigate deep language past surpassing the time limits of the comparative method (Reesink et al. 2009), to test hypotheses about putative language relationships (Bowern 2012) and to study variation among dialects (Syrjänen et al. 2016) and languages of a language family (Norvik et al. 2022). I use this model as a way of testing the performance of different sets of structural features in recovering the five language families and as a tool for comparing the level of admixture (or diffusion) in feature sets spanning across different language levels.

1.7 Aims of the thesis

Before I continue with the aims of the thesis, I would like to point out the aims that I do not pursue in this thesis because of the scope, the type of data used and the language sample. As was mentioned before, basic vocabulary and inflectional morphology are the most common data used to establish genealogical relationships. When using these, it is sufficient to have the data on the languages, for which we test the genealogical relationship. Should it be proven that structural features can be adduced for testing hypotheses about deep genealogical relationships between languages or even well-established language families, we would need languages that do not belong to the group in order to show that the languages in question are related. For example, were we to test the relatedness of the Transeurasian languages using structural features, we would need to include languages from other language families spoken in the area, such as Uralic, Yukaghiric, Chukotko-Kamchatkan, Nivkh, Ainu and probably others. Only if we can show that Turkic, Mongolic, Tungusic, Japonic and Koreanic cluster together more closely with each other than with other language families, can we suppose that there might be other links than merely geographical proximity and language contact with subsequent borrowing. Since my language sample contains only Transeurasian languages, I can investigate the relationships *among* the Transeurasian languages, but cannot test the unity as such. Therefore, testing the status of Transeurasian languages as a genealogical grouping is not one of the aims of this thesis.

This thesis pursues the following aims:

1) define the typological profile of the Transeurasian languages based on the information from grammatical descriptions of the respective languages,

2) test Bayesian tree building with structural features as data on the example of the Transeurasian unity (under the tentative presumption that these languages are related),

3) investigate the stability of structural features in terms of phylogenetic signal and evolutionary rate,

4) investigate the differences in stability among language levels, functional categories and parts of speech,

5) reconstruct ancestral states of structural features at the proto-language level for each of the five language families,

6) test the applicability of admixture model to structural data,

7) investigate the differences between language levels (phonology, morphology, syntax) in terms of amounts of admixture,

8) investigate the differences in the correct assignment of languages to language families with each of the feature sets, reduced to phonological, morphological and syntactic features respectively.

1.8 Overview of the chapters

The main part of this thesis is composed of three publications: Chapter 2 is published as a chapter in the Oxford Guide to the Transeurasian languages, Chapter 3 is published as an article in the Royal Society Open Science and Chapter 4 is accepted for publication with minor revisions as a research article in the Journal of Language Evolution.

Chapter 2 presents the typological profile of Transeurasian languages, supported by examples for each feature. In this chapter, I provisorily calculate the phylogenetic signal in structural features and compare it for real and simulated data. I use Bayesian tree-building methods (Bouckaert et al. 2014) to construct two phylogenetic trees of the Transeurasian languages: one based on the whole data set and one based on the set of stable features, determined as having a high phylogenetic signal ($D < 0.5$).

In Chapter 3, I calculate the phylogenetic signal in structural features using the metric D (Fritz and Purvis 2010) and the evolutionary rate of change (feature gain and feature loss) in structural features using the R package `caper` (Beaulieu et al. 2020). I calculate the correlation between the phylogenetic signal and the rate of loss and gain. Furthermore, I compare the stability of features across language levels, functional categories, and parts of speech and determine the most stable categories. I reconstruct the ancestral states of structural features at the language family level, i.e. for Proto-Turkic, Proto-Mongolic, Proto-Tungusic, Proto-Japonic and Proto-Koreanic, and compare the reconstructability across different categories.

Chapter 4 investigates the performance of structural features spanning over different language levels (phonology, morphology, syntax) in recovering language families. I apply an admixture model from population genetics (Pritchard et al. 2000) and obtain admixture profiles for the languages of the sample for the assumed number of populations (K) from 2 to 10. I compare the level of admixture and the precision of the assignment of languages to their respective language families at each of the levels and determine the language level with the least amount of admixture and the most precise genealogical classification of languages.

2. Typological profile of the Transeurasian languages

Author's contribution

This chapter has been published as a peer-reviewed chapter in the *Oxford Guide to the Transeurasian Languages* in June 2020. The manuscript was written by Nataliia Hübler under the supervision of Martine Robbeets, Eurasia3angle research group, Max Planck Institute for the Science of Human History, Jena. The data were collected and analyzed by Nataliia Hübler. The author would like to thank Alexander Savelyev, Sofia Oskolskaya and Orkhan Mehraliev for their help in data collection as well as Simon Greenhill for his advice on the phylogenetic comparative methods. The manuscript profited from valuable comments by Martine Robbeets and Ilya Gruntov.

The data, the scripts and other materials are available at: *https://doi.org/10.5281/zenodo.7244011.*

Typological profile of the Transeurasian languages from a quantitative perspective

Abstract

This chapter provides an overview of the typological features of the Transeurasian (Turkic, Mongolic, Tungusic, Japonic, Koreanic) languages, including brief descriptions of the phonology and morphosyntax of these languages. By applying phylogenetic comparative methods, I delimit a set of structural features with a high phylogenetic signal. These features can be assumed to be genealogically stable. I compare the trees achieved by Bayesian tree-sampling based on all 226 features and on the 97 structural features with a high phylogenetic signal and come to the conclusion that the data set with presumably stable structural features does not provide a tree that is compatible with the language history assumed by classical historical linguists. Neither full nor reduced feature set provides a reliable internal classification of the Turkic, Mongolic, Tungusic and Japonic language families.

Keywords: Transeurasian languages, typological features, phylogenetic signal, Bayesian tree-sampling

1 Introduction

It is common knowledge that most languages of Northeast Asia exhibit similarities in their structure, among them verb-final word order, strong head-marking, agglutinative suffixing morphology, lack of gender distinctions. The main discussion concerns the question whether all these similarities can be attributed to areal dispersal or whether some are residue of inheritance from a proto-language.

Although there is still no full consensus on the status of the Transeurasian unity as a Sprachbund or a language family, the genealogical relatedness of the Transeurasian languages is gradually gaining acceptance in the literature. See Robbeets (2020d) for the view that the Transeurasian languages are related and Vajda (2020) for the view that Transeurasian languages represent an area of diffusion. Moreover, scholars who agree on the relatedness of

Transeurasian languages suggest different topologies for the Transeurasian macrofamily (see Robbeets 2020b).

Classical comparative linguists rely on basic vocabulary and cognate grammatical morphemes when postulating language relationships. There are basic vocabulary (Robbeets 2020a) and cognate grammatical morphemes (Robbeets 2020c) in support of Transeurasian genealogical affiliation. Among the reasons why historical linguists do not wish to take abstract grammatical features into account are the following. First, structural features are more prone to borrowing than basic vocabulary or form-function matches in morphology. Second, the number of states structural features take (namely two: absent or present) facilitates convergent evolution (Heggarty 2006: 187, 193, Greenhill et al. 2017: 5). Third, the possible functional dependencies between features may lead to non-informative branch lengths (Heggarty 2006: 186). Fourth, a high rate of change leads to frequent switches between the states and the impossibility of predicting the states for the latest common ancestors (Greenhill et al. 2017). The answers to the questions such "Do structural features change faster than basic vocabulary?" and "How easily are structural features borrowed?" differ drastically. Some scholars state that structural features contain a deeper phylogenetic signal than basic vocabulary (Dunn et al. 2005), others add that it is impossible to disentangle genealogical signal from the one coming from ancient contact events (Wichmann and Holman 2009: 221) or that a group of features cannot define a genealogical unit (Reesink et al. 2009: 8).

In this chapter, I will not use structural evidence to establish language relatedness, but examine whether a set of stable structural features can replicate a topology of the individual Transeurasian families based on basic vocabulary and phonological correspondences, and compare the performance of structural features in providing tree structures that represent true language relations to that of basic vocabulary (as in Savelyev 2020 and Whaley and Oskolskaya 2020).

Robbeets (2020d) delimits a core of structural features that are shared by the Transeurasian languages and seem to be more easily explainable by inheritance than by borrowing. My approach is different from that of Robbeets, as I apply Bayesian inference to reach the topology of the Transeurasian languages and calculate the phylogenetic signal in the structural features along the topology in Fig. 4. Bayesian inference and phylogenetic comparative methods have not yet been applied to the structural features of the Transeurasian languages to find a historical signal in them and build the topology of the Transeurasian languages. Among the studies that applied Bayesian methods to structural data cross-linguistically, we find Dunn et al. (2008), Dediu and Levinson (2012), Reesink et al. (2009). Wichmann (2015) and Greenhill et al. (2017) concentrate on the rate of change of structural features.

The chapter is structured as follows. In Section 2 I present the language sample used for the typological description of the languages in question and the phylogenetic analysis in the following sections. Section 3 provides an overview of the typological similarities and differences between 38 Transeurasian languages. In Section 4 I apply phylogenetic comparative methods to delimit a set of structural features with a high phylogenetic signal and compare the topology of the Transeurasian languages based on all the features to the one based on the delimited set of structural features with a high phylogenetic signal. I summarize the findings in Section 5.

2 Data

The language sample is heterogeneous in terms of geography and genealogical affiliation. The sample covers 13 Turkic languages, 10 Tungusic, 5 Mongolic, 9 Japonic languages and Korean (see Table 1 and Figure 1).

In the description of the typological type of the Transeurasian languages that follows I will refer to the doculects[1] of the sample. Any generalizations about the overall presence or absence of a feature in a language family will only take into account the doculects mentioned in Table 1. Cases, where the presence of a feature is debatable or unknown, will also be excluded from generalisations.

[1] The information on a language in this chapter refers to a particular language as it is documented in the language description. The current state of the language can therefore deviate from the form described in the language grammar, which was used for this study. By referring to a particular "language" I thus mean a "doculect" if not noted otherwise.

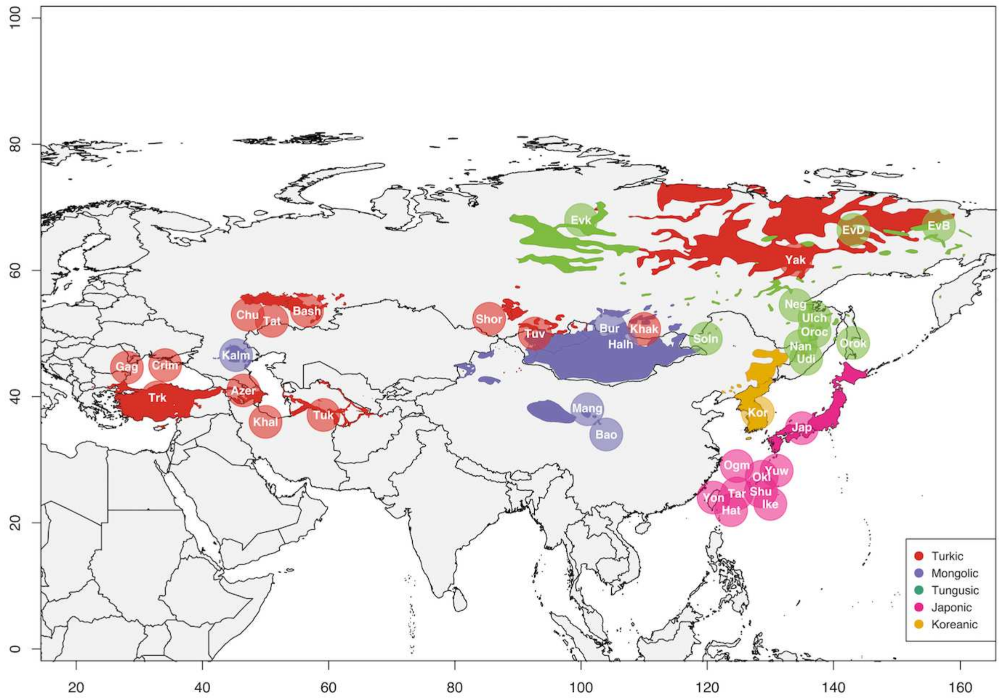

Figure 1: Geographical distributionm of the languages of the sample. Abbreviations: EvB = Even(Beryozovka dialect), EvD = Even (Dogdo-Chebogalahskiy dialect), Evk = Evenki, Nan = Nanai, Neg = Negidal, Oroc = Oroch, Orok = Orok, Udi = Udihe, Ulch = Ulch, Soln = Solon, Azer = Azerbaijani, Bash = Bashkir, Chu = Chuvash, Crim = Crimean Tatar, Gag = Gagauz, Khak = Khakas, Khal = Khalaj, Shor = Shor, Trk = Turkish, Tuv = Tuvan, Yak = Yakut, Tat = Tatar, Tuk = Turkmen, Jap = Japanese, Ogm = Ogami, Shu = Shuri, Tar = Tarama, Hat = Hateruma, Ike = Ikema, Oki = Okinoerabu, Yon = Yonaguni, Yuw = Yuwan, Bao = Bao'an, Halh = Khalha, Mang = Mangghuer, Kalm = Kalmyk, Bur = Buriat, Kor = Korean

Table 1: Language sample: Classification according to Johanson (2020) (Turkic); Whaley and Oskolskaya (2020) (Tungusic); Heinrich et al. (2015) (Japonic)

Turkic	Bulgharic: Chuvash
	Oghuzic: Turkmen, Azerbaijani, Gagauz, Turkish, Khalaj
	Siberian: Yakut, Tuvan, Khakas, Shor
	Kipchak: Crimean Tatar, Tatar, Bashkir
Mongolic	Khalkha, Kalmyk, Buriat, Mangghuer, Bao'an
Tungusic	Northern: Even (Beryozovka dialect), Even (Dogdo-Chebogalahskiy dialect), Evenki, Solon, Negidal
	Southern: Udihe, Oroch, Nanai, Ulch, Orok
Japonic	Northern Ryukyuan: Shuri, Yuwan, Okinoerabu
	Southern Ryukyuan: Ogami, Yonaguni, Hateruma, Tarama, Ikema
Koreanic	Modern Korean

3 Typological overview

3.1 Phonology

3.1.1 Vowels

Japonic (apart from Yonaguni), Tungusic, Mongolic languages as Buriat, Kalmyk, Khalkha, Siberian Turkic languages and Khalaj exhibit the vowel length distinction, e.g. Buriat (Mongolic, Poppe (1960: 6)): *tohon* 'fat, butter' - *to:hon* 'dust', *dara* 'press (imperative)' - *da:ra* 'freeze (imperative)' The most common type of vowel harmony synchronically is palatal harmony, which is present in all Turkic, some Mongolic, some Tungusic languages and Korean (1) (for a detailed discussion on the vowel harmony and beyond, see Joseph et al. 2020).

(1) Korean (Koreanic, Sohn 1999: 181)

a. *cwuk-essta*
die-PST
'died'

b. *nol-assta*
play-PST
'played'

Tungusic and some Mongolic languages also exhibit tongue root vowel harmony (2). For a broader discussion, see Janhunen (2012: 78–79), Svantesson (2020), Oskolskaya (2020), and Robbeets (2020d).

(2) Even (Tungusic, Kim 2011: 40)

a. *nɔŋan-dʊ*
3SG-DAT
'to him/her'

b. *min-du*
1SG-DAT
'to me'

3.1.2 Positional constraints

Initial velar nasals are not permitted word-initially in Transeurasian languages, apart from most Tungusic languages (except for Solon) and Bao'an. Initial trill /r/ in native words is restricted to Bao'an and Mangghuer. Initial consonant clusters are only permitted in some Japonic and Mongolic languages, and even if so, then most commonly the second consonant is a glide.

3.1.3 Phoneme inventories

Two separate liquid phonemes are present in all Mongolic and Turkic languages as well as in some Tungusic languages. They are absent in Japonic, Korean and Negidal, Orok, Oroch and Udihe. Typical of Tungusic languages is presence of voicing distinctions in stops, but not in fricatives (apart from Oroch, where there is a voicing distinction in dorsal fricatives). Most Turkic languages have both distinctions, apart from Chuvash (has none) and Yakut (has no voicing distinction in fricatives). Among Mongolic languages, at least Manghhuer is a special case with the voicing distinctions neither in plosives nor in fricatives, whereas most Mongolic languages have this distinction in plosives. Japonic languages mostly exhibit a voicing distinction in stops (apart from Ogami), but only some have it in fricatives. Korean has no voicing distinction in plosives or fricatives.

Transeurasian languages have two laryngeal contrasts for stops: voiced and voiceless. The only Transeurasian language exhibiting three laryngeal contrasts for stops (voiced, voiceless, aspirated) is Korean.

3.2 Agglutination and position of bound morphemes

Transeurasian languages in the sample are languages with agglutinative morphology, with the bound morphology being mostly suffixing.

3.3 Noun

In all Mongolic, Turkic languages and Korean nouns can be marked for plural. Among Tungusic languages, this holds for all Northern Tungusic languages, Nanai and Ulch. In Japonic languages nouns can be marked for plural, but this is mostly restricted to animate nouns. Southern Tungusic languages have a plural marker for animate nouns (apart from Ulch), Nanai has both a productive plural marker and a plural marker for kinship terms. The markers are typically regular, i.e. the plural form can be predicted from the singular form, with some phonological variation, e.g. plural formation in Yakut is accomplished by means of the suffix *-lar* and its allomorphs *-tar*, *-dar*, *-nar* (3). For lexicalization of plural markers, see Gruntov and Mazo (2020).

(3) Yakut (Turkic, Kharitonov 1982: 191)
at-tar
horse-PL
‘horses’

Transeurasian languages do not have any marking for any other number than plural, except for Bao’an, which has dual and paucal marking on nouns in addition to plural marking (4).

(4) Bao’an (Mongolic, Fried 2010: 68)
au=ʁala silaŋ=da o-tɕo
man=DU Xining=LOC go-IPFV.OBJ
‘The (two) men are going to Xining.’

The plural marker can have an associative meaning in Japonic, most Turkic languages (5) and Korean (*-tul*).

(5) Chuvash (Turkic, Krueger 1961: 94)
ivanov-zem
Ivanov-PL
'members of the Ivanov's family'

Some Mongolic (Bao'an, Mangghuer), Tungusic languages, Khalaj, Shor, Yakut and Korean have a special associative plural marker, e.g. compare the associative and the plural marker in Even (6a and 6c).

(6) Even (Tungusic, Lebedev 1978: 43–44)

a. *ami-ja*
father-ASSOC
'father and his relatives'

b. *orïr*
deer.SG
'a deer'

c. *orïl*
deer.PL
'deer'

Most Transeurasian languages have a pattern of derivation of action (7a), agent (7b) and object (7c) noun from a verb.

(7) Khalkha (Mongolic, Janhunen 2012: 97–98)

a. *saa-ly*
milk-NMLZ
'milking'

b. *bic-e:c*
write-NMLZ
'scribe'

c. *bic-ig*
write-NMLZ
'script'

Morphological core case (S, A, P argument) marking is common in most Transeurasian languages. Japonic languages and Korean mark grammatical relations by clitics. In this study, they are treated as morphological case marking, given their phonological boundness. Oblique arguments are marked either by a case suffix, by a postposition or by both.

In Transeurasian languages, noun reduplication serves the expression of a collective meaning (8a), plurality (8b) or distribution (8c).

(8) a. Kalmyk (Mongolic, Benzing 1985: 143)
ükr∼mükr
COW∼COLL
‘cows of different kinds’

b. Azerbaijani (Turkic, Shiraliev 1971: 43)
dästä∼dästä čičäk
bunch∼PL flower
‘bunches and bunches of flowers’

c. Korean (Koreanic, Sohn 1994: 386)
cip∼cip
house∼DISTR
‘every house’

Diminutive derivation (9) is productive across all Transeurasian languages with only a few exceptions and missing information for some languages.

(9) Shuri (Japonic, Shimoji 2012: 354)
taru:-gwa:
Taruu-DIM
‘a little Taruu’

The languages, where it is present, but is not a productive process, include Chuvash and Khalkha. Augmentative derivation is only found in Northern Tungusic languages, as, e.g. in Negidal (10), and Yonaguni.

(10) Negidal (Tungusic, Tsintsius 1982: 21)
bəje-xa:ja:
human-AUG
‘a huge human’

3.4 Pronoun

Some Tungusic languages and Bao’an (among the Mongolic languages in the sample) exhibit an inclusive/exclusive distinction in the first person plural, e.g. Udihe (Tungusic, Girfanova 2002: 18): *minti* 1PL.INCL, *bu* 1PL.EXCL. This distinction is present in Buriat and Kalmyk diachronically only. There is no gender distinction in personal pronouns in all Transeurasian languages, apart from Japanese (Japonic, Hinds 1986: 239): *kare* 3SG.M, *kanojo* 3SG.F, which entered the Japanese language relatively late, in Middle Japanese, and increased in frequency after the 16th century under the influence of Dutch.

Possessive pronouns not formed by a regular process are not well-spread throughout Transeurasian languages. In most Tungusic languages oblique pronominal stems fulfill their function. In Mongolic languages they are usually formed from a stem, different both from nominative and oblique pronominal stem and a genitive marker: Mangghuer (Mongolic, Slater 2003: 83): *mu=ni* 1SG=GEN, *namei=du* 1SG=DAT, *bi* 1SG.NOM. There is no synchronically detectable pattern in the spread of possessive pronouns across the Turkic languages.

Special logophoric pronouns are not common in Transeurasian languages. For Ogami, Pellard (2009) reports the existence of a reflexive pronoun that is used to indicate that the subject of the subordinate clause is the same as the subject of the first clause. In the example in (11) the reflexive pronoun *naa* is used to indicate that the 3rd person reporting the speech refers to a group of people including himself, whereas the reflexive pronoun *tuu* cannot be used logophorically. In Bao'an (Fried 2010: 121), logophoric pronouns are not obligatory.

(11) Ogami (Japonic, Pellard 2009: 122)
kanu psta=a naa-ta ik-a-tεεn=ti aɯɾ-i=ɯ
DIST nobody=TOP REFL-PL go-IRR-ACOM=QUOT say-CONV=IPFV
'He says, they will not go.'

Most Transeurasian languages possess a phonologically independent reflexive pronoun. Reciprocal pronouns are only rarely mentioned in descriptive works. Tungusic, Mongolic, some Japonic and Turkic (apart from Yakut) languages all form the oblique pronominal stem with a dental nasal, e.g. Negidal (Oskolskaya p.c. 2017): *bi* 1SG.NOM, *min* 1SG.OBL; Buriat (Poppe 1960: 50): *bi* 1SG.NOM, *mini:* 1SG.GEN; Turkish (Kornfilt 1997: 281): *o* 3SG.NOM, *on-a* 3SG-DAT (see also Schwarz et al. 2020). This is not the case in most Japonic languages and Korean. In northern Ryukyuan dialects, the first person pronoun uses *wa:-* as the nominative and genitive base and extended *waN-* in the oblique cases (Robbeets 2020d).

3.5 Demonstrative

Demonstratives in Mongolic and Tungusic languages have a two-way distance contrast. Japonic (Japanese, Shuri), Turkic languages (Chuvash, Shor, Turkish and Yakut) and Korean possess three demonstratives expressing a three-way distance relationship, e.g. Japanese (Hinds 1986: 232): *kono* ‘this’, *sono* ‘that’, *ano* ‘that over there’. Invisibility seems to be an accompanying meaning of the distal demonstrative in some Turkic languages. The only demonstrative with the dedicated function of expressing invisibility is present in Bao’an in the sample: *ənə* ‘this’, *nokə* ‘that’, *thər* ‘that out of sight’ (Fried 2010: 143). In some Tungusic languages demonstratives agree with the noun in number (12a), in Mongolic languages this is only the case in Buriat (see example 12c for Buriat in contrast to 12b for Kalmyk), although it had been a standard agreement in Middle Mongolian (Orlovskaya 1999: 27).

(12) a. Evenki (Nedjalkov 1997: 294
tari-l-va *beje-l-ve*
that-PL-ACC.DEF man-PL-ACC.DEF
‘those people’

b. Kalmyk (Benzing 1985: 133)
ter *ger-müd*
this house-PL
‘these houses’

c. Buriat (Poppe 1960: 110)
te-de *gern-ü:d*
that-PL house-PL
‘those houses’

All Mongolic, some Tungusic, Siberian Turkic languages and Gagauz possess a verb for content interrogation (meaning ‘do what?’). Japanese (Japonic, Hinds 1986: 29) has a compound *do:-si-te* how-do-PTCP ‘why’.

3.6 Article

Nouns are not obligatorily modified by definite articles in the whole area. Indefinite articles are optional in some Turkic languages (Khalaj, Khakas, Crimean Tatar, Turkish, Gagauz, Turkmen), Mangghuer, Bao’an and Oroch. Their position varies though: in Turkic languages there are only indefinite prenominal articles (13b), Manghhuer (13a, indefinite), Bao'an (indefinite) and Udihe (definite) have only postnominal articles.

(13) a. Mangghuer (Mongolic, Slater 2003: 99)
shuguo beghe ge
big tree SG:INDEF
'a big tree'

b. Khalaj (Turkic, Doerfer 1988: 94)
bi: ki-ni:
one day-ACC
'on one day'

3.7 Adjective

In Korean and most Japonic languages adjectives can receive the same marking as verbs used both predicatively (14a) and attributively (14b), in Turkic languages adjectives in predicative position can receive the same marking as verbs (14c).

(14) a. Japanese (Japonic, Hinds 1986: 345)
ano eiga=wa omosiroka-tta
that movie=TOP interesting-PST
'That movie was interesting.'

b. Japanese (Japonic, Hinds 1986: 346)
omosiroka-tta eiga
interesting-PST movie
'an interesting movie'

c. Turkish (Turkic, Kornfilt 1997: 83)
termiz-di-m
clean-PST-1SG
'I was clean.'

Reduplication of adjectives is a common process in Transeurasian languages; mostly it expresses intensification of the quality (15).

(15) Yakut (Turkic, Kharitonov 1982: 156)
χap∼χara
black∼INT
'very black'

Adjectives normally do not agree with nouns in number, with the exception of some adjectives in Buriat (16), Even (at least Dogdo-Chebogalahskiy dialect) and Evenki. It was very common in Middle Mongolian, in Buriat it might, however, be either an archaism or the influence of Modern Russian (Gruntov p.c. 2018).

(16) Buriat (Mongolic, Sanzheev 1953: 137)
hain-ü:d mori-d
good-PL horse-PL
'good horses'

3.8 Numeral system

The only numeral system represented in Transeurasian languages in the sample is the decimal one.

3.9 Verb

3.9.1 Tense-aspect-mood-evidentiality marking

TAME marking is accomplished by means of suffixation. Most Transeurasian languages have present (or non-past, i.e. not dedicated to marking present tense) and past tense marking (17).

(17) a. Bashkir (Turkic, Yuldashev 1981: 273)
al-dï-m
take-PST.INDEF-1SG
'I took (it).'

b. Chuvash (Turkic, Andreev 1997: 485)
yurla-d-əp
sing-PRS-1SG
'I am singing.'

Japonic languages lack dedicated future tense marking, whereas Tungusic languages, Korean, Bao'an and some Turkic languages mark it. Some Japonic and Turkic languages possess a free-standing particle for marking mood, Ogami (Japonic) and Turkmen (Turkic) for marking aspect, Yakut, Crimean Tatar (Turkic) and Bao'an (Mongolic) for marking tense. Most Transeurasian languages have a morphological distinction between perfective/imperfective aspect and morphological marking of mood. The verb form in the 2nd person imperative mood is identical to the root of the verb in Mongolic and Turkic languages, whereas Japonic, Tungusic languages and Korean have a dedicated suffix marking imperative mood (18).

(18) Korean (Sohn 1999: 276)
mek-ela
eat-IMP
'Please eat.'

Evidentiality marking is moderately common in Turkic languages (Chuvash (19), Yakut, Khakas, Crimean Tatar, Tatar, Gagauz), in Japonic languages (Hateruma, Ogami, Okinoerabu, Yonaguni), in Mongolic languages (Mangghuer, Kalmyk) and Korean.

(19) Chuvash (Turkic, Savelyev p.c. 2017)
a. *və^{w}l kay-nə*
3SG go-EVID
'He went (apparently).'
b. *və^{w}l kay-rэ*
3SG go-PST.3SG
'He went.'

3.9.2 Valency-changing operations

The only valency-increasing strategy across Transeurasian is causativization, which is accomplished by means of suffixation exclusively. As for other strategies of adding arguments to a verb, some Transeurasian languages possess locative markers. Ogami has a "purposive" converb marked by *-ka*, which introduces an argument for a goal of motion. A motion suffix is common in North Tungusic languages (Pakendorf and Aralova 2020: 300) and Oroch (Avrorin and Boldyrev 2001: 282). A morphologically marked passive voice (20) is available as a valency-decreasing strategy for all Transeurasian languages, excluding Chuvash, Nanai, Bao'an and Mangghuer.

(20) Shuri (Japonic, (Shimoji 2012: 376))
ari=nkai sugur-at-ta-n
3SG=DAT hit-PASS-PST-IND
'Someone was hit by her/him.'

Some Transeurasian languages (among them Korean and Even) use the same marker for passivization and causativization (21). However, as this marker became lexicalized (Robbeets 2007: 160), it is not a common isomorphism in modern Transeurasian languages.

(21) a. Korean (Koreanic, Sohn 1999: 367)
po-i-ta
see-CAUS/PASS-DECL
'be seen; show'

b. Even (Tungusic, Lebedev 1978: 84)
maa-v-daji
kill-CAUS/PASS-PTCP
'be killed'

c. Even (Tungusic, Lebedev 1978: 84)
i:-v-deji
enter-CAUS/PASS-PTCP
'carry in'

The agent in a passive construction is most often marked the same way as the recipient in a ditransitive construction, i.e. either as a dative case marker or as a dative particle (22).

(22) Okinoerabu (Japonic, van der Lubbe and Tokunaga 2015: 361–362)
Mariko=ga Taroo=ni ʔabi-ra-tta-mu
Mariko=NOM Taroo=DAT call-PASS-PST-IND
'Mariko was called by Taroo.'

Incorporation of nouns into verbs is not a common intransitivising strategy in Transeurasian languages. In all the Transeurasian languages antipassive marking is absent.

3.9.3 Verb morphology in subordinate clauses

Most Transeurasian languages use infinite verbal morphology to indicate subordinate clauses, with the verb marked for finiteness in the main clause, i.e. clause chaining, which is only in rare cases described as such. The converb strategy for marking the distinction between simultaneous and sequential clauses (23) is very common across all Transeurasian languages.

(23) Ulch (Tungusic, (Petrova 1936: 58))
buə ŋənə-məri jaja-ha-pu
1PL walk-SIM.PL sing-PST-1PL
'We were singing while we were walking.'

Among Transeurasian languages of the sample, three subgroups possess an existential verb that is different from the equative copula: most Turkic (Azerbaijani *var*), Japonic (Japanese *aru/iru*) languages and Korean (*issta*). In Mongolic and most Tungusic languages it appears to be identical with the copula (apart from the cases of missing data): Tungusic **bi-*, Mongolic **bu-*, **a-*.

3.9.4 Reduplication

Apart from Bao'an (24), which employs verb reduplication for expressing a continuous action, and Tuvan, where verb reduplication "indicates an extension of the action for a definite period of time" (Krueger 1997: 141), verb reduplication is not a common phenomenon in Transeurasian languages (note that only cases where reduplicated verbs constitute a single phonological word are taken into account).

(24) Bao'an (Mongolic, Fried 2010: 102)
atɕaŋ khəl∼khəl-tɕə
3SG speak∼CONT-PFV
'He talked and talked (for a long time).'

3.10 Attributive possession

In cases, where the possessor is marked on the possessed, pronominal possessors follow their heads (suffixes), nominal possessors precede the possessed across Transeurasian languages. In Turkic, Mongolic and Tungusic languages the possessor is indicated on the possessed by a suffix in attributive possession. Japonic, some Tungusic, most Mongolic and Turkic languages and Korean (25) indicate the possessor with a genitive marker, which can be either a clitic or a suffix.

(25) Korean (Koreanic, Sohn 1994: 174)
na=uy yenphil
1SG=GEN pencil
'my pencil'

In most Tungusic languages the possessor is unmarked (26).

(26) Even (Tungusic, Kim 2011: 62)
svinija ulrə-n
swine meat-3SG
'swine's meat, pork'

Only Tungusic languages and Chuvash (debatable, see Savelyev 2020 for Chuvash) have different marking for alienable and inalienable possession. Tungusic languages have special marking for alienable possession (-**ŋi*) in addition to the person of the possessor.

3.11 Predicative possession

Transeurasian languages show a variety of ways to express predicative possession: (i) with a transitive "habeo"-verb (some Japonic languages, e.g. 27a), (ii) with a locative-marked possessor (a common strategy in all subgroups of the Transeurasian unity, e.g. 27b), (iii) with a dative-marked possessor (available in Japonic, Mongolic, Tungusic languages, e.g. 27c), (iv) with a possessor coded as an adnominal possessor (Korean, Tungusic, Mongolic, Turkic languages, Yuwan, e.g. 27d), (v) with a possessor coded as a comitative argument (the least common strategy, available in Yakut, some Mongolic and Tungusic languages, e.g. 27e).

(27) a. Japanese (Japonic, Hinds 1986: 138)
watasi=wa kuruma=o motte iru
1SG=TOP car=ACC possess.PTCP be
'I have a car.'

b. Korean (Koreanic, Sohn 1999: 284)
halapeci=kkey chayk=i manh-ayo
grandpa=LOC book=NOM many-POL
'Grandpa has many books.'

c. Evenki (Tungusic, Bulatova and Grenoble 1999: 9)
bəjətkə:n-du: kniga bisi-n
boy-DAT book be-3SG
'The boy has a book.'

d. Azerbaijani (Turkic, Mehraliev, p.c.)
män-im pišiy-ïm var
1SG-POSS.1SG cat-POSS.1SG exist
'I have a cat.'

e. Kalmyk (Mongolic, Benzing 1985: 56)
surhulc nain denš-tä
student eighty kopecks-COM
'The student has 80 kopecks.'

3.12 Alignment

All Transeurasian languages have accusative (S/A P) alignment of marking of core arguments (28).

(28) Japanese (Japonic, Ishizuka 2012: 3, 192)

a. *keisatu=ga ken=o tukamae-ta*
police=NOM Ken=ACC catch-PST
'The police caught Ken.'

b. *kondo=wa kiji=ga tonde-ki-mashi-ta*
next=TOP pheasant=NOM fly-come-POL-PST
'Next a pheasant came flying down [to them].'

Parallel to it, all Turkic and Mongolic languages also have neutral S/A/P alignment of marking due to their differentiation between definite and indefinite objects: indefinite objects do not receive accusative marking and are thus unmarked (the same way as the S/A arguments), e.g. (29).

(29) Shor (Turkic, Dyrenkova 1941: 59)
aŋčï kaŋdus aŋnapča
hunter otter hunt
'A hunter hunts otter.'

Korean allows omission of all case-marking particles. Some Tungusic languages also exhibit neutral marking, as for them the accusative marking is optional. In Udihe, it can be omitted i) for phonological reasons, ii) if the object is non-specific (30), iii) if the verb is in the imperative.

(30) Udihe (Tungusic, Nikolaeva and Tolskaya 2001: 123)
ipaNene-mi ogbö wa:-ni
go-INF elk kill.PST-3SG
'On the way, he killed an elk.'

The A/S argument is often indexed on the verb by a suffix across Transeurasian languages (31).

(31) Chuvash (Turkic, Andreev 1997: 484)
pïradə-p
go-1sg
'I go.'

There is variation in the alignment of marking the recipient of a ditransitive construction and the patient of a transitive verb. Chuvash, Mangghuer, Japanese and Korean allow the same marker for both constructions. Other Transeurasian languages employ different markers for these roles.

3.13 Negation

Most Transeurasian languages mark negation on the verb by means of a suffix (32).

(32) Bashkir (Turkic, Poppe 1964: 94)
min unï kyr-mä-ne-m
1sg 3sg see-neg-pst-1sg
'I didn't see him.'

Some Mongolic languages, such as Bao'an, Kalmyk and Mangghuer, do not have inflectional morphology for negation, they mark it by a particle instead. For one of the strategies in Kalmyk, see example (33).

(33) Kalmyk (Mongolic, Benzing 1985: 90)
es bosna:
neg stand.up.prs
'He doesn't stand up.'

All Tungusic and some Japonic languages (Tarama and Yonaguni) possess an auxiliary for marking standard negation (34).

(34) Evenki (Tungusic, Nedjalkov 1997: 96)
bejumimni homo:ty-va e-če-n va:-re
hunter bear-acc.def neg-pst-3sg kill-ptcp
'The hunter didn't kill the bear.'

It is possible to mark prohibitive and declarative negation (transitive declarative clauses) in the same way in Mongolic and some Turkic languages. Japanese uses the marker *-nai* for some types of declarative negation as well as for prohibitive negation (prohibitive also requires the infiniteness marker

-de), but in general different markers for both negation types are employed in Japonic languages and Korean. Most Transeurasian languages employ different negation markers for verbal vs. locative/existential/nominal negation (35), apart from Korean, Nanai and several Japonic languages.

(35) Turkish (Turkic, Kornfilt 1997: 123–125)

a. *hasan kitab-ï oku-ma-dï*
Hasan book-ACC read-NEG-PST
'Hasan didn't read the book.'

b. *ben hasta deyil-im*
1SG sick NEG.COP-1SG
'I am not sick.'

c. *ben ev-de yok-tu-m*
1SG home-LOC NEG.EXIST-PST-1SG
'I was not at home.'

3.14 Word order

In most cases in Transeurasian languages modifiers precede their heads, thus adjectives, numerals and demonstratives usually precede the noun (see Sections 3.5 and 3.7 for examples). In most Japonic and Mongolic languages a numeral can both precede and follow the noun. In Korean, the standard position for the numeral is the one after the noun. The modifier-head structure also holds for relative clauses: in all Transeurasian languages, apart from Azerbaijani and Khalaj, the relative clause precedes the noun it modifies. In simple pragmatically unmarked clauses the word order is verb-final both for transitive and intransitive clauses for Transeurasian languages. Clausal objects typically appear in the same position as nominal objects in Transeurasian languages, apart from some Turkic (Azerbaijani borrowed this construction from Persian) and Tungusic languages. The order of main arguments in transitive declarative clauses is rigid in some Mongolic and almost all Turkic languages (apart from Gagauz), whereas Tungusic, Japonic languages and Korean allow variation in the order of A and P, as long as these are appropriately marked for their function. Content interrogatives most often occur in situ in Transeurasian languages.

3.15 Interrogation

Marking interrogation by a clause-final question particle is the most common strategy across Transeurasian languages (36).

(36) Shor (Turkic, Dyrenkova 1941: 244)
ol taiga-da aŋ köp-pe
that taiga-LOC animal many-Q
'Are there many animals in taiga?'

A minor strategy is marking it by intonation, which is present in some Turkic, Tungusic and Japonic languages.

3.16 Comparative construction

Comparison is mostly accomplished by means of one kind of locative marking of the standard of comparison. The most common case used in this function is ablative (37).

(37) Azerbaijani (Turkic, Shiraliev 1971: 47)
bakï kirovabad-dan böyükiür
Baku Kirovabad-ABL big
'Baku is bigger than Kirovabad.'

All Japonic languages, Ulch, Nanai, Orok and Korean have a marker that has neither locational meaning nor the meaning 'surpass/exceed' (38).

(38) Japanese (Japonic, Kaiser et al. 2013: 42)
gyu:niku=ga butaniku yori yasui
beef=NOM pork COMP cheap
'Beef is cheaper than pork.'

The adjective in a comparative construction is unmarked in most Transeurasian languages or marked optionally, apart from a number of Turkic languages, both Even dialects and Evenki (39).

(39) Even (Lebedev 1978: 55)
bii hin-duk egʤe-tmïr
1SG 2SG-ABL high-COMP
'I am higher than you.'

3.17 Coordination and conjunction

Conjunction vs. coordination marking has internal discrepancies among Transeurasian languages. Nanai, Orok, Evenki, Buriat, Khalkha, Kipchak Turkic languages, Yakut, Khakas and Azerbaijani use different morphemes to express conjunction and comitative (e.g. 40).

(40) Evenki (Tungusic, Bulatova and Grenoble 1999: 12, 56)

a. *bi: əkin-nu:n-mi: təwlə:-m*
1SG sister-COM-REFL.SG collect.berries-1SG
'I went with my sister to pick berries.'

b. *bi: taduk girki-w ollo-mo:-čo:-wun*
1SG and.then friend-POSS.1SG fish-go-PST-1PL.EXCL
'My friend and I went fishing.'

3.18 Obligatoriness of S/A argument

Most Transeurasian languages allow omission of the S/A argument (41).

(41) Mangghuer (Mongolic, Slater 2003: 124)
ning ge khuba di ge-jiang
this do divide eat do-OBJ:PFV
'Like this (they) divided and ate (him).'

3.19 Derivation of adpositions

Adpositions are often derived from place nouns by locational suffixes (42).

(42) a. Buriat (Mongolic, Sanzheev 1962: 301–304)
bäe-hä:n
body-ABL
'from the side'

b. Evenki (Tungusic, Bulatova and Grenoble 1999: 13)
amut daga-la:-n
lake close-LOC-POSS.3SG
'closer to the lake'

3.20 Classifiers

Numeral classifiers are the only type of classifiers present in Transeurasian languages. These are common in Japonic languages, some Tungusic languages, such as Evenki, Ulch, Negidal (the latter probably under the influence of Ulch, Oskolskaya p.c. 2017), Nanai, Turkic (Crimean Tatar), Mongolic (Mangghuer) and Korean. In Evenki there are numeral classifiers differentiating human and non-human counted entities (43).

(43) Evenki (Tungusic, Nedjalkov 1997: 283)

a. *nadan-i:*
seven-CLF:HUM
'seven people (together)'

b. *nada-ngna*
seven-CLF:NONHUM
'seven objects (together)'

4 Phylogenetic analysis

4.1 Coding procedure

The current study encompasses a heterogeneous language sample consisting of 38 doculects and 226 binary structural features. I use 189 formulations of the features from the Grambank database (Hammarström et al. 2017), 10 binarised versions of Grambank features on word order and 27 features relevant for Transeurasian languages (partly from Robbeets 2017). I coded the features based on descriptive works, dictionaries and personal correspondence with language experts. The data set with the coding for each individual language for each feature as well as the description of the structure of the data set can be found in the online supplementary materials.

The feature set provides an extensive coverage of morphosyntax of the language (e.g. person and number marking on nouns, possessive constructions, interrogation, negation, derivation patterns, valency operations, numeral systems, comparison, argument marking, deixis) as well as phonology (voicing distinction in plosives and fricatives, l/r distinction, constraints on initial consonants, availability of initial consonant clusters, vowel harmony, vowel length).

The four main criteria for feature selection are: i) stability, ii) informativity, iii) codability, iv) logical independence. The first criterion is fulfilled by the preselection of the features for their being stable cross-linguistically in Grambank. The "Transeurasian" features are assumed to be stable by Robbeets (2020d). The second criterion foresees informativity of the features. The features, which are not part of Grambank, were added based on their variation in the language sample. This aims at resolving the internal relationships between the languages in question. The third criterion takes into account the coverage of the respective topic by reference grammar. In this way, languages with extensive descriptions available can be included as well as those with only grammar sketches. Grambank features have been preselected to meet this criterion. Some "Transeurasian" features were excluded a-posteriori due to the low coverage of the respective topics in the descriptive works. According to the fourth criterion, the value of one particular feature has to be independent of the value of another feature, i.e. neither triggered nor predicted by it.

4.2 Stability of structural features

I delimit structural features stable in Transeurasian languages by extracting the features with a high genealogical signal. To avoid circularity, I calculate the signal along the tree based on lexical data and phonological correspondences (see Figure 4). For each feature with moderate variation (149 features in total), I calculate the phylogenetic signal with the metric Fritz and Purvis' D using the function `phylo.d` from the package `caper` in R. This method takes into account the distribution of the feature values in sister branches: if sister languages have the same feature value, the D value will be low and thus the phylogenetic signal will be high.

To set a cut-off point for the stability of the features, I compare the distribution of the D values for the real data and the randomized data (see Figure 2). I set this point to the two standard deviations from the mean of the randomized data, i.e. 0.53.

Sixty-five percent of the features have a D value smaller than 0.53 and can thus be considered relatively stable. The impact of the language domain, which the feature covers, the genealogical attribution of the languages in question and the proportion of 0's and 1's for a particular feature can impact the estimated phylogenetic signal in the feature. We will have a closer look at the influence of the part of speech on the amount of the phylogenetic signal

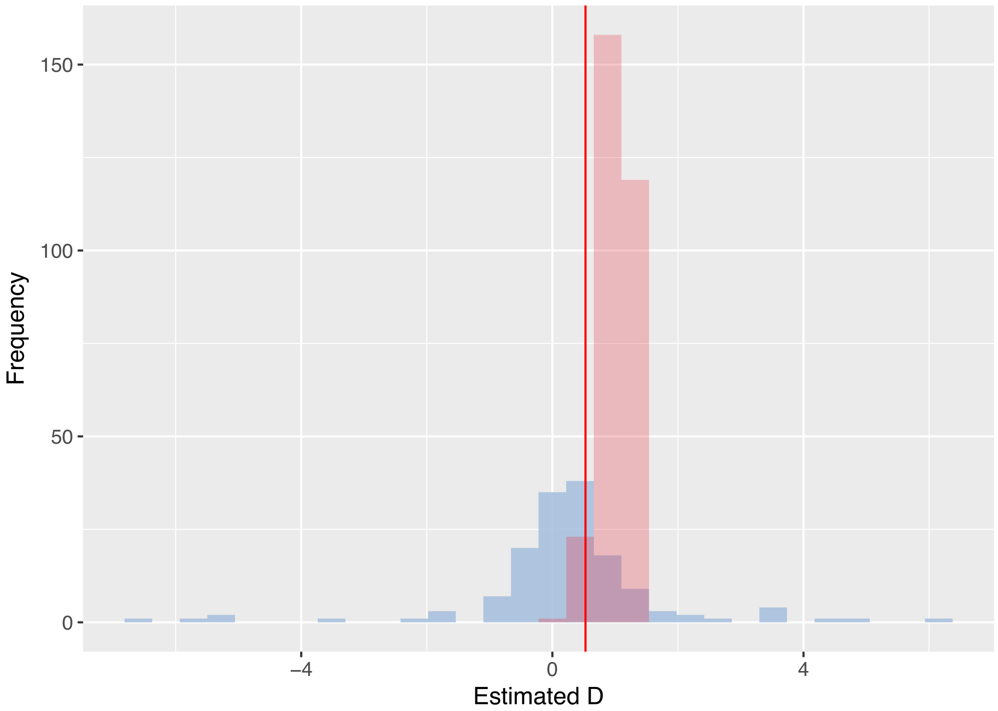

Figure 2: Estimated D values for real data (blue) and randomized data (red). The vertical line indicates the two standard deviations from the mean threshold.

in the feature and its interaction with other factors.

The phylogenetic signal differs across features covering different parts of speech (see Figure 3). The differences in the distribution of D values across parts of speech can be explained by at least two factors. First, the number of the features that correspond to a particular language domain differs (6 features on adjectives, 2 on articles, 3 on demonstratives, 23 on nouns, 2 on numerals, 17 on pronouns, 42 on verbs). Second, extremely low values of D are often due to high uniformity of features, e.g. 36 out of 38 languages have the same value for a particular feature and this leads to underestimation of D values. This is particularly the case for pronouns, where most of the features have the same value 0 for all the languages except one or two.

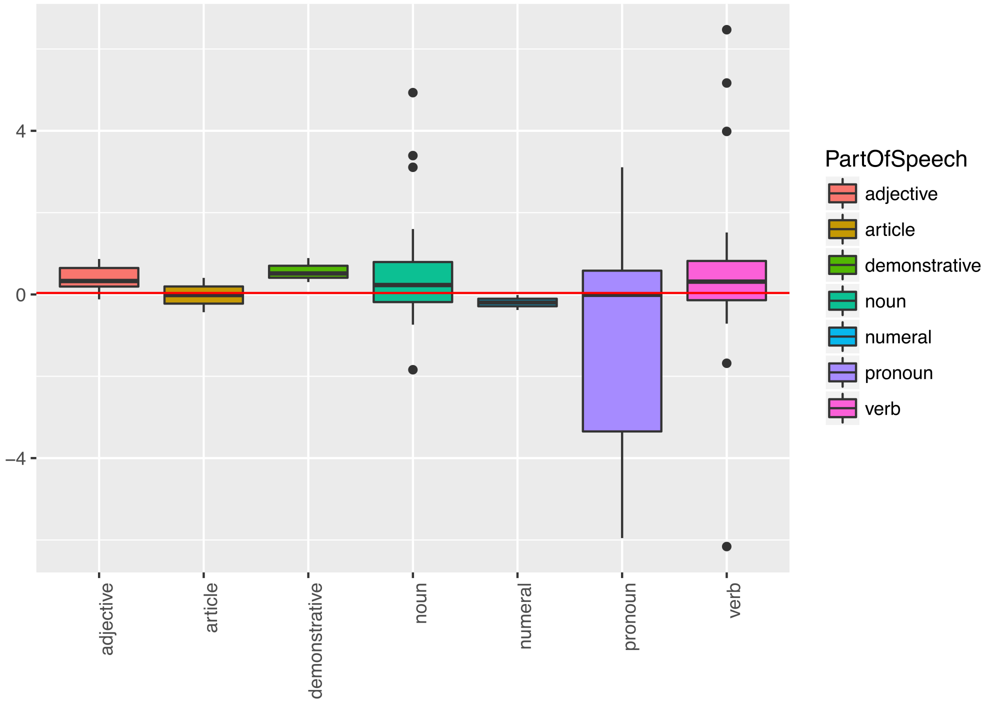

Figure 3: Estimated D values across parts of speech.

4.3 Bayesian approach to the classification of the Transeurasian languages

I use the whole data set and the data set with only the stable features delimited according to the procedure described further in Section 4.2 to build two topologies of the Transeurasian languages (compare Figures 5 and 6). The underlying Bayesian analysis derives a distribution of trees instead of a single tree. The more often a particular clade (a language grouping) appears in this distribution, the higher is the credibility of the clade and the lower the uncertainty within the clade.

The traditional affiliation of languages to the respective language families is replicated in the topology based on the whole data set, except for Yakut. As the Mongolo-Yakut branch is short and the posterior probability for the clade is low, Yakut must have split from the Turko-Mongolic ancestor at approximately the same time as Mongolic and Turkic split into two branches. The posterior probabilities for the individual language families are moder-

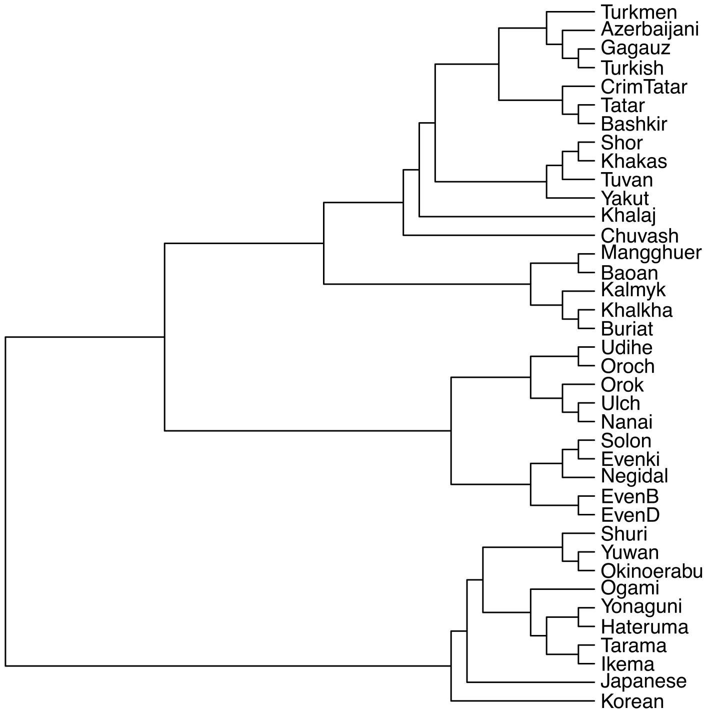

Figure 4: The tree used for the estimation of the phylogenetic signal

ately high: 1.00 for Koreano-Japonic, 0.83 for Altaic, 0.98 for Tungusic, 0.77 for Mongolo-Turkic, 0.7 for Mongolic and 0.71 for Turkic excluding Yakut.

The internal structure of each smaller-level language family is replicated to a different extent, which is reflected in the high uncertainty (i.e. low posterior probability estimates) in the clades. There might be several explanations for this. First, it may be a result of horizontal transmission, i.e. a high number of borrowing events between the languages. Second, the branches may be

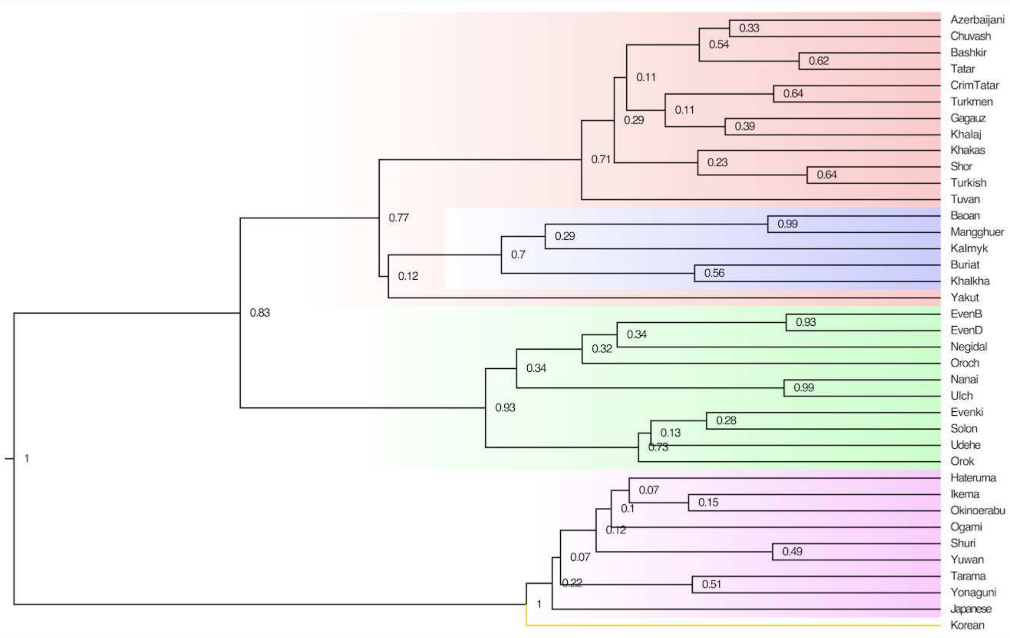

Figure 5: Topology of the Transeurasian languages based on the whole data set

so closely related that it is difficult for the algorithm to resolve them (Dunn et al. 2008). Both explanations are valid for some branches in the Transeurasian topology. For example, Yakut appears outside the Turkic cluster owing to its known history of contact with Tungusic languages. Turkic languages are structurally too similar to one another for the algorithm to reliably establish the individual groupings. The same explanation might also be valid for Japonic languages. On the positive side, the known close interrelatedness of Ulch and Nanai and two Even dialects respectively is replicated in the high posterior probabilities in the tree.

Some of these relationships are replicated if the features with the higher D values, i.e. the ones assumed to be unstable, are excluded. The main structure of the tree, Japono-Koreanic vs. Altaic branch, Tungusic languages splitting off first from the Proto-Altaic ancestor, remains intact. High posterior probabilities for Japono-Koreanic and Tungusic branches are also preserved. After the reduction of feature number to 97, Korean does not appear as a separate branch anymore, Yakut disrupts the structure of the

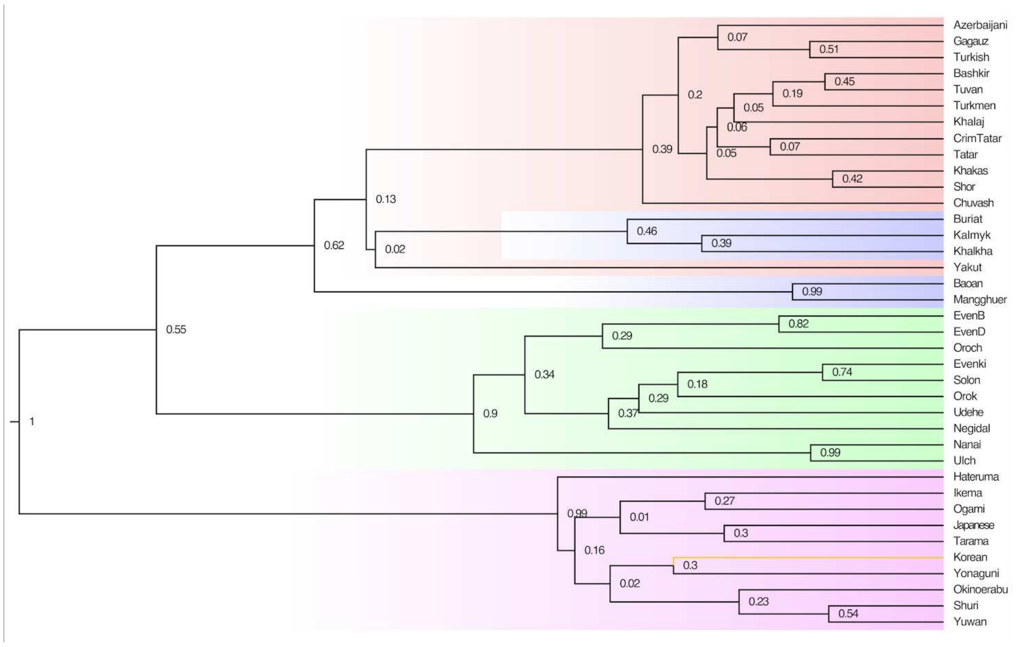

Figure 6: Topology of the Transeurasian languages based on stable structural features

Mongolic language family by appearing inside the Buriat-Khalkha-Kalmyk cluster and separating it from Bao'an and Mangghuer, posterior probabilities for Altaic, Turkic, and Mongolic branches drop. This result might be due to the following methodological restrictions. First, the metric functions reliably for language samples with 50 languages and more (Fritz and Purvis 2010: 1050). Second, a different strategy for the delimitation of the features could lead to a better topology.

The exclusion of "unstable" structural features does not improve the internal classification of the individual language families in terms of making it more similar to the classification based on lexical data and phonological correspondences as expected.

Comparing the Bayesian classifications of the individual language families (Savelyev 2020, Whaley and Oskolskaya 2020), I come to a conclusion that structural features perform worse in terms of confidently disentangling the internal structure of the lower-level branches, but well enough to replicate the affiliation of most languages within their respective language families, if the number of features is sufficiently high (or at least as high as the vocabulary lists used in this volume, i.e. around 200).

5 Conclusions

The chapter addressed the following research questions: What does a typological profile of the Transeurasian languages look like? Do structural features provide a reliable tree of the Transeurasian languages? Are there differences in structural features in terms of their phylogenetic signal? Do structural features with a high phylogenetic signal provide a “better” tree than the whole set? These questions aimed at filling the gap in the debate on the internal structure of the Transeurasian unity, on the suitability of structural features for building trees that represent true language history (ideally genealogical relationships between languages) and on the stability of structural features.

There is a consensus on the fact that the Transeurasian languages share structural similarities, but no quantitative approach has been applied to the structural data to address the issue of the exact interrelationships between these languages. The previous research suggested a genealogical relationship between Japonic and Koreanic languages and a genealogical grouping of Turkic, Mongolic and Tungusic languages. The results of the study show that the distribution of the structural features among Transeurasian languages supports a division between Altaic and the Japono-Koreanic unities. This split is also reflected in the results of the Bayesian analysis through the binary structure of the Transeurasian tree with the Altaic and the Japono-Koreanic branches. There is also a noteworthy tendency of Tungusic languages to follow either an Altaic or a Japono-Koreanic type in a number of features.

The features in Table 2 can be assumed to constitute the Transeurasian language type synchronically, according to the frequency of their occurrence across Transeurasian languages. Features that are common only in Altaic languages are still listed in Table 2, as they are frequent in 3 of 5 branches of the Transeurasian unity.

There is an ongoing debate on the stability of structural features. Despite the discrepancies in the results achieved in previous studies, most scholars agree upon the fact that there is at least a set of genealogically stable structural features. This study has measured the phylogenetic signal in the structural features of the Transeurasian languages by applying the metric

Table 2: Typological profile of the Transeurasian languages

Phonology	vowel length distinction; vowel harmony; no word-initial velar nasals and consonant clusters; two-fold division of the distribution of the distinction in liquids
Nominal morphosyntax	regular plural marking; associative plural marking; rich derivational morphology; morphological core case marking; nominal reduplication; oblique pronominal stem with a nasal; no agreement in number between the noun and adjective/demonstrative; no plural marking on the noun in numeral-noun phrases; GEN marking of the possessor; possessor indicated on the possessed by a suffix; accusative alignment of marking of main arguments; ABL case marking of the standard of comparison in a comparative construction; adpositions derived from place nouns marked with locative cases; NP word order: modifier-head;
Verbal morphosyntax	passivization and causativization by morphological means; clause chaining; morphological marking of negation; verb agreement with the S/A argument in person and number
Clause	SOV word order; pro-drop languages; clause-final particle for marking interrogation; LOC/DAT marking of the possessor (lit. 'The cat is on/to me.') or coding of the possessor as an adnominal possessor (lit. 'My cat exists.') in a predicative possession construction

Fritz and Purvis's D. The most commonly discussed range of the D values is between 0 (strong phylogenetic signal) and 1 (the feature is distributed randomly on the tree). The analysis of the stability of the Transeurasian structural features has shown that the features vary in terms of the phylogenetic signal and more than a half of the features with moderate variation have a high genealogical signal. The current study has thus provided a summary of the typological profile of the Transeurasian languages, suggested an internal structure of the Transeurasian unity based on the structural features and calculated the phylogenetic signal in the structural features. The internal structure of the Transeurasian unity achieved in this study goes in line with the proposal of Robbeets (2018) on the Transeurasian tree consisting of the Japono-Koreanic and Altaic branches, with Altaic splitting further into Mongolo-Turkic and Tungusic branches. The structural features with a high phylogenetic signal do not point to a tree of Transeurasian languages suggested by historical comparative linguists. In order to account for the source of the similarities between languages, a further study is needed, where the geographical location of languages (synchronically) and the nodes (diachronically) are controlled for. Neither the tree based on the whole set of structural features nor the tree based on the stable set of features provide a reliable structure of Turkic, Mongolic, Tungusic and Japonic language families.

Abbreviations

1	first person	DIST	distal
2	second person	DISTR	distributive
3	third person	DU	dual
ABL	ablative	EVID	evidential
ACC	accusative	EXCL	exclusive
ACOM	anticommissive	EXIST	existential
ASSOC	associative	F	feminine
AUG	augmentative	GEN	genitive
CAUS	causative	HUM	human
CLF	classifier	IMP	imperative
COLL	collective	INCL	inclusive
COM	comitative	IND	indicative
COMP	comparative	INDEF	indefinite
CONT	continuous	INF	infinitive
CONV	converb	INT	intensive
COP	copula	IPFV	imperfective
DAT	dative	IRR	irrealis
DECL	declarative	LOC	locative
DEF	definite	M	masculine
DIM	diminutive	NEG	negative

NMLZ	nominalizer
NOM	nominative
NONHUM	non-human
OBJ	objective
OBL	oblique
PASS	passive
PFV	perfective
PL	plural
POL	polite
POSS	possessive
PRS	present
PST	past
PTCP	participle
Q	question particle
QUOT	quotative
REFL	reflexive
SG	singular
SIM	simultaneous
TOP	topic

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.

References

Andreev, Ivan A. 1997. Chuvashskij jazyk [The Chuvash language]. In *Jazyki mira: Tjurkskije jazyki [The languages of the world: Turkic languages]*, ed. E.R. Tenishev, 480–491. Bishkek: Izdatelskij dom "Kyrgystan".

Avrorin, Valentin A., and Boris V. Boldyrev. 2001. *Grammatika orochskogo jazyka [A grammar of the Oroch language]*. Novosibirsk: Izdatel'stvo SO RAN.

Benzing, Johannes. 1985. *Kalmückische Grammatik zum Nachschlagen*, volume 1. Otto Harrassowitz Verlag.

Bulatova, Nadezhda Yakovlevna, and Lenore Grenoble. 1999. *Evenki*, volume 141 of *Languages of the World/Materials*. Lincom Europa.

Dediu, Dan, and Stephen C Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PloS One* 7:e45198.

Doerfer, Gerhard. 1988. *Grammatik des Chaladsch*, volume 4. Wiesbaden: Otto Harrassowitz Verlag.

Dunn, Michael, Stephen C Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84:710–759.

Dunn, Michael J., Angela Terrill, Ger P. Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072 – 2075.

Dyrenkova, Nadezhda P. 1941. *Grammatika Šorskogo jazyka [A grammar of the Shor language]*. Moskva: Izdatel'stvo Akademia Nauk SSSR.

Fried, Robert Wayne. 2010. *A grammar of Bao'an Tu, a Mongolic language of northwest China*. State University of New York at Buffalo.

Fritz, Susanne A, and Andy Purvis. 2010. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24:1042–1051.

Greenhill, Simon J, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C Levinson, and Russell D Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114:E8822–E8829.

Gruntov, Ilya, and Olga Mazo. 2020. A comparative approach to nominal morphology in Transeurasian. Case and plurality. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 522–553. Oxford University Press.

Hammarström, Harald, Hedvig Skirgård, Jeremy Collins, Hannah Haynie, Alena Witzlack, Stephen C. Levinson, Russell Gray, Jakob Lesage, Richard Kowalik, Robert Forkel, Linda Raabe, Suzanne van der Meer, Jana Winkler, Ger Reesink, Tessa Yuditha, Patience Epps, Luise Dorenbusch, Hilário de Sousa, Cheryl Akinyi Oluoch, Claire Bowern, Giada Falcone, Eloisa Ruppert, Martin Haspelmath, Nataliia Hübler, Karolin Abbas, Jesse Peacock, Hugo de Vos, Olga Krasnoukhova, Robert Borges, Stephanie Petit, Michael Dunn, Carolina Kipf, Jay Latarche, Nancy Bakker, Roberto Herrera, Johanna Nickel, Giulia Barbos, Kristin Sverredal, Tim Witte, Ruth Singer, Michael Dunn, Janina Klingenberg, Sören Danielsen, Swintha Pieper, and Damian Blasi. 2017. *Grambank: A world-wide typological database. Electronic database under development*. Max Planck Institute for the Science of Human History.

Heggarty, Paul. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data—and to dating language. In *Phylogenetic methods and the prehistory of languages*, ed. Peter Forster and Colin Renfrew, 183–194. McDonald Institute for Archaeological Research: Cambridge.

Heinrich, Patrick, Shinsho Miyara, and Michinori Shimoji, ed. 2015. *Handbook of the Ryukyuan languages: History, structure, and use*, volume 11. Walter de Gruyter GmbH & Co KG.

Hinds, John. 1986. *Japanese*. Croom Helm Descriptive Grammars. London: Croom Helm.

Ishizuka, Tomoko. 2012. *The passive in Japanese: A cartographic minimalist approach*, volume 192. John Benjamins Publishing.

Janhunen, Juha A. 2012. *Mongolian*, volume 19 of *London Oriental and African Language Library*. Amsterdam: John Benjamins Publishing.

Johanson, Lars. 2020. The classification of the Turkic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 104–114. Oxford University Press.

Joseph, Andrew, Seongyeon Ko, and John Whitman. 2020. A comparative approach to the vowel systems and harmonies in the Transeurasian languages and beyond. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 486–508. Oxford University Press.

Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto. 2013. *Japanese: a comprehensive grammar*. London and New York: Routledge.

Kharitonov, Luka N. 1982. *Grammatika sovremennogo jakutskogo literaturnogo jazyka. Fonetika i morfologija [A grammar of the modern standard Yakut language. Phonetics and morphology]*. Moskva: Nauka.

Kim, Juwon. 2011. *A grammar of Ewen*, volume 6 of *Altaic Languages Series*. Seoul. Korea: Seoul National University Press.

Kornfilt, Jaklin. 1997. *Turkish*. London and New York: Routledge.

Krueger, John R. 1961. *Chuvash manual: Introduction, grammar, reader, and vocabulary*, volume 7 of *Indiana University Publications, Uralic and Altaic Series*. Bloomington: Indiana University Press.

Krueger, John R. 1997. *Tuvan manual*, volume 126 of *Uralic and Altaic Series*. Bloomington: Indiana University Press.

Lebedev, Vasilij D. 1978. *Jazyk evenov jakutii [The language of the Even people of Yakut region]*. Leningrad: Nauka.

van der Lubbe, Gijs, and Akiko Tokunaga. 2015. Okinoerabu grammar. In *Handbook of the Ryukyuan languages: History, structure, and use*, ed. Patrick Heinrich, Michinori Shimoji, and Shinsho Miyara, 345–377. Berlin: De Gruyter Mouton.

Nedjalkov, Igor. 1997. *Evenki*. Descriptive Grammars. London & New York: Routledge.

Nikolaeva, Irina, and Maria Tolskaya. 2001. *A grammar of Udihe*, volume 22. Walter de Gruyter.

Orlovskaya, Mariya N. 1999. *Yazyk mongolskih tekstov XIII-XIV vv [The language of Mongolic texts of the 13th-14th centuries]*. Moskva: Institut vostokovedenija RAN.

Oskolskaya, Sofia. 2020. Nanai and the Southern Tungusic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 303–320. Oxford University Press.

Pakendorf, Brigitte, and Natalia Aralova. 2020. Even and the Northern Tungusic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 288–304. Oxford University Press.

Pellard, Thomas. 2009. Ōgami: Éléments de description d'un parler du sud des ryūkyū. Doctoral Dissertation, Paris: École des Hautes Études en Sciences Sociales, Paris.

Petrova, Taisija I. 1936. *Ul'čskij dialekt nanajskogo jazyka [The Ulch dialect of the Nanai language]*. Moskva: Gosudarstvennoje uchebno-pedagogicheskoe izdatelstvo.

Poppe, Nicholas N. 1960. *Buriat grammar*, volume 2 of *Uralic and Altaic Series*. Indiana University Publications.

Poppe, Nicholas N. 1964. *Bashkir manual*, volume 68 of *Research and Studies in Uralic and Altaic Languages*. Bloomington: Indiana University.

Reesink, Ger, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biol* 7:e1000241.

Robbeets, Martine. 2007. The causative-passive in the Trans-Eurasian languages. *Turkic Languages* 11:235–278.

Robbeets, Martine. 2017. The Transeurasian languages. In *The Cambridge Handbook of Areal Linguistics*, 586–626. Cambridge University Press.

Robbeets, Martine. 2020a. Basic vocabulary in the Transeurasian languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 645–659. Oxford University Press.

Robbeets, Martine. 2020b. The classification of the Transeurasian languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 31–39. Oxford University Press.

Robbeets, Martine. 2020c. A comparative approach to verbal morphology in Transeurasian. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 511–521. Oxford University Press.

Robbeets, Martine. 2020d. The typological heritage of the Transeurasian languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 127–144. Oxford University Press.

Sanzheev, Garma D. 1953. *Sravnitelnaja grammatika mongolskih jazykov [A comparative grammar of Mongolic languages]*, volume 1. Moskva: Izdatelstvo akademii nauk SSSR.

Sanzheev, Garma D. 1962. *Grammatika buriatskogo jazyka: Fonetika i morfologija [A grammar of the Buriat language: Phonetics and morphology]*. Moskva: Izdatelstvo vostochnoj literatury.

Savelyev, Alexander. 2020. A Bayesian approach to the classification of the Turkic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 115–124. Oxford University Press.

Schwarz, Michal, Ondřej Srba, and Václav Blažek. 2020. A comparative approach to the pronominal system in Transeurasian. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 554–584. Oxford University Press.

Shimoji, Michinori. 2012. Northern Ryukyuan. In *The languages of Japan and Korea*, ed. Nicholas Tranter, 351–380. New York: Routledge.

Shiraliev, M. 1971. *Grammatika azerbaijanskogo jazyka [A grammar of the Azerbaijani language]*. Baku: Elm.

Slater, Keith W. 2003. *A grammar of Mangghuer: A Mongolic language of China's Gansu-Qinghai sprachbund*. London and New York: Routledge.

Sohn, Ho-min. 1994. *Korean: a descriptive grammar*. London and New York: Routledge.

Sohn, Ho-min. 1999. *The Korean language*. Cambridge Language Surveys. Cambridge University Press.

Svantesson, Jan-Olof. 2020. Khalkha Mongolian. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 334–335. Oxford University Press.

Tsintsius, Vera I. 1982. *Negidal'skij jazyk [The Negidal language]*. Leningrad: Nauka.

Vajda, Edward. 2020. Transeurasian as a continuum of diffusion. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 726–734. Oxford University Press.

Whaley, Lindsay J, and Sofia Oskolskaya. 2020. The classification of the Tungusic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 80–91. Oxford University Press.

Wichmann, Søren. 2015. Diachronic stability and typology. In *The Routledge Handbook of Historical Linguistics*, ed. Claire Bowern and Bethwyn Evans, 212–214. London and New York: Routledge.

Wichmann, Søren, and Eric W Holman. 2009. *Temporal stability of linguistic typological features*. Lincom Europa.

Yuldashev, Ahnef A. 1981. *Grammatika sovremennogo bashkirskogo literaturnogo jazyka [A grammar of the modern standard Bashkir language]*. Moskva: Nauka.

3. Phylogenetic signal and rate of evolutionary change in language structures

Author's contribution

This chapter was published as a peer-reviewed research article in Royal Society Open Science (*https://doi.org/10.1098/rsos.211252*) in March 2022. Part of the data for the analyses was taken from the data set used in the previous chapter and the other part was collected afterwards specifically for this chapter. Robert Forkel and Christoph Rhymski curated the data and brought it into the cross-linguistic data format (cldf). The data were analysed by Nataliia Hübler under the supervision of Simon Greenhill. The manuscript was written by Nataliia Hübler. The author would like to thank Simon Greenhill for his valuable input on running the Bayesian phylogenetic analysis, the ancestral state reconstruction and for his continuous support during the revision process. The author also thanks Robert Forkel and Christoph Rzymski for their help in data curation. The valuable comments on the manuscript by Jakob Lesage and Volker Gast were also highly appreciated. James Boyko supported the author with the specifics of the `corHMM` package.

The data are available at:
https://doi.org/10.5281/zenodo.7135936.
The supplementary materials are available at:
https://doi.org/10.6084/m9.figshare.c.5898486.
The scripts and other materials are available at:
https://doi.org/10.5281/zenodo.6257956.

ROYAL SOCIETY
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Hübler N. 2022 Phylogenetic signal and rate of evolutionary change in language structures. *R. Soc. Open Sci.* **9**: 211252. https://doi.org/10.1098/rsos.211252

Received: 30 July 2021
Accepted: 28 February 2022

Subject Category:
Organismal and evolutionary biology

Subject Areas:
computational biology

Keywords:
phylogenetic signal, evolutionary rate, structural features, ancestral state reconstruction, Transeurasian, Altaic

Author for correspondence:
Nataliia Hübler
e-mail: nataliia_huebler@eva.mpg.de

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5898486.

THE ROYAL SOCIETY
PUBLISHING

Phylogenetic signal and rate of evolutionary change in language structures

Nataliia Hübler[1,2]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, Jena 07745, Germany
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103

NH, 0000-0002-0013-563X

Within linguistics, there is an ongoing debate about whether some language structures remain stable over time, which structures these are and whether they can be used to uncover the relationships between languages. However, there is no consensus on the definition of the term 'stability'. I define 'stability' as a high phylogenetic signal and a low rate of change. I use metric *D* to measure the phylogenetic signal and Hidden Markov Model to calculate the evolutionary rate for 171 structural features coded for 12 Japonic, 2 Koreanic, 14 Mongolic, 11 Tungusic and 21 Turkic languages. To more deeply investigate the differences in evolutionary dynamics of structural features across areas of grammar, I divide the features into 4 language domains, 13 functional categories and 9 parts of speech. My results suggest that there is a correlation between the phylogenetic signal and evolutionary rate and that, overall, two-thirds of the features have a high phylogenetic signal and over a half of the features evolve at a slow rate. Specifically, argument marking (flagging and indexing), derivation and valency appear to be the most stable functional categories, pronouns and nouns the most stable parts of speech, and phonological and morphological levels the most stable language domains.

1. Introduction

Tracing the history of languages and their speakers is a challenging undertaking. Linguists draw on various aspects of language to track these histories, often relying on 'basic' or 'core' vocabulary as a marker of language history [1–5]. However, recently more studies have been using structural features of languages to answer questions about language history and population movements [6–12].

Although it seems clear that structural features provide another source of information on the history of languages, there is an ongoing debate about whether they can recover history as well as or at a deeper level than basic vocabulary [6, p. 2073]; [11, pp. 1–2].

Some critiques argue that language structures primarily reflect the history of contact between the languages in question due to their susceptibility to borrowing and high rates of chance similarities given a limited set of states, often only 'present' or 'absent' [13, p. 3923] [8]. In fact, the stability of a set of structural features cannot be directly compared with the stability of a basic vocabulary list, as the latter was preselected for the known stability of the concepts [14, p. 122]; [15, p. 68]. A set of structural features can rather be compared with a random list of words in a language, where basic concepts and words with a high borrowability level are included.

To tackle the problem described above, there have been several attempts at defining a set of stable structural features. Nichols [16, pp. 209–210] points to items she claims are stable, comprising inclusive/exclusive opposition, head/dependent marking and alignment. Greenhill *et al.* [12] compared the rate of change in basic vocabulary and structural features and came to a conclusion that structural features (grammar and phonology) change faster than basic vocabulary on average. Nevertheless, they state that there is a core of grammatical features that evolve at a slow rate. Dediu & Levinson [11] constructed stability profiles for the features from *World Atlas of Language Structures* [17]; however, Greenhill *et al.* [18, p. 2449] argue that these data have serious limitations due to coding scheme (high level of categorization in WALS versus direct presence/absence coding in the current study, which follows the guidelines of Grambank [19]).

A major problem is that *stability* is a complex concept, often conflated with either phylogenetic signal (i.e. how well a given trait fits onto a given language phylogeny) or with evolutionary rate (i.e. how fast a trait changes on a given phylogeny). Because of this complexity, Revell *et al.* [20, p. 591] strongly encourage studies to treat stability (or conservatism), phylogenetic signal and evolutionary rate separately, as their results suggest that there is no correlation between either of these. Instead, Revell *et al.* [20, p. 591] define, for evolutionary biology, phylogenetic signal as 'the statistical non-independence among species trait values due to their phylogenetic relatedness'. For our purposes, we could translate this definition in linguistics terms: if a feature value of one language depends on the feature value of another language due to the relatedness of these languages, then the feature has a phylogenetic signal.

As for evolutionary rate, I investigate both directions of change, feature loss and feature gain, and calculate not only the transition rates between the two states, present and absent, but also the probability of states being absent or present at the nodes corresponding to proto-languages. Reconstructing some parts of the vocabulary and the phonological system of a proto-language is an ordinary procedure in comparative linguistics, but the field of linguistic and cultural evolution is still far from routinely reconstructing grammatical structures or cultural phenomena to ancestral stages. So far, there are only several studies using phylogenetic comparative methods to reconstruct individual abstract aspects of language, comprising word order [21], numeral systems [22], colour terms [23] and Indo-European grammar [24].

The two main competing forces in language evolution are inheritance and language contact, therefore it is not sufficient for our understanding of the evolutionary dynamics of structural features to study language families with a highly tree-like structure [13] and a low conflicting signal to conclude that some structural features evolve at a slow rate [24, p. 586]—we need to compare the performance of structural features across different language families, with high and low proportions of borrowing in their languages. The languages spoken across (mainly) Northern Asia provide a perfect sample for this endeavour, because they cover a large enough area, with well-known contact relations between them, and exhibit the perfect degree of genealogical heterogeneity. The languages are known to share a set of typological similarities, but the source of these similarities is unclear: there are hypotheses suggesting their genealogical relationship [25,26] as well as studies discarding these hypotheses and attributing the similarities to borrowing and chance [27,28] (the lists of the supporters and critics are by no means exhaustive). Even though the approach in this study does not aim at proving or discarding any hypotheses on the relatedness of these language families, the language sample is nevertheless highly suitable for investigating stability of structural features because of the known areal effects. If, despite the high levels of contact between the languages in the sample and a high potential for feature transfer, we can show that some structural features have a phylogenetic signal, then it would indicate that structural features convey a historical signal that is due to genealogical relationships rather than language contact.

2. Material and methods

2.1. Materials

The language sample contains languages belonging to five language families: 12 Japonic, 2 Koreanic, 14 Mongolic, 11 Tungusic, 21 Turkic languages (see figure 1 for the geographical distribution of the

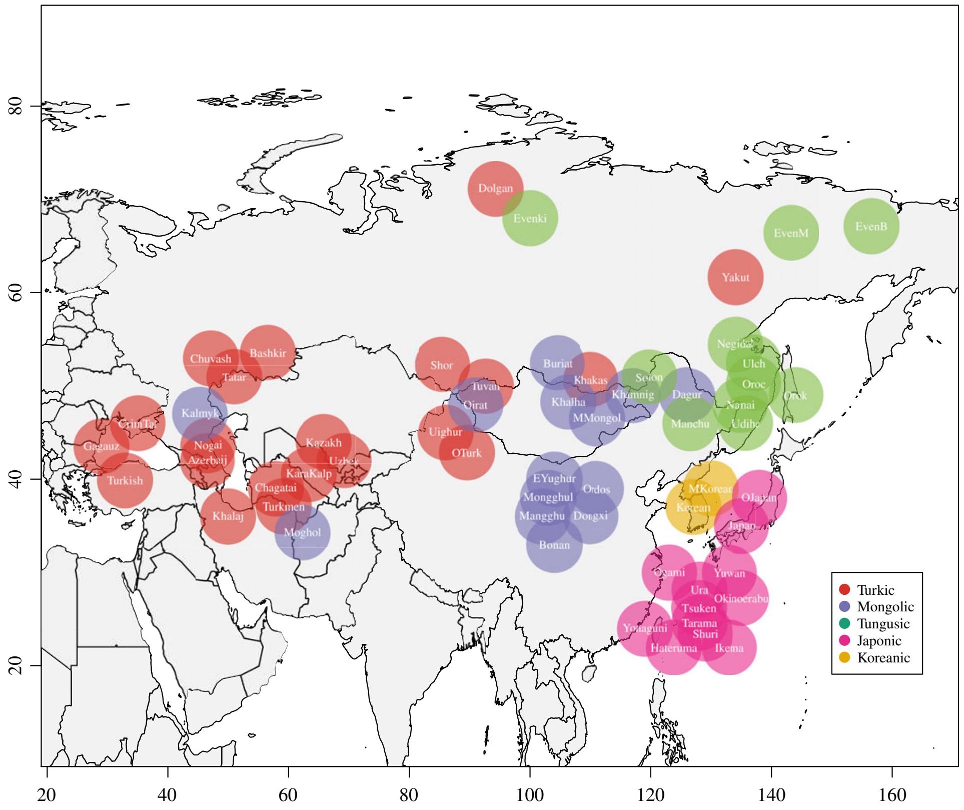

Figure 1. Distribution of the languages considered in the study. Some language names are represented as short versions of full names: Azerbaij = Azerbaijani, CrimTat = Crimean Tatar, Dongxi = Dongxiang, EvenB = Beryozovka Even, EvenM = Moma Even, KaraKalp = Kara-Kalpak, Khamnig = Khamnigan, MKorean = Middle Korean, MMongol = Middle Mongol, OJapan = Eastern Old Japanese, OTurk = Old Turkic, Uzbek = Northern Uzbek. Coordinates of languages adapted from Glottolog [29].

languages). Each language[1] was coded for 224 features. I used 189 structural features from the Grambank database [19] and binarized six Grambank features on word order (from 'What is the order of X and Y?' to 'Can X precede Y?' and 'Can Y precede X?'). I added 35 features on phonology and formal representation to increase the variability among language families. Eight of these feature formulations (TE004–TE008, TE018, TE019, TE027) are based on the feature set from Robbeets [30] and show some variation in the region. Each feature received the value 1, if the feature question could be answered with a 'yes', and the value 0, if the feature question could be answered with a 'no'. If there was not enough information on the feature in the grammar, the feature was coded as '?' (replaced by 'NA' in further analysis). Out of 224 features, 53 features were absent in all languages and were therefore excluded from further analysis (the algorithm can only be applied to the features with a value 'present' in at least some languages). The final feature set comprises therefore 171 features. Out of the 171 features, more than a half of the languages could be coded for 95% of features (162 features), around two-thirds of the languages could be coded for more than 78% of features (134 features) (see electronic supplementary material, figure S1 for the relationship between the amount of 'present', *D* and rates, and electronic supplementary material, figure S2 for the relationship between missing data, *D* and rates).

To investigate the evolutionary dynamics of features in more detail and to compare it across a relatively big number of features, I divided the features into 17 functional categories, five language domain categories and 10 part of speech categories. A short overview of the categorization and the main idea behind individual categories follow below, for the categorization of individual features, see electronic supplementary material, table S1.

[1] For the sake of simplicity, I use the term 'language' throughout instead of the more accurate term 'doculect', i.e. a language as it is described in the grammar. The information available in the grammar may be different from the current state of the language or variety holding the same name or different from the variety known to the reader.

The four chosen language domains (or levels) are: 'phonological shape' (14 features), 'word' (71 features), 'nominal phrase' (21 features), 'clause' (63 features) and 'other' (4 features). Features that target the form of the word (phonological shape) comprise vowel harmony (4 features), phonotactic constraints (3 features), voicing/aspiration distinctions in consonants (4 features), l/r distinction. The category 'clause' comprises features that have the whole clause as their scope. Most features have to do with phonologically free marking; in some features, there is variation, e.g. the feature on negation marking appearing clause-finally versus clause-initially: in many languages in the sample, negation is marked by a suffix on the verb, and, due to subject–object–verb (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 feature sets included are: comparative construction (4 features), predicative possession (5 features), interrogation (7 features), negation (5 features). The category 'word' comprises features that target the word and where the presence of the feature is realized by a bound marker. The most prominent feature sets in this category include case marking, indexing, derivation, number and possession marking, morphological tense–aspect–mood (TAM) marking, and valency markers on verbs. The category 'NP' covers word order and agreement in the noun phrase. Features on the adpositions, articles and nominal conjunction are also included in this category. The features that could not be assigned to any of the above-mentioned categories were categorized as 'other'.

Parts of speech are a highly disputed topic in linguistics and it is by no means a trivial matter to assign words of one language to a particular part of speech, let alone when we deal with 60 languages at a time. For the current exploratory purposes, they nevertheless appear to be a useful proxy for explaining stability of particular features. The part of speech categories include features that could be described as targeting 'adjective' (5), 'article' (4), 'demonstrative' (3), 'noun' (19), 'particle' (12), 'pronoun' (15) and 'verb' (54). A significant number (15) of features concerns both nouns and pronouns, therefore a separate category ('noun/pronoun') appeared worthwhile. Features targeting the presence of pre- and post-positions and ideophones could not be conflated with any other part of speech and form their own category 'other' (3). A number of features (41) could not be assigned to a part of speech. These are mostly features that otherwise fall into the functional categories 'phonological distinctiveness' and 'word order' and language domain categories 'nominal phrase' and 'clause'.

The functional categories are: 'argument marking (core)' (10), 'argument marking (non-core)' (8), 'deixis' (15), 'derivation' (5), 'interrogation' (8), 'modification' (6), 'negation' (7), 'phonological distinctiveness' (14), 'possession' (11), 'quantification' (17), 'TAME+' (23), 'valency' (11), 'word order' (22) and 'other' (14). In the categories 'argument marking (core)' and 'argument marking (non-core)' both features on flagging (marking on the nouns and pronouns) and indexing (argument marking on the verbs) are included. 'Deixis' covers features on articles, pronouns (except case marking), and demonstratives. 'Derivation' includes features on deverbal and denominal derivation (action/state, agent and object derivation, diminutive and augmentative marking). 'Interrogation' covers features on the manner of expression of interrogation as well as position of the interrogation markers. 'Modification' includes features on the comparative construction and on adjectives acting as verbs. 'Negation' covers features on the negation of verbs and other types of predicators. 'Phonological distinctiveness' overlaps completely with the category 'phonological shape' from the language domain categorization. 'Possession' includes features on attributive (ways of expressing 'my goat') and predicative (ways of expressing 'I have a goat') possession. 'Quantification' spans over numeral systems, classifiers, nominal number marking and agreement in number in a nominal phrase. 'TAME+' covers both tense–aspect–mood–evidentiality marking (phonologically free and bound) and other non-derivational marking on or modification of verbs. 'Valency' includes features on causatives, applicatives, passives and other valency-related phenomena. 'Word order' spans from the order of components in the nominal phrase to order in the clause and the position of the relative clause according to the noun. Features that would require opening up small categories were grouped together in the category 'other'. Features on reduplication, verbal compounding, copula for predicate nominals, existential verb, ideophones, clause chaining, light verbs and others are included in this category.

2.2. Methods

To serve as the 'gold standard' reconstruction of the relationships between these languages, I constructed a phylogeny from the classification taxonomy from Glottolog (v. 4.2.1) [29] for each of the five language families in the dataset. Glottolog is an independent catalogue of language relationships and references.

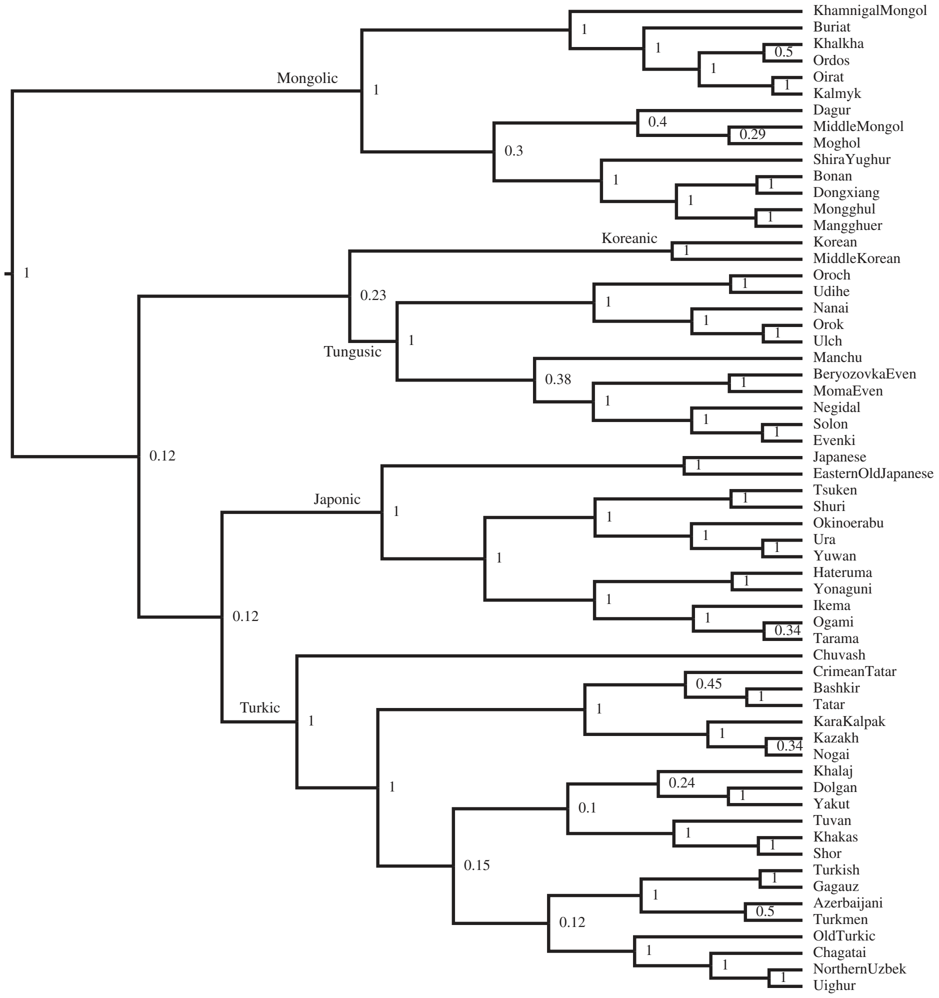

Figure 2. The maximum clade credibility tree: low probability on the node indicates either an agnostic view on the order of the splits or that a branch split into more than two further branches (i.e. a non-binary structure of the tree); high probability on the node (1) indicates the monophyletic constraint on the node according to the Glottolog (v. 4.2.1) [29] classification, except for the root probability value (1), which is an artefact of the fact that no languages, apart from Transeurasian, were included in the sample (i.e. there is no outgroup).

The subgroupings in the Glottolog classification were used to enforce monophyletic clade constraints (figure 2). The classifications are based on Johanson [31, pp. 161–162] for Turkic, Rybatzki [32, pp. 386–389] for Mongolic, and Pellard [33, pp. 5–8] for Japonic. The Tungusic classification is based on the three-branch proposal of the family by Doerfer [34] referred to by Whaley [35, p. 397].

I used BEAST (v. 2.5.1) [36] to build a data-free 'pseudo-posterior' of trees from this classification using a covarion model of evolution [37] and a relaxed clock model [38]. As there was no linguistic data beyond the tree topology, I ran this analysis for 10 000 000 generations, sampling a tree from the posterior every 1000 generations. This procedure provided a posterior probability distribution of the trees with each node in the Glottolog classification having a posterior probability of 1.0, while unresolved groupings were assigned low probabilities but—importantly—retaining the uncertainties in the language subgroupings. This allows me to adopt an agnostic view on the order of splits of branches. For example, according to Glottolog, the South Kipchak branch comprises three languages: Kara-Kalpak, Kazakh and Nogai. The branch itself has the probability 1.0, because these languages definitely belong to the same branch, but the probability of the Kazakh-Nogai branch is low, namely 0.35, which means that the clustering of Kazakh and Nogai together is arbitrary and Kazakh and Kara-Kalpak could have also belonged to one branch with similarly low probability. As there is no information in Glottolog about relationships above the language family level like Transeurasian, the relationships between the language families are likewise arbitrary and all deep groupings are equally likely to appear in the posterior sample (note that posterior probabilities for the higher nodes are below 0.25).

To measure the phylogenetic signal, I use metric *D* [39]. This metric calculates the sum of the differences between related branches in a tree. *D* is the sum of sister clade differences across the tree and its values normally fall in the range between 0 and 1. In a trait that is strongly phylogenetically structured, sister languages will share the same value (and have no difference). If the trait is not phylogenetically patterned, then sister languages will have different values: the *D* value will be high and the phylogenetic signal will be low. I computed the *D* statistic on a sample of 1000 trees from the posterior probability distribution described above, using R (v. 4.0.1) [40] and the function *phylo.d* in the package *caper* (v. 1.0.1]) [41].

To measure the evolutionary rate and reconstruct ancestral states, I used the Hidden Rates model as implemented in the function *corHMM* from the package *corHMM* (v. 2.4) [42]. The function allows to choose between two models: *ARD* (= all rates differ) and *ER* (= equal rates). As there are no strong grounds to assume that linguistic features are gained and lost at the same rate and as the differences between these rates will be particularly interesting for explaining linguistic diversity in future research, I chose the model *ARD*. The rate of gain ($0 \rightarrow 1$) is therefore allowed to be different from the rate of loss ($1 \rightarrow 0$). I set the root prior following Maddison *et al.* [43] and FitzJohn *et al.* [44]. The Hidden Rates model foresees in our case two rate classes, a fast one (F) and a slow one (S), and two possible values: 0 and 1. A feature value has an equal probability of belonging to a slow and to a fast rate class. Each observed feature value would therefore belong to one of the classes and have a particular value, e.g. 1F or 1S. Features belonging to different classes can potentially have different transition rates. According to this model, there are eight possible transition rates: 1F to 0F, 0F to 1F, 1S to 0S, 0S to 1S, 1F to 1S, 1S to 1F, 0F to 1S and 1F to 0S. The rates cannot be observed directly—the affiliation with different rate classes can only be derived from the states, therefore such a model is commonly known as a hidden Markov model [45, p. 726]. Ancestral states were estimated using marginal approach, which integrates over the states at other nodes and calculates the likelihood of state at each node [42].

I estimated Kendall's τ to measure the correlation between the phylogenetic signal and the evolutionary rate.

3. Results

Over a half of all features (63%) have a median *D* value below 0.5 and 37% of features have a median *D* value over 0.5. If we want to categorize the distribution of *D* values with more precision, we can divide them into four categories, depending on the range the value falls into: $D < 0$ in overclumped features (46%), *D* between 0 and 0.5 in features with a phylogenetic signal (17%), *D* between 0.5 and 1 in randomly distributed features (23%) and $D > 1$ for overdispersed features (14%).

As for the rate, over half of the features are gained and lost at a slow rate: 68% for feature loss and 75% for feature gain. Only approximately one third of the features evolve at a fast rate: 32% for feature loss and 25% for feature gain.[2]

For a more fine-grained categorization, one could divide the features into three categories: below −0.5 'slow', between −0.5 and 0.5 'medium' and above 0.5 'fast' (table 1). We see that the group of features lost at a fast rate is bigger than the group of features gained at a fast rate and a reverse trend for the slow rate: there are more features gained at a slow rate than lost (see table 2 for the measures of centre and dispersion and figure 3 for the distribution of the *D* and rate values; the *D* and rate values for individual features can be found in electronic supplementary material, tables S2 and S3).

Despite the observations made by studies in evolutionary biology [20], I find a moderate positive correlation between the phylogenetic signal and the evolutionary rate (figure 4): τ for the rate of loss and gain is 0.51 and 0.5 respectively, *p* values approximate 0, i.e. features with a high phylogenetic signal tend to evolve at a slower rate, while features with low phylogenetic signal tend to evolve at a fast rate. There are 58 (gain) and 34 (loss) features that have rate equal to 0 prior to $\log_{10}$

[2] I use the $\log_{10}$ transformed values of evolutionary rate for the further analysis of the results, because the transformation returns a normal distribution, which allows for a better visualization and interpretation of the distribution. I assume that features with the $\log_{10}$ value below 0 evolve rather slowly and those above 0 evolve rather fast. Since $\log_{10}$ transformation of 0 would produce infinite values, I replaced rates equal to 0 by a value close to 0 (0.0000000001) prior to $\log_{10}$ transformation.

Table 1. Evolutionary rate (proportion per category).

category	gain	loss
slow	0.65	0.47
medium	0.2	0.28
fast	0.15	0.25

Table 2. Phylogenetic signal and evolutionary rate, summarized based on the median value per tree per feature.

metric	min	median	max	s.d.
D	−4.97	0.04	2.34	1.43
rate of loss	0	0.33	99.98	17.35
rate of gain	0	0.14	95.59	15.15
rate of loss ($\log_{10}$ transformed)	−10	−0.48	2	4.08
rate of gain ($\log_{10}$ transformed)	−10	−0.85	1.98	4.74

transformation. These features are thus almost never lost or gained: they are either present in all languages (e.g. 56 out of 56 languages with enough information on the feature or 60 out of 60 languages, i.e. the whole sample) or absent in all but one or two languages (e.g. 'present' in 1 out of 60 or 2 out of 24 languages). For the distribution of the *D* values across these features, see electronic supplementary material, figure S7.

We see an overall trend of an increase in evolutionary rate and in *D* with an increase in language level up to the nominal phrase and then a slight decrease at the clause level (see figure 5 and table 3). There are no levels evolving at a fast rate (above 0) on average, apart from the features that could not be attributed to any level ('other'). The category 'other' is also distributed randomly on the phylogeny, alongside 'nominal phrase' (these are the only categories without a phylogenetic signal). Features operating on the levels 'word' and 'phonological shape' are lost at the lowest rate on average. The category 'word' maintains its dominant position in terms of slow rate of change also for feature loss, but is followed by the category 'nominal phrase'. In terms of phylogenetic signal, 'phonological shape' is the most overclumped category on average, followed by the category 'word'.

The majority of functional categories have median *D* below 0.5 (i.e. the features have a phylogenetic signal or are overclumped), except for 'interrogation', 'quantification' and 'TAME+' (see figure 6 and table 4). Features belonging to the categories 'argument marking (non-core)' and 'derivation' are gained at the slowest rate on average, but the category 'argument marking (core)' is ahead of other functional categories in terms of rate of loss, followed by 'argument marking (non-core)', 'valency' and 'derivation'. In terms of phylogenetic signal, 'modification' is the most clumped functional category, followed by 'argument marking (core)' and 'valency'. The fastest changing and most overdispersed functional category is 'interrogation'. Functional categories 'deixis' and 'TAME+' follow 'interrogation' in the rate of loss and in the reverse order ('TAME+', then 'deixis') in the rate of gain. They also belong to the few overdispersed functional categories in terms of phylogenetic signal, alongside 'interrogation' and 'quantification'.

The most slowly evolving parts of speech are pronoun, noun and 'other' (i.e. adpositions and ideophones) (see figure 7 and table 5). 'Pronoun' is also the most clumped category in terms of phylogenetic signal, followed by the category 'noun/pronoun' (see §2 for clarification). There is no part of speech that would be lost fast: the median for all of them lies below 0. The relatively fast lost parts of speech are demonstrative, adjective, article and particle. The same four parts of speech are the only categories that have the rate of gain above zero. Apart from adjective ($D = 0.4$), these same categories are also overdispersed in terms of phylogenetic signal ($D > 0.5$), complemented by the category 'other'.

The reconstructability of features does not show remarkable variation across language families (figure 8): the proportions of features reconstructable as 'present' or 'absent' are similar enough in all five language families for a joint summary. About one fifth of the features (19–22%) in the feature set

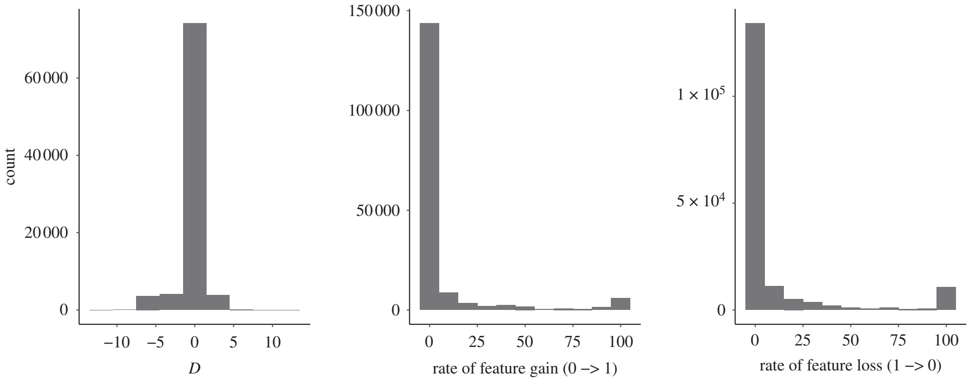

Figure 3. Distribution of phylogenetic signal (*D*) and evolutionary rate (transition from 0 to 1 and from 1 to 0) across 1000 trees and 171 features.

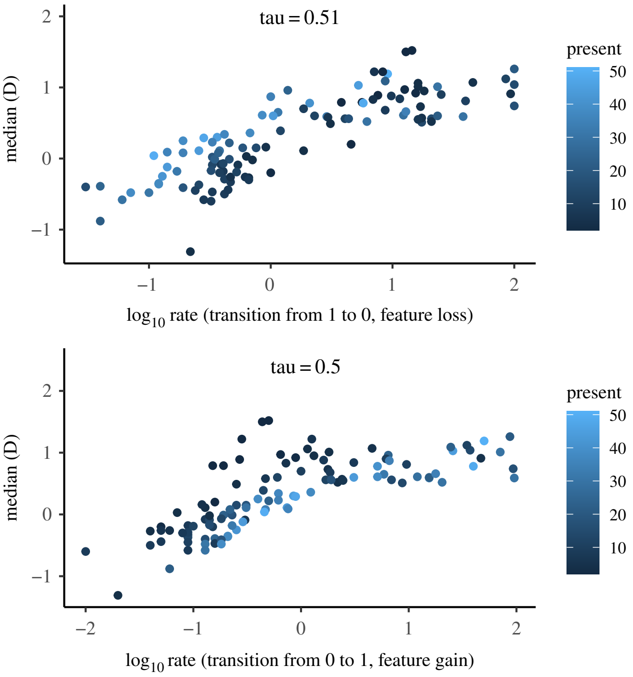

Figure 4. Correlation between phylogenetic signal and transition rates. Dark colour of the data points indicates features with most values 'absent' and light colour indicates features with most values 'present'. Features with the rates equal to zero prior to the $\log_{10}$ transformation (58 for feature gain and 34 for feature loss) are not included in the plot: these features are either present in all languages or absent in all but one or two languages.

can be reconstructed with 95% certainty, and around one third of the features (33–38%) can be reconstructed with 75% certainty of being present in the proto-language. About one fifth to one fourth (21–26%) of the features can be reconstructed with 95% certainty and almost a half (44–50%) with 75% certainty of being absent in the proto-language. Overall, there is only a small range of features (17–22%) that cannot be reconstructed as either 'present' or 'absent' (between 25% and 75% certainty in reconstruction as 'present') in the proto-language.

In order to make the reconstructability of features belonging to particular categories comparable across categories, I normalized the number of features reconstructable as 'present' (≥95%) by the number of features belonging to each category. This was necessary to eliminate the impact of some categories with a high number of features (e.g. 'word' (71), 'clause' (63), 'verb' (54), 'not assignable' (41), see §2 for details). There are some differences in the reconstruction across language domains and

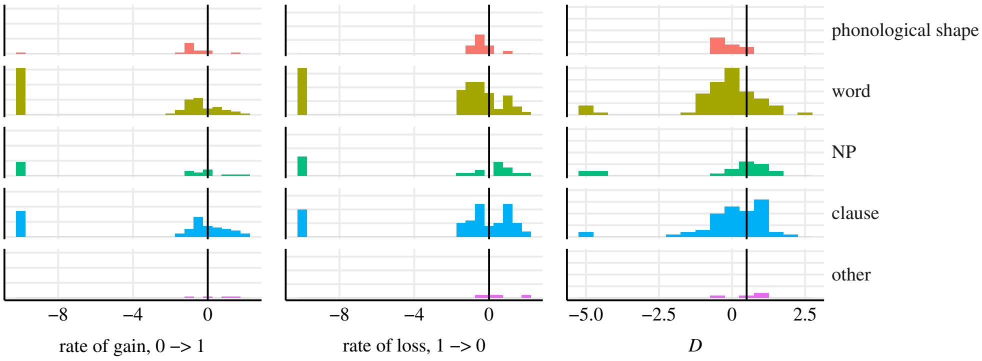

Figure 5. Feature loss and gain across different language levels. The vertical line marks the division between low and high $\log_{10}$-transformed evolutionary rate and the division between phylogenetic signal and random distribution on the phylogeny.

Table 3. Phylogenetic signal and evolutionary rate across language levels.

level	median rate (loss)	median rate (gain)	median *D*
phonological shape	−1.02	−0.36	−0.24
word	−1.22	−0.92	−0.01
NP	−0.89	−0.39	0.57
clause	−0.57	−0.38	0.22
other	0.29	0.25	0.77

Table 4. Phylogenetic signal and evolutionary rate across functional categories.

level	median rate (loss)	median rate (gain)	median *D*
argument marking (core)	−5.44	−1.46	−0.42
argument marking (non-core)	−10	−1.31	−0.07
deixis	−0.52	0.27	0.16
derivation	−10	−1.05	0.16
interrogation	0.68	0.97	0.84
modification	−0.87	−0.82	−0.44
negation	−0.57	−0.59	0.08
other	−5.52	−0.85	0.04
phonological distinctiveness	−1.02	−0.36	−0.24
possession	−0.74	−0.38	0.04
quantification	−0.82	−0.44	0.8
TAME+	−0.22	0.02	0.6
valency	−1.05	−1.15	−0.29
word order	−1.17	−0.52	−0.05

across language families (figure 9), but the overall trends overlap in most language families: the categories 'nominal phrase' and 'word' have the highest proportion of reconstructable features among level categories, 'pronoun' and 'other' (adpositions and ideophones) among parts of speech and 'derivation' among functional categories. Features in the category 'phonological shape' are better reconstructable for Mongolic and Tungusic languages than for the other families. There is a striking difference in the proportion of well-reconstructable features belonging to the categories 'verb' and

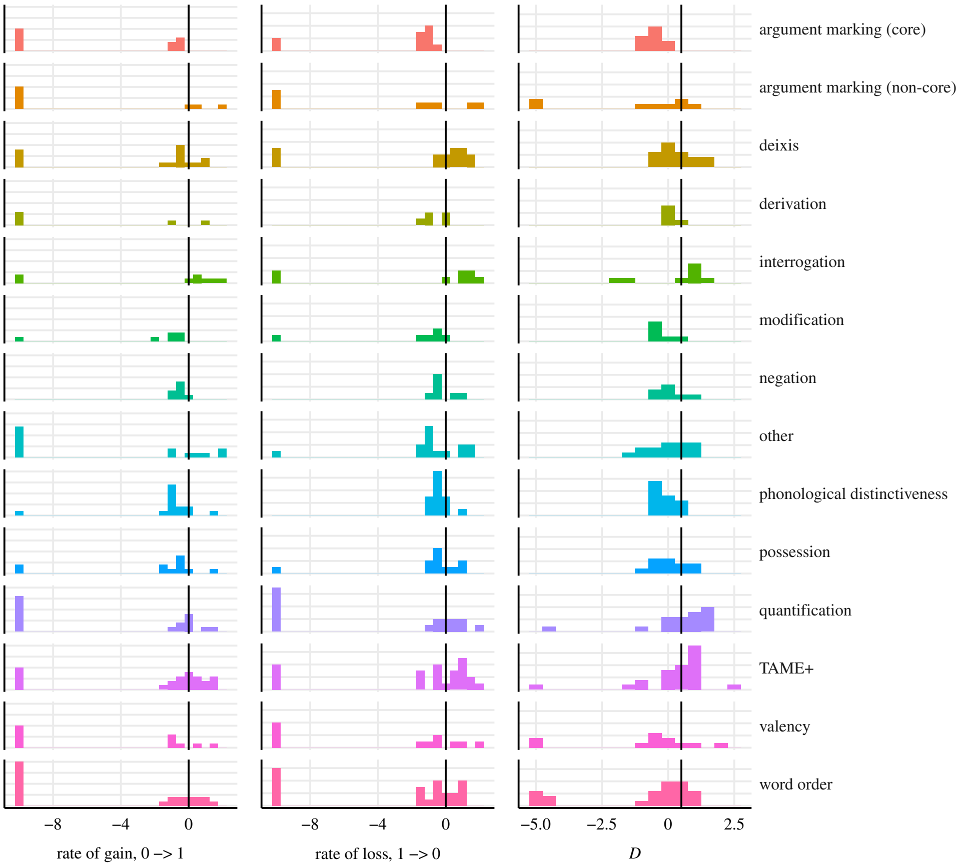

Figure 6. Feature loss and gain across different functional categories. The vertical line marks the division between low and high $\log_{10}$-transformed evolutionary rate and the division between phylogenetic signal and random distribution on the phylogeny.

Table 5. Phylogenetic signal and evolutionary rate across parts of speech.

part of speech	median rate (loss)	median rate (gain)	median *D*
adjective	−0.35	0.08	0.4
article	−0.45	0.92	0.68
demonstrative	−0.3	0.61	0.56
not assignable	−1.05	−0.43	−0.08
noun	−10	−1	0.15
noun/pronoun	−0.74	−0.77	−0.18
other	−10	−10	0.51
particle	−0.58	0.62	0.93
pronoun	−10	−0.85	−0.28
verb	−0.7	−0.59	0.11

'pronoun': 'pronoun' is the best reconstructable category in so-called Altaic languages, with 'verb' being rather moderately represented, but the distribution is the opposite for Japonic and Koreanic languages: here verbs are approximately equally well reconstructable as pronouns.

As might have been expected, the proportion of features that can be reconstructed to the proto-language with 95% probability for pairs of language families falls out slightly lower than the

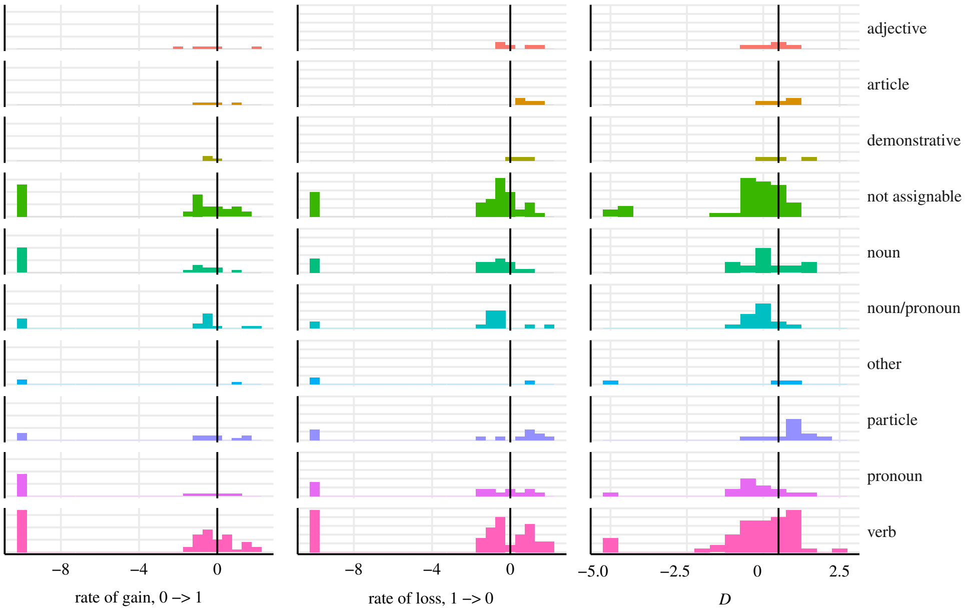

Figure 7. Feature loss and gain across different parts of speech. The vertical line marks the division between low and high $\log_{10}$ evolutionary rate and the division between phylogenetic signal and random distribution on the phylogeny.

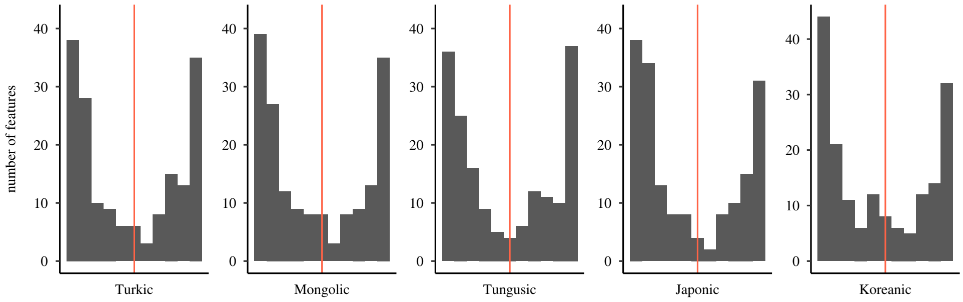

Figure 8. Reconstructed states per language family: *X* axis from 0 (absent) to 1 (present), the red line indicates the point of highest uncertainty in reconstruction, 0.5.

proportion of features reconstructed for individual language families, e.g. we can reconstruct the same 33 features (19.14% of features) for Turkic and Mongolic together and 36 features for Turkic and Mongolic separately. A pairwise comparison of features reconstructable to proto-languages indicates that Turkic/Tungusic and Mongolic/Tungusic pairs have the highest number of shared features (20.3%) that can be reconstructed with 95% probability as 'present' in the proto-language (see table 6 for the counts on other pairs). The tentative grouping of Japonic and Koreanic has the same amount of well-reconstructed features (17.4%) as Japonic/Turkic, Japonic/Tungusic and other pairs.

4. Discussion

In the terms used in this study, a feature is stable if it evolves at a relatively slow rate and has a high phylogenetic signal. The results have shown that 66% of the features in the dataset have a *D* value below 0.5 and evolutionary rate below 0 (after the $\log_{10}$ transformation). Therefore, more than half of the features can be called relatively stable.

Why are some features more stable than others? Greenhill *et al.* [12, p. 4] and Arnold [46, p. 76] both independently propose the availability of a feature for reflection and analysis as a tentative explanation for its (in)stability. I see some support for this idea in my results: we can speculate that number marking

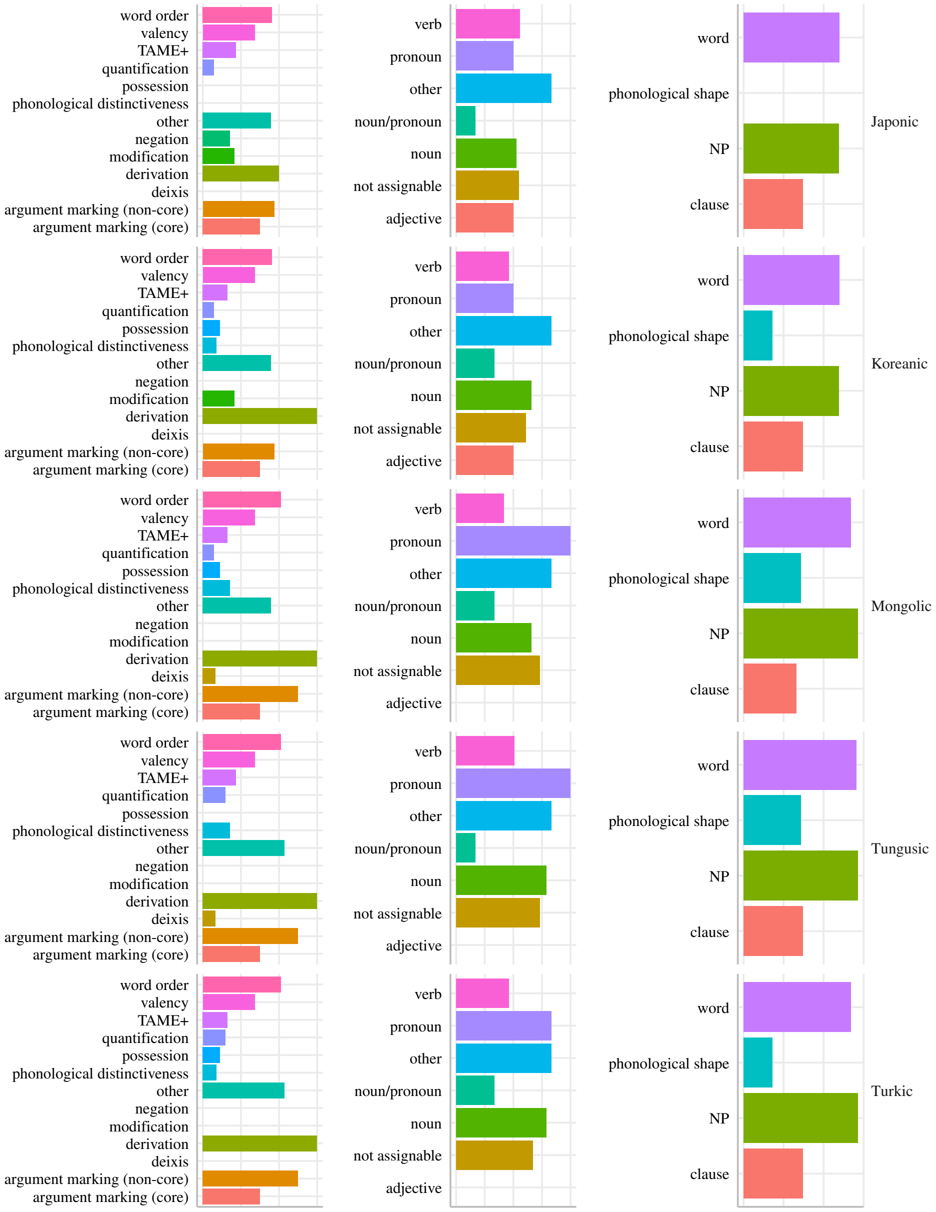

Figure 9. Proportion of features reconstructed with 0.95% certainty as 'present' across categories and language families. The number of features reconstructable per language family per category was normalized by the number of features belonging to each category.

Table 6. Overlaps in reconstructed features for pairs of language families at 95% probability of 'present', in %.

language family	Turkic	Mongolic	Tungusic	Koreanic	Japonic
Turkic	X	19.14	20.3	16.82	17.4
Mongolic	19.14	X	20.3	17.4	16.24
Tungusic	20.3	20.3	X	17.4	17.4
Koreanic	16.82	17.4	17.4	X	17.4
Japonic	17.4	16.24	17.4	17.4	X

is more analysable for speakers than derivation of nouns from verbs, core argument marking is less conscious than oblique argument marking, valency is less analysable than tense–aspect–mood marking. Interrogation probably needs most conscious processing by speakers compared with other categories and is thus more prone to change, be it as a result of a single innovation in the community of speakers or of a borrowing event. One could hypothesize that several ways of interrogation marking could exist in parallel in a language (possibly with one way dominant at a time) and the choice of marking could depend on the pragmatic situation (and the speaker could therefore decide upon the marking spontaneously). We could apply a similar explanation also to language levels: the higher the level, the more conscious are the speakers of their language use. It follows from the results that the features operating on the phonological and word level are the most stable. This trend does not extend on the nominal phrase and clause: here we see an opposite situation with 'clause' being more stable than 'NP', but it does not mean that the explanation of reflection cannot be applied here. On the contrary, the word order in the noun phrase, the use of adpositions, articles and conjunctions might be more available to analysis than the position of the negation marker, the verb, the way one expresses possession of the type 'I have a dog'. One should note that 'clause' appears relatively stable despite the fact that it includes features from the most unstable functional category, 'interrogation'. The ease of analysis of particular language categories by speakers would profit from a thorough investigation by other disciplines.

In some cases, variation in stability could also be explained by the frequency of use (on the relationship between frequency and stability, see Bybee & Thompson [47] and Diessel [48, p. 118]). We could apply this explanation to the parts of speech: pronouns and nouns appear to be the most stable and the most frequently used parts of speech. Articles, on the contrary, are neither obligatory nor frequent in the languages of the sample, therefore, as one might expect, they are rather unstable both in terms of rate and phylogenetic signal. This might be different for a different language family, where articles are obligatory. Following the logic of frequency of use, 'phonological distintiveness' should be the most stable functional category, but it appears to be rather intermediate both in terms of rate and phylogenetic signal—we would need to develop an explanation for it in further research. In order to connect the stability of interrogation and other functional categories to the frequency of use, one would need a corpus of the languages in the sample (e.g. to compare the frequency of declarative and interrogative clauses and draw conclusions from the frequency distributions).

Another possible explanation is the areal spread of the feature. Word order in noun phrases appears to be identical in the whole area: all modifiers appear before the noun, and most of the languages in question have been in contact either with each other, or a neighbouring language, often Mandarin Chinese or Russian. It is therefore difficult to conclude that constituent order is *per se* a genealogically stable feature, if language contact cannot contribute to variation. As for clausal word order, it has some variation in the area: all Transeurasian languages have OV order, whereas influential neighbouring languages have VO order. In some Transeurasian languages, also VO order is possible, due to borrowing, but most of the languages could resist the influence of dominant neighbouring languages and retained OV order.

In order to better assess the results being generalizable to world's languages, it is worth comparing the stability of the features in my results to those of the previous studies. According to Nichols [49, p. 353], the most genetically[3] stable features are the alignment of head/dependent marking, inclusive/exclusive oppositions, gender, number oppositions in the noun, and detransitivation processes. For head/dependent marking, we consider features on attributive possession, indexing and core case marking (argument marking (core)). The results of this study support the findings of Nichols [49] in this respect: features belonging to the functional category 'argument marking (core)' ($D = -0.42$, $q01 = -1.46$, $q10 = -5.44$)[4] and features on indexing ($D = -0.35/-0.36$, $q01 = -0.68$, $q10 = -0.92$) are both lost and gained at a slow rate and are overclumped in terms of phylogenetic signal. Features on attributive possession evolve at a slow rate, but differ in their phylogenetic signal: marking of possession by a suffix on the possessed is overclumped ($D = -0.9$), on the possessor has a phylogenetic signal ($D = 0.04$) and marking the possessed with a prefix is extremely dispersed ($D = 1.04$, present only in one language, Hateruma). Inclusive/exclusive distinction (in the current language set: in pronouns only) is almost never gained ($q01 = -10$) and lost at a slow rate ($q10 = -0.42$), it is 'overclumped' in terms of phylogenetic signal ($D = -0.15$). There is no gender in the

[3] Terminology of Nichols [49] is preserved.

[4] $q01$ stands for the transition from 0 (absent) to 1 (present), or feature gain, $q10$ for the transition from 1 (present) to 0 (absent), or feature loss. The value −10 after the $\log_{10}$ transformation corresponds to the value 0 before the $\log_{10}$ transformation.

languages in question, therefore we can only say that gender is difficult to gain, despite neighbouring languages (e.g. Russian) having gender. The features in the functional category ‘quantification’, which includes more features than mentioned in Nichols (1993) [49], evolve at a relatively slow rate, but have no phylogenetic signal. Some features of interest here are: associative plural marker ($D = 0$, $q01 = -10$, $q10 = -1$), suppletion for number ($D = -1.24$, $q01 = -10$, $q10 = -10$) and non-phonological allomorphy of noun number markers ($D = 0.15$, $q01 = -0.6$, $q10 = -0.12$), which all evolve at a slow rate and have a phylogenetic signal. Plural marking on nouns ($D = 0.6$, $q01 = -0.07$, $q10 = -0.44$) also evolves at a slow rate, but is more dispersed than the other features on nominal number. As for the detransitivising processes, features on valency are gained and lost at a slow rate ($q01 = -1.15$, $q10 = -1.05$) and have a high phylogenetic signal ($D = -0.05$). Morphological passive marking evolves slowly and is overclumped in terms of phylogenetic signal ($D = -0.29$, $q01 = -10$, $q10 = -1.15$). Even if these are not the only features with a high phylogenetic signal and evolving at a slow rate, the results of the current study do not contradict the findings of Nichols [49] in most cases.

In her 1992 book [16, p. 167], Nichols mentioned word order as being very genetically unstable. This finding is only partially supported by my results: the category ‘word order’ as a functional category evolves at a slow rate ($q01 = -0.52$, $q10 = -1.17$) and has a high phylogenetic signal ($D = -0.05$), but the stability varies for particular orders of constituents (OV is more stable than VO). Nichols [49, p. 353] classifies clause word order as areally stable, and this is definitely true for the current language sample: the word order is verb-final in all languages in the sample ($D = -4.95$, $q01 = -10$, $q10 = -10$), with four languages (Gagauz, Khalaj, Beryozovka Even and Moghol) also allowing verb-medial word order due to borrowing ($D = 1.07$, $q01 = 0.06$, $q10 = 1.21$). It is difficult to conclude that verb-final word order is more stable than verb-medial word order: in a language sample, where 60 out of 60 languages have verb-final word order, this feature has not changed, but this might well be the case for verb-medial word order in an area or a language family, where this particular order dominates.

Greenhill *et al.* [12] quantify the stability of structural features in terms of evolutionary rate. Based on a sample of 81 Austronesian languages, this study sets apart the following features as being particularly stable: inclusive versus exclusive distinctions, gender distinction in third person only in pronouns, tone, future marking on the verb, conflation of categories (e.g. alignment, conflation of second and third persons in non-singular numbers), which mostly overlap with those of Nichols [49]. According to my results, morphological future marking has a phylogenetic signal and is lost and gained at a slow rate ($D = 0.11$, $q01 = -0.51$, $q10 = -0.42$). Gender in third person pronouns ($D = -0.43$, $q01 = -10$, $q10 = -10$) is almost never distinguished in the languages in question, apart from Japanese, where pronouns are generally omitted (and third person pronouns the more so). There is no data on the conflation of second and third persons in non-singular numbers available, and this is not relevant for the area in question. The current results thus go in line with the findings of Greenhill *et al.* [12].

One of the conclusions of a recent study on the evolutionary rate in structural features based on Indo-European languages [24] is that morphological features evolve slower than syntactical features. The current study provides evidence in support of this conclusion: the level ‘word’ evolves at the slowest rate among all level categories and takes in the second position in terms of phylogenetic signal, giving way only to ‘phonological shape’.

Since we see support of the presented results in previous studies, which did not focus on the same region, and since this study covers five language families (albeit with hypothesized genealogical relationships or at least forming a sprachbund), we can say that stability patterns of different areas of grammar might have a cross-linguistic component. Nevertheless, we can only draw conclusions on the stability of particular features for one unit at a time, be it a language family or an area, because we can only measure features that are present in the given family or area. We cannot fully compare these results with those for Indo-European or Austronesian, because there are typical Austronesian, Indo-European and ‘Transeurasian’ features.

Already, at the stage of data collection, it becomes obvious that the area is very homogeneous: many features in the questionnaire are either invariable or deviate for very few languages (whether present or absent): 118 features are present in 50 languages (out of 60) or more, 53 features of the initial 224 features appeared to not to be present in the area. This can be explained by the nature of the questionnaire: at the core of the feature set from Grambank is the selection of features exhibiting some variation in the languages of Island Melanesia [6]. Therefore, features interesting for that area are uninformative for Northern Eurasia and were discarded in further analysis.

This development is not new in linguistics: there were several adjustments to the basic vocabulary list after it was shown that not all items on the lists are universal and present in all languages, as they were originally claimed to be. Some languages appeared e.g. to lack words for ‘snow’, ‘ice’, ‘freeze’ and ‘sea’

[50,51]. The same way as these words typical for cold regions of the Earth tend to be stable in the languages, where they are present, some features are typical and stable in one part of the world, but completely absent in the other part of the world—and thus irrelevant for the studies on stability of structural features in that region.

We could think of stability not as a universal phenomenon, but as a trend that depends on the area and genealogical affiliation: one could expect language families with similar typological profiles (e.g. more synthetic or more analytic) to show similar stability patterns in their grammars, e.g. that morphology will tend to be stable if the language makes use of extensive morphological marking. If the language rather uses free marking more often than bound marking, then one might hypothesize that this marking will be more stable due to its higher frequency. The current study is thus a step towards a list of universally stable structural features (if it were ever to be determined), but not the final destination in the construction thereof: it provides ground for testing new hypotheses for stability patterns in other language families.

Given the correlation between the rate and the phylogenetic signal mentioned in §3, there is often an overlap between the rate of loss, gain and phylogenetic signal across categories. Can we conclude from this that it is inessential to investigate these two dynamics simultaneously and one metric would have been sufficient? It appears that we would lose information if we only considered either of the two: if we only accounted for the rate of loss, we would conclude that articles are definitely stable, because they are difficult to lose, but considering also the rate of gain we see that they are far more easy to gain and do not have a phylogenetic signal, but are rather distributed randomly on the tree. Therefore, a straightforward conclusion that articles are clearly stable would be misleading. Overall, there is a discrepancy between the phylogenetic signal and evolutionary rate in 40 features, i.e. some features with high D evolve slowly and some features with low D evolve fast. The method used in this study allows us to get a more complete picture of the evolutionary dynamics of structural features and prevents us from making precipitous conclusions.

We have seen that more than a half of the structural features bear a high phylogenetic signal and evolve at a slow rate. If the features are preserved relatively well, can we reconstruct them to the proto-language level? It might be interesting to compare the reconstructability of features across language families to see if there are family-specific trends, e.g. is there an especially innovative language family, where very few features can be reliably reconstructed, or an overly conservative language family, where most of the features can be traced back to the proto-language? We could also determine the features that can be reconstructed as ‘present’ to the proto-language level across all five families.

Ancestral state reconstruction can be performed most reliably if most members of the family are sampled. Unfortunately, we cannot know how many languages once belonged to the five language families in the sample, but we can assume that the current number of languages in these families is only a small fraction of the languages once spoken in the area, especially given that nowadays most Tungusic languages are severely endangered, most Mongolic, Turkic and Japonic (Ryukyuan) languages do not enjoy high prestige, or speakers are discouraged from using their native languages, there are only two Koreanic languages in the sample, one of which is merely the older stage of Modern Korean. This is a limitation of the study, which we have to bear in mind when interpreting the results, but there is no known possibility at the moment to account for it.

As for the question on the conservatism in particular language families, I find that the proportion of features that can be reconstructed as ‘present’ or ‘absent’ at the proto-language level does not vary substantially across language families (within 3% for ‘present’ and 5% for ‘absent’ at 95% probability), i.e. any of the five proto-languages can be reconstructed equally well. If we lower the probability boundary to 75%, then Turkic with 38% of features reconstructable as ‘present’ stands out slightly as being more conservative than other language families with their 33.3%–34.5% of reconstructable features.

In §3, I discussed categories of features that can be reconstructed well (≥ 95%) for each of the language families, approximately one fifth of the features. It is not surprising that most categories that are best reconstructable in all language families also are the categories that evolve at the slowest rate in most cases (‘argument marking non-core’ and ‘derivation’ among functional categories, ‘pronoun’ among parts of speech). We see an unexpected pattern in the language levels: we would assume that relatively few features operating on the level of the noun phrase should be reconstructed well to the proto-language level, because this is the fastest evolving category, but in terms of reconstruction, the category ‘noun phrase’ can well compete with ‘word’.

Apart from well-reconstructable features, there is one fifth of features that can be reconstructed rather poorly (around 0.5 probability of being ‘present’). Poor reconstruction of features is partly indirectly due to an unequal distribution of presence and absence across branches, e.g. present in 30 languages out of

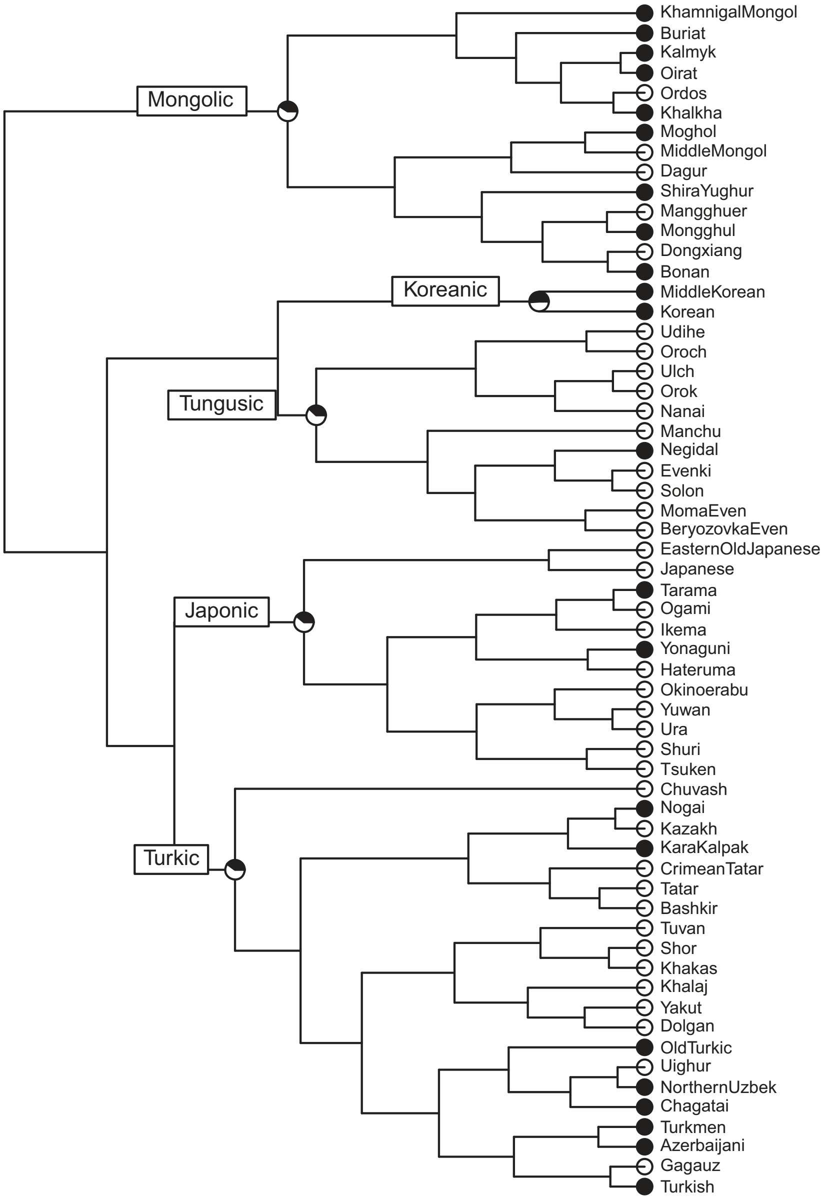

Figure 10. Distribution of the feature TE027 on the tree. The circles on the tips indicate feature values for individual languages and the circles on the nodes the reconstructed ancestral states for five proto-languages: 'present' in black, 'absent' in white.

60 languages. Features with such an unequal distribution comprise, among others, marking of comitative and conjunction, productive plural marking, future tense marking on the verb, tense marking by an auxiliary verb, etc. There might also be a joint effect of a random distribution on the tree and a high amount of missing data. Such features include a class of patient-labile verbs, an inclusory construction, a morpho-syntactic distinction between controlled vs. uncontrolled events, etc. These features are codable as 'present' only for several languages and are either absent in other languages or there is no information on their presence available.[5] Therefore, if a feature has only 50/50 of '1' and

[5] The decision on coding the feature as absent versus unknown is often made based on the overall quality of the grammar description work. For example, if there is no information on the feature in a grammar sketch of 30 pages, then the feature usually receives a '?'. If the feature is not mentioned in an extensive 600 pages grammar, then the feature is most probably absent.

'?' values inside a family, the reconstruction will not be only 1—it will incorporate the uncertainty by allowing some chance of the feature being '0' at the node level: the model assumes polymorphic '0/1' values at the tips with '?'. This way, we are not making false assumptions of full reconstructability of a feature if there are several '1' values and much missing data.

From the computational side, poor reconstruction is an effect of the interaction between the evolutionary rate and the values at the tips: if the model itself is not sure about the ancestral state, like the 'equal rates' model, it lets more information flow into the ancestral state. With the 'all rates differ' model, the reconstruction is less susceptible to slight differences in tip values [52, p. 476]. We will take feature TE027 'Can 1PL marker be augmented by a collective plural marker?' as an illustrative example for the interaction between the model assumptions and the tips (see figure 10 for the distribution of the feature on the tree and the reconstructed ancestral states). It is probable that in three language families (Turkic, Tungusic and Japonic) the feature was innovated on some branches, whereas it is present in both Koreanic and most Mongolic languages and we would intuitively think it was lost in some of the Mongolic languages. Nevertheless, this feature is reconstructed with high certainty in none of the families: even in Mongolic and Koreanic the probability of this feature being '1' at the proto-language level is not higher than 0.51.

Why is the feature TE027 reconstructed for Proto–Koreanic almost as a 50/50 chance of being present if both Middle Korean and Modern Korean have it and why is its probability of being 'present' so low for Mongolic, even though most Mongolic languages seem to have an additional plural marker on the 1PL personal pronoun? It is most probably due to the interaction effect between the evolutionary rate and the tip values described above: the ancestral state reconstruction is informed about the evolutionary rate of this feature, which is relatively high, and the chosen model is *ARD*, therefore the impact of the tip values is moderate. We get thus a rather high uncertainty in the reconstructed node value for all families. In this case, the model is not necessarily useless: it is warning us to reconstruct this fast evolving feature with great care.

The combination of reconstructed states and the rate of change of particular features can allow further research to contextualize the rates in time, if there is enough information on the age of the proto-language. For example, the age of Proto-Turkic was estimated to be around 2100 years before present [5]. In the first step, one extracts the features reconstructed as 'present' in Proto–Turkic (95% probability of being 'present' or higher). In the next step, one calculates the distance between Proto-Turkic and its children languages on these traits. This procedure provides us with a number of differences between Proto–Turkic and each child language.

5. Conclusion

Structural features as another tool for gaining information on the relationships between languages are gaining importance in the field of historical linguistics. In order for structural features to be competitive, they need to have a comparable performance for reconstructing ancient relationships (i.e. stability) as basic vocabulary does. We can test this performance by analysing the dynamics of change in structural features, best measured as the phylogenetic signal and the rate of change. Even though the study presents results on five language families (Turkic, Tungusic, Mongolic, Japonic and Koreanic), the type of data and the transparent methodology make it possible for the results to be replicated on other language families to obtain a cross-linguistically stable set of structural features. Extracting the features with a high phylogenetic signal and evolving at a slow rate would enable us to compare the performance of the most stable vocabulary with the most stable structural features instead of a random set of features. This feature set can then be applied for testing hypotheses about language history on relatively equal terms with basic vocabulary.

Data accessibility. The scripts used in the study are available in the following GitHub repository: https://github.com/NataliiaHue/stability. The data are provided in electronic supplementary material [53]. The complete database of structural features with sources, comments and examples is available on https://doi.org/10.5281/zenodo.5720838 and the code for the analysis and other materials on https://doi.org/10.5281/zenodo.6257956.

Competing interests. I declare I have no competing interests.

Funding. The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 646612) granted to Martine Robbeets and from the Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History.

Acknowledgements. I am grateful to Simon Greenhill for his valuable input on the implementation of analysis, targeted problem solving and his continued general support. I also thank Volker Gast and Jakob Lesage for their insightful comments on the paper, Robert Forkel and Christoph Rzymski for their help in data formatting and curation and James Boyko for his help with the corHMM package.

References

1. Gray RD, Atkinson QD. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)
2. Gray RD, Drummond AJ, Greenhill SJ. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)
3. Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert R, Gray RD, Verkerk A. 2018 A Bayesian phylogenetic study of the Dravidian language family. *R. Soc. Open Sci.* **5**, 171504. (doi:10.1098/rsos.171504)
4. Robbeets M, Bouckaert R. 2018 Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. *J. Lang. Evol.* **3**, 145–162. (doi:10.1093/jole/lzy007)
5. Savelyev A, Robbeets M. 2020 Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *J. Lang. Evol.* **5**, 39–53. (doi:10.1093/jole/lzz010)
6. Dunn MJ, Terrill A, Reesink GP, Foley RA, Levinson SC. 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075. (doi:10.1126/science.1114615)
7. Reesink G, Singer R, Dunn M. 2009 Explaining the linguistic diversity of Sahul using population models. *PLoS Biol.* **7**, e1000241. (doi:10.1371/journal.pbio.1000241)
8. Reesink G, Dunn M. 2012 Systematic typological comparison as a tool for investigating language history. In *Melanesian languages on the edge of Asia: challenges for the 21st Century* (eds N Evans, M Klamer), pp. 34–71. Honolulu, HI: University of Hawai`i Press.
9. Wichmann S, Holman EW. 2009 *Temporal stability of linguistic typological features*. Munich, Germany: Lincom.
10. Dediu D. 2010 A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proc. R. Soc. B* **278**, 474–479. (doi:10.1098/rspb.2010.1595)
11. Dediu D, Levinson SC. 2012 Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE* **7**, e45198. (doi:10.1371/journal.pone.0045198)
12. Greenhill SJ, Wu C-H, Hua X, Dunn M, Levinson SC, Gray RD. 2017 Evolutionary dynamics of language systems. *Proc. Natl Acad. Sci. USA* **114**, E8822–E8829. (doi:10.1073/pnas.1700388114)
13. Gray RD, Bryant D, Greenhill SJ. 2010 On the shape and fabric of human history. *Phil. Trans. R. Soc. B* **365**, 3923–3933. (doi:10.1098/rstb.2010.0162)
14. Swadesh M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)
15. Tadmor U. 2009 The Leipzig–Jakarta list of basic vocabulary. In *Loanwords in the world's languages: a comparative handbook* (eds M Haspelmath, U Tadmor), pp. 68–75. Berlin, Germany: Walter de Gruyter.
16. Nichols J. 1992 *Linguistic diversity in space and time*. Chicago, IL: University of Chicago Press.
17. Dryer MS, Haspelmath M, editors. 2013 *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. See http://wals.info/.
18. Greenhill SJ, Atkinson QD, Meade A, Gray RD. 2010 The shape and tempo of language evolution. *Proc. R. Soc. B* **277**, 2443–2450. (doi:10.1098/rspb.2010.0051)
19. Skirgård H *et al*. Submitted. Grambank data reveal global patterns in the structural diversity of the world's languages.
20. Revell LJ, Harmon LJ, Collar DC. 2008 Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* **57**, 591–601. (doi:10.1080/10635150802302427)
21. Dunn M, Greenhill SJ, Levinson SC, Gray RD. 2011 Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* **473**, 79–82. (doi:10.1038/nature09923)
22. Zhou K, Bowern C. 2015 Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proc. R. Soc. B* **282**, 20151278. (doi:10.1098/rspb.2015.1278)
23. Haynie HJ, Bowern C. 2016 Phylogenetic approach to the evolution of color term systems. *Proc. Natl Acad. Sci. USA* **113**, 13 666–13 671. (doi:10.1073/pnas.1613666113)
24. Carling G, Cathcart C. 2021 Reconstructing the evolution of Indo-European grammar. *Language* **97**, 561–598. (doi:10.1353/lan.2021.0047)
25. Robbeets M. 2015 *Diachrony of verb morphology: Japanese and the Transeurasian languages*, vol. 291. Berlin, Germany: Walter de Gruyter GmbH.
26. Starostin SA, Dybo A, Mudrak O, Gruntov I. 2003 *Etymological dictionary of the Altaic languages*. Leiden, The Netherlands: Brill.
27. Vovin A. 2005 The end of the Altaic controversy (in memory of Gerhard Doerfer). *Central Asiatic J.* **49**, 71–132.
28. Doerfer G. 1974 Ist das Japanische mit den altaischen Sprachen verwandt? *Z. Dtsch. Morgenl. Ges.* **124**, 103–142.
29. Hammarström H, Forkel R, Haspelmath M, Bank S. 2020 *Glottolog 4.2.1*. Jena: Max Planck Institute for the Science of Human History. See https://glottolog.org/ (accessed 3 June 2020).
30. Robbeets M. 2017 The Transeurasian languages. In *The Cambridge handbook of areal linguistics*, pp. 586–626. Cambridge, UK: Cambridge University Press.
31. Johanson L. 2006 Turkic languages. In *Encyclopedia of language and linguistics*, vol. 13, 2nd edn (ed. K Brown), pp. 161–164. Amsterdam, The Netherlands: Elsevier.
32. Rybatzki V. 2003 Intra-Mongolic taxonomy. In *The Mongolic languages*. Routledge Family Series (ed. J Janhunen), pp. 362–390. London, New York: Routledge.
33. Pellard T. 2009 Ōgami: Éléments de description d'un parler du Sud des Ryūkyū. PhD thesis, École des Hautes Études en Sciences Sociales, Paris, France.
34. Doerfer G. 1978 Some classification problems of Tungus. *Tungusica* **1**, 1–26.
35. Whaley L. 2012 Deriving insights about Tungusic classification from derivational morphology. In *Copies versus cognates in bound morphology* (eds M Robbeets, L Johanson), pp. 395–409. Leiden, The Netherlands: Brill.
36. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537. (doi:10.1371/journal.pcbi.1003537)
37. Penny D, McComish BJ, Charleston MA, Hendy MD. 2001 Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* **53**, 711–723. (doi:10.1007/s002390010258)
38. Drummond AJ, Suchard MA. 2010 Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8**, 1–12. (doi:10.1186/1741-7007-8-114)
39. Fritz SA, Purvis A. 2010 Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24**, 1042–1051. (doi:10.1111/j.1523-1739.2010.01455.x)
40. R Core Team. 2020 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org/.
41. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2018 Caper: comparative analyses of phylogenetics and evolution in R. R Package Version 1.0.1. See https://CRAN.R-project.org/package=caper.
42. Beaulieu J, O'Meara B, Oliver J, Boyko J. 2020 CorHMM: hidden Markov models of character evolution. R Package Version 2.1. See https://CRAN.R-project.org/package=corHMM.
43. Maddison WP, Midford PE, Otto SP. 2007 Estimating a binary character's effect on

speciation and extinction. *Syst. Biol.* **56**, 701–710. (doi:10.1080/10635150701607033)

44. FitzJohn RG, Maddison WP, Otto SP. 2009 Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* **58**, 595–611. (doi:10.1093/sysbio/syp067)

45. Beaulieu JM, O'Meara BC, Donoghue MJ. 2013 Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Syst. Biol.* **62**, 725–737. (doi:10.1093/sysbio/syt034)

46. Arnold L. 2012 Investigations in item stability: in pursuit of the optimal meaning list for use in the initial stages of the comparative method. Master's thesis, The University of Edinburgh, Edinburgh, UK.

47. Bybee J, Thompson S. 1997 Three frequency effects in syntax. In *Proc. of the 23rd Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, vol. 23, pp. 378–388. (doi:10.3765/bls.v23i1.1293)

48. Diessel H. 2007 Frequency effects in language acquisition, language use, and diachronic change. *New Ideas Psychol.* **25**, 108–127. Modern Approaches to Language. (doi:10.1016/j.newideapsych.2007.02.002)

49. Nichols J. 1993 Diachronically stable structural features. In *Historical linguistics 1993* (ed. Hennig Andersen), pp. 337–356. Amsterdam, The Netherlands: John Benjamins. (doi:10.1075/cilt.124.27nic)

50. Hoijer H. 1956 Lexicostatistics: a critique. *Language* **32**, 49–60. (doi:10.2307/410652)

51. Oswalt RL. 1971 Towards the construction of a standard lexicostatistic list. *Anthropol. Linguist.* **13**, 421–434.

52. Boyko JD, Beaulieu JM. 2021 Generalized hidden Markov models for phylogenetic comparative datasets. *Methods Ecol. Evol.* **12**, 468–478. (doi:10.1111/2041-210X.13534)

53. Hübler N. 2022 Phylogenetic signal and rate of evolutionary change in language structures. Figshare.

4. Modelling admixture across language levels to evaluate deep history claims

Author's contribution

This chapter is published as a research article in the Journal of Language Evolution *(https://doi.org/10.1093/jole/lzad002)* in March 2023. It presents additional analyses on the data collected for the two precious chapters. Nataliia Hübler designed and performed research, Nataliia Hübler and Simon Greenhill analyzed the data and wrote the manuscript.

The research leading to this article 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. The authors 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. They also thank two anonymous reviewers for helpful discussion.

The data are available at:
https://doi.org/10.5281/zenodo.7135936.
The scripts and other materials are available at:
https://doi.org/10.5281/zenodo.7515103.

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

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

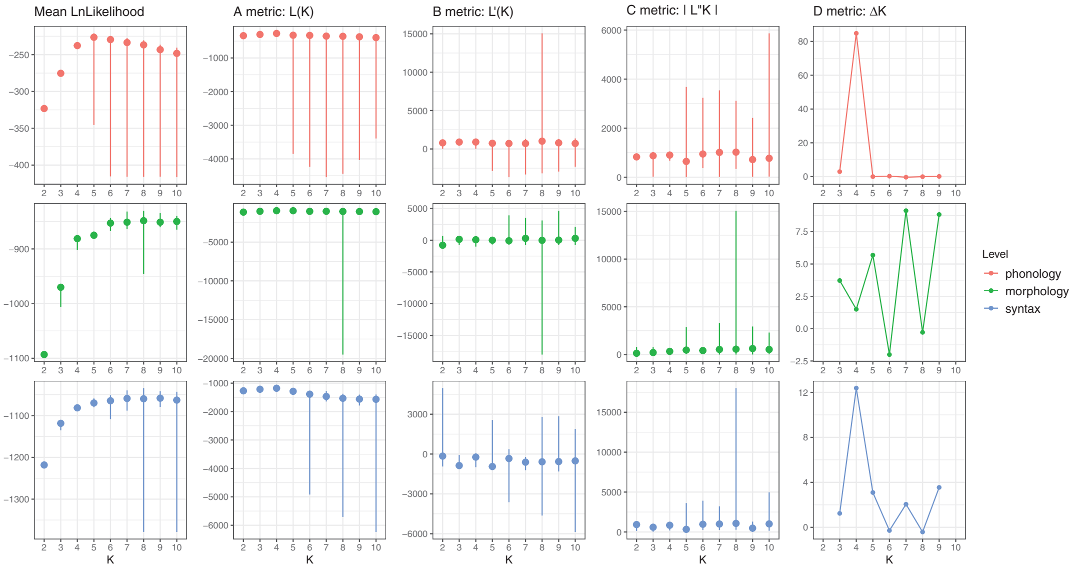

Figure 1 Variation in the log likelihood of K from 2 to 10 across 50 runs and three language levels. The bars indicate the whole range of values (from minimum to maximum value) and the points indicate the median value. Each row represents a language level: phonology, morphology and syntax. Each column 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).

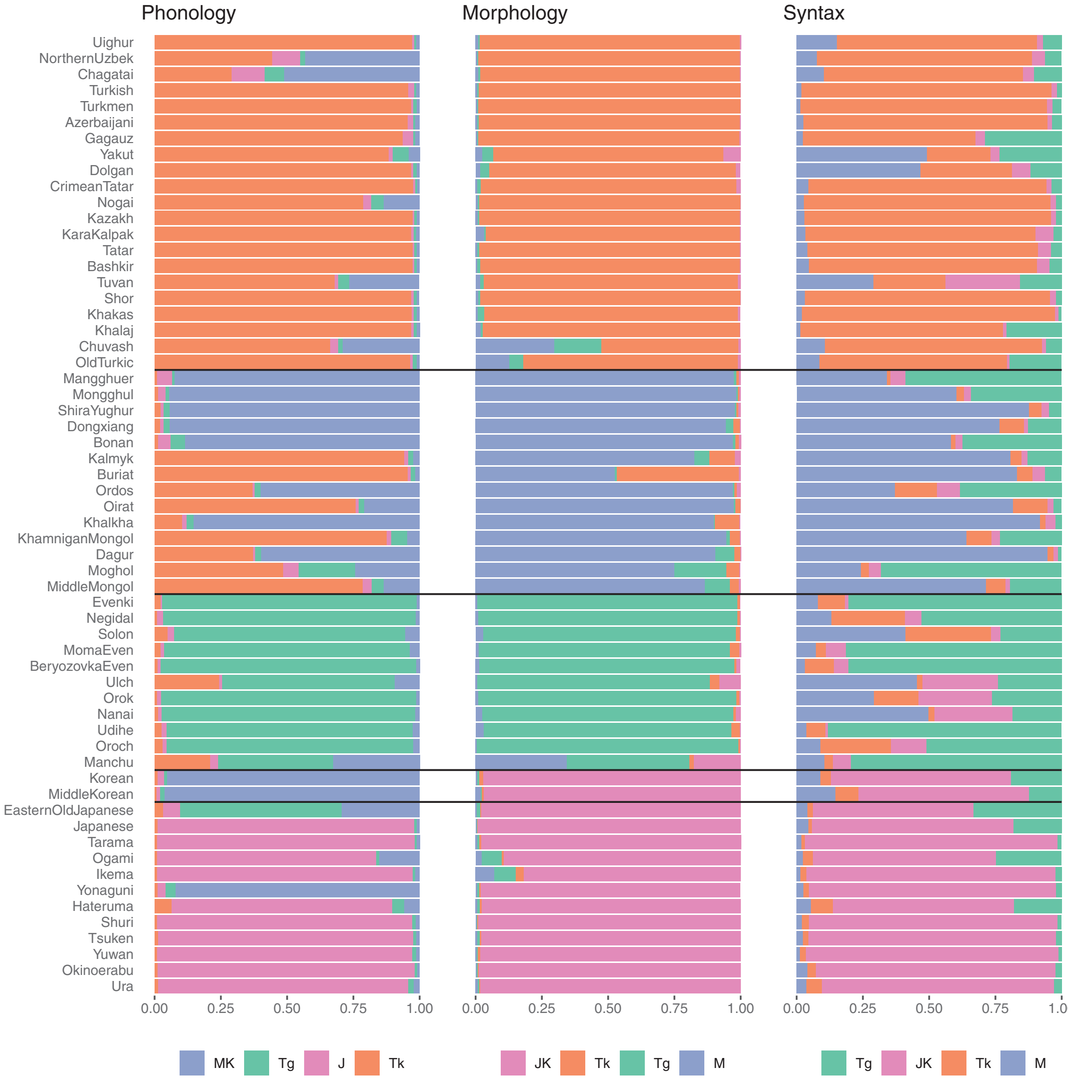

Figure 2 Population structure at $K = 4$. Each row corresponds to a language and each column corresponds to a language level. Languages appear in the order of 1) language families, divided by a black horizontal line: Japonic (from Ura to Eastern Old Japanese), Koreanic (Middle Korean and Korean), Tungusic (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

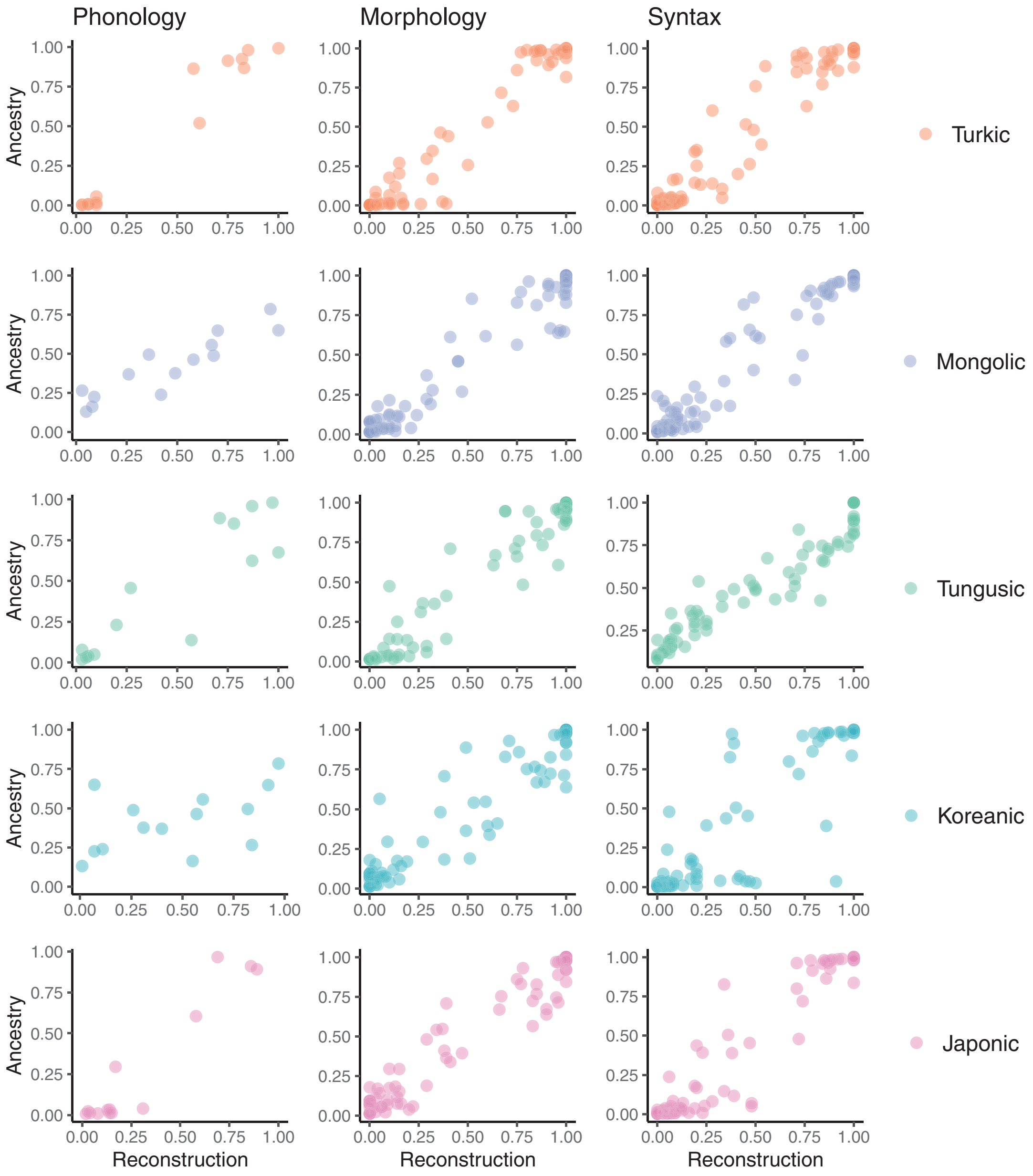

Figure 3 Estimated feature frequencies at *K* = 4 shown as proportions. Abbreviations: Tk = Turkic, Tg = Tungusic, M = Mongolic, K = Koreanic, J = Japonic. The horizontal bar corresponding to each feature consists of feature frequencies (presence) in each of the four assumed ancestries. Each frequency 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

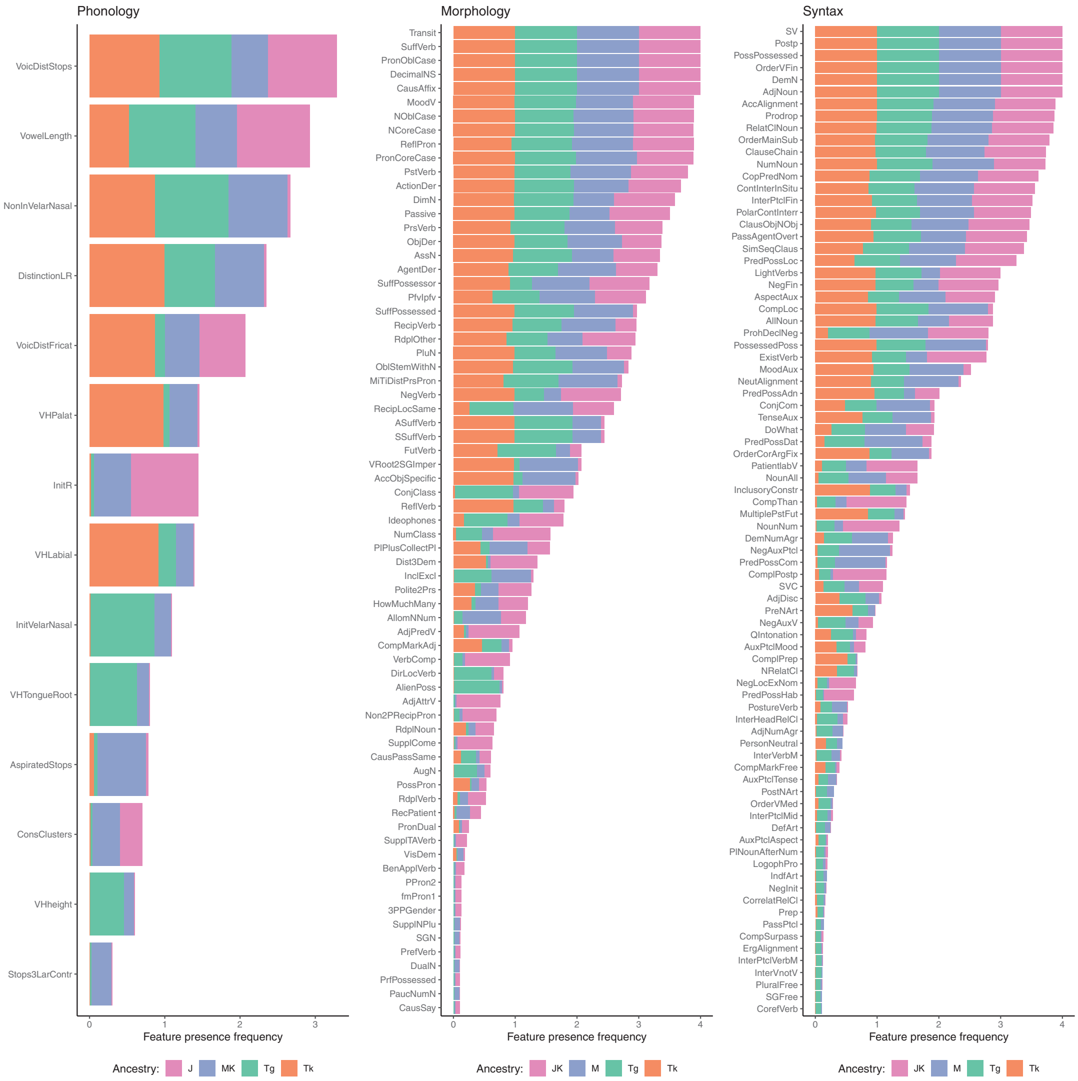

Figure 4 Reconstructability of features as being present/absent in the proto-language, following the results of ancestral state reconstruction by Hübler (2022), vs. contribution of features to the ancestries. The reconstruction of features corresponds to their contribution to respective ancestries: if 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.

5. Discussion

5.1 General discussion

One of the aims of historical linguistics is to understand the relationships between languages and to classify them into language families. There are over 200 language families and over 150 language isolates, i.e. languages with no attested/living "relatives" (Hammarström et al. 2020). Historical linguists have put forward hypotheses about possible genealogical relationships between language families and language isolates, but have encountered difficulties providing the necessary evidence in support of these relationships. One of the challenges lies in the restricted time depth, after which we cannot reliably test hypotheses about the relationships between languages. For most of the attested language families, basic vocabulary was used to establish the relatedness of the languages belonging to them. Over time, the number of changes that accumulate in the basic vocabulary grows, and after several thousand years it is no longer possible to track genealogical relationships between languages. Some researchers suggest that using structural features might "push back the time barrier" (Gray 2005, Greenhill et al. 2010) and allow us to establish relationships even beyond language family level. However, in contrast to basic vocabulary, which was pre-selected for its resistance to change, the stability of structural features is not yet fully clarified. Therefore, investigating the stability of structural features is a necessary step towards testing deep inter-family relations.

Before safely using structural features to test relationships between language families, we have to investigate the amount of phylogenetic signal in them, the rate, at which they evolve, and show that they point at genealogical and not areal groupings. Alternatively, we have to be able to prove the opposite and discard structural features as a source of information on the genealogy of languages. In this thesis, I used the structural data collected for the so-called "Transeurasian" languages (comprising Turkic, Mongolic, Tungusic, Koreanic and Japonic language families) to tackle the question of the stability of structural features. First, I described the typological profile of these languages and tested the tree model in correctly determining the relationships between languages (Chapter 2). The tree model did not recover the relationships between languages accurately enough when applied to structural data. Next, I measured the phylogenetic signal and the rate of evolutionary change in structural features (Chapter 3) and showed the differences in stability across language levels, parts of speech and functional categories. Based on the results of this chapter, we cannot state that all structural features are equally good at reflecting genealogical relationships, but that there is a set of structural features that are more stable than others. Additionally, I have shown (Chapter 4) that structural features covering the morphological level point at genealogical relationships between languages and have low levels of diffusion. These results provide ground for further investigation of the external relationships between not only Altaic/Transeurasian, but also other language families.

5.2 Typological profile of the Transeurasian languages

To recap the structural similarities between the Transeurasian languages, a typological profile of these languages is presented in Chapter 2. The data collected for this chapter were extended by several languages and converted to the cross-linguistic data format, published as an openly accessible data set on zenodo (Hübler and Forkel 2022), and added to the Grambank database (Hammarström et al. 2017), which will be accessible online with the release of the database and will allow large-scale cross-linguistic comparisons.

Phonologically, these languages are characterised by a length distinction in vowels, vowel harmony (palatal and tongue root) and absence of velar nasals and consonant clusters word-initially (predominantly). In terms of morphology, these languages make frequent use of morphological plural marking (restricted to animate nouns in Japonic), derivational morphology, core and oblique case marking, nominal reduplication, genitive marking of the possessor, possessive suffixes, ablative marking of the standard of comparison, passive and causative suffixes. Most often, the verb agrees with the S/A argument in person and number (not in Japonic and Koreanic), the pronoun is omitted and the noun does not take plural number if combined with a numeral. The clausal word order is SOV (apart from several languages allowing additionally SVO order) and the order in noun phrases is modifier - head. Interrogation is most commonly marked by a clause-final particle. In complex clauses, only the verb in the main clause carries the TAM-marking. In predicative possession, the possessor is marked either by LOC/DAT (lit.

'The cat is on/to me.') or as an adnominal possessor (lit. 'My cat exists.'), except for Japonic and Koreanic, which make use of a 'habeo'-verb.

The typological profile suggests the division of the Transeurasian languages into Altaic (Turkic, Mongolic, Tungusic) and Japonic/Koreanic languages. A phylogenetic tree of the Transeurasian languages[1], achieved by Bayesian tree building as implemented in BEAST (Bouckaert et al. 2014), supports the division between the Altaic and Japono-Koreanic languages (they have posterior probabilities of 0.83 and 1.0 respectively), but shows low resolution within the individual language families. For example, Mongolic and Turkic languages do not appear as two separate branches, but rather as one poorly resolved branch; the support for most internal branches is low (median of approx. 0.3, standard deviation of 0.3). It stands to reason that these methods do not advance our understanding of the relationships between these languages either due to the nature of the languages (the evolution is not tree-like) or to the nature of the data (structural features are ill-suited for building phylogenies). The poor performance of the tree model on these data raises the question of the compatibility of the methods and the data and suggests that we have to investigate the stability of structural features more thoroughly and consider alternative methods.

5.3 Phylogenetic signal and evolutionary rate

We have seen that a combination of structural features as data and a phylogenetic tree as a model might be not the approach of choice to describe either the diversification of or the relationships between the Transeurasian languages. The practical failure has its root in the theory: the data do not conform to the assumptions of the tree model. First, the debate on the relatedness of the Transeurasian languages has not been resolved as of today (Robbeets et al. 2021, Tian et al. 2022). Second, we do not understand the evolutionary dynamics of structural features well enough to use them as the only data type when constructing phylogenies.

The second study (Chapter 3 of this thesis) tackles the question of the stability of structural features by measuring the phylogenetic signal and the evolutionary rate in structural features. It might seem redundant to study rate of gain, rate of loss and phylogenetic signal separately, but the results show that there are differences in how fast features are lost, gained and their phylogenetic signal. Previous studies (Nichols 2003, Carling and Cathcart

[1] A phylogenetic tree always has the relatedness of the languages it is applied to as an assumption. This assumption follows the views of Robbeets and Bouckaert (2018) on the genealogical relationship between the languages.

2021) suggested that features can be divided into several categories, based on their stability "pattern": apart from highly stable (rarely lost and rarely gained) and highly unstable (frequently lost and frequently gained) features, there are also "recessive" features, which are unlikely to be inherited and are often lost (high rate of loss and low rate of gain), and "attractive" features, which are likely to be inherited, are rarely lost and more often gained (low rate of loss and high rate of gain). In the current language sample and linguistic area, there are more features that tend to be recessive (15) than those that tend to be attractive (2). Articles, plural marking on nouns preceded by a numeral, a logophoric pronoun, marking of direct evidence, a bound comparative degree marker and other features are rather recessive in the current language sample. A vowel length distinction and a difference in marking polar and content questions are rather attractive features.

The results show that more than half of the features (63%) have a phylogenetic signal and evolve at a slow rate (68% are lost at a slow rate and 75% are gained at a slow rate). Of all features, 19%–22% can be reconstructed as "present" and 21%–26% as "absent" in the proto-language (with 95% probability). There are notable differences in stability across the functional categories, parts of speech and language levels: features on core argument marking (flagging and indexing), derivation and valency are more stable than those on interrogation and quantification, features targeting nouns and pronouns are more stable than those targeting articles and demonstratives and features operating on the phonological and morphological levels are more stable than those operating on the level of NP and clause. Even though there are no grounds to postulate these features as cross-linguistically stable, the future research will be able to replicate these results on other language families by using the open source code provided along with the article. The results presented in this chapter will be directly comparable to those of the future studies if these also use the features stored in the Grambank database, which now covers almost 2,500 languages coded for 195 features. The only domain that cannot be compared directly is phonology: Grambank does not include any phonological features, which will have to be coded separately. Comparing stability of features across multiple language families will bring us closer to a basic set of structural features (similar to a basic vocabulary list), which can be used to test hypotheses about deep language relationships.

5.4 Admixture

The third study, presented in Chapter 4, focused on the performance of structural features in replicating genealogical relationships on the language family level. I used a method favoured in population genetics, the Bayesian clustering algorithm STRUCTURE. I compared the performance of structural features across three language levels: phonology, morphology and syntax, and found differences not only in the precision in the assignment of languages to the respective ancestries (comparable to language families), but also in the amount of admixture at each of the levels.

The results show high levels of admixture in syntax for most languages and thus provide tentative support of the findings of previous research, which suggest that borrowing plays a more important role and is more frequent in syntax than in other language domains (Thomason and Kaufman 1988, Campbell 2013). Although we see admixture, most probably, as a result of frequent borrowing in syntax, the nature of it remains unclear. Ringe (2013) suggests that changes in phonology and morphology might trigger changes in syntax – this is a statement that can only be tested in smaller-scaled studies on individual language families and languages. Phonological features take in an intermediate position between syntax and morphology in terms of levels of admixture, but we cannot draw ultimate conclusions on their specific propensity for borrowing because of the small number of features (only 14). Moreover, we would rather have to admit that phonological features perform surprisingly well given the low number of features.

The central conclusion of the study is that morphological features convey the most precise information on the genealogical relations between languages: most of the languages are correctly assigned to their respective language families. They also showed comparatively low levels of admixture, once again supporting their potential usability for testing hypotheses about deep relationships between languages.

One of the advantages of admixture model implemented in STRUCTURE is that the languages do not need to be related – it can thus be applied to a set of languages not previously recognised as a language family. Borrowing does not constitute a problem for running STRUCTURE or interpreting its results, which makes it especially suitable for structural data. One of the drawbacks of this approach is that STRUCTURE can only interpret the data it was given the best way possible: if we do not include languages from other language families, obviously, we cannot see their influence on the languages of our sample and might misinterpret the admixture as resulting from language contact *within* our language sample.

5.5 Conclusions and outlook

In this thesis, I defined the typological profile of the Transeurasian languages based on the information from grammatical descriptions of the respective languages and built a phylogenetic tree with Bayesian methods based on structural features coded for Transeurasian languages as data. The phylogenetic tree did not significantly advance our understanding of the relationships and the history of the languages in question.

A necessary prerequisite for using structural features as data for investigating genealogical relationships between languages is the understanding of their stability. As a way forward, I investigated the stability of structural features by measuring the phylogenetic signal they entail and the rate, at which they are gained and lost. I found differences in stability across language levels, functional categories and parts of speech and determined the most stable categories. Overall, features targeting phonologically bound elements, as is the case in morphology, tend to be especially stable. Furthermore, I reconstructed ancestral states of structural features at the proto-language level for Proto-Turkic, Proto-Mongolic, Proto-Tungusic, Proto-Koreanic and Proto-Japonic.

As an alternative to the tree model, I applied an admixture model to three feature sets and compared the performance of phonological, morphological and syntactic features in their assignment of languages to language families. I have shown that admixture provides more accurate results than the tree model and is better suited as method given structural features as data. The combination of morphological features and STRUCTURE as method shows a high potential for further application in investigating relationships between language families. The results of this thesis can also be used as a ground for further research on the stability of structural features with data from other language families and on the deep relationships between languages or language families, since the code has been made publicly available and the data is stored in a standardised way.

6. Bibliography

Atkinson, Quentin D, and Russell D Gray. 2005. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology* 54:513–526.

Baskakov, A, Nikolaj. 1981. *Altaiskaja sem'ja jazykov i ee izučenie*. Moscow: Nauka.

Beaulieu, Jeremy, Brian O'Meara, Jeffrey Oliver, and James Boyko. 2020. *corhmm: Hidden markov models of character evolution*. URL `https://CRAN.R-project.org/package=corHMM`, r package version 2.1.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology* 10:e1003537.

Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences* 279:4590–4595.

Bowern, Claire, and Bethwyn Evans. 2015. *The Routledge Handbook of Historical Linguistics*. Routledge.

Bulatova, Nadezhda Yakovlevna, and Lenore Grenoble. 1999. *Evenki*, volume 141 of *Languages of the World/Materials*. Lincom Europa.

Calude, Andreea S., and Annemarie Verkerk. 2016. The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution* 1:91–108.

Campbell, Lyle. 2003. How to show languages are related: Methods for distant genetic relationship. *The handbook of Historical Linguistics* 19:262–282.

Campbell, Lyle. 2013. *Historical linguistics: An introduction*. Edinburgh University Press, 3 edition.

Carling, Gerd, and Chundra Cathcart. 2021. Reconstructing the evolution of Indo-European grammar. *Language* 97.

Darwin, C. 1871. *The descent of man*. London: Murray.

Dediu, Dan, and Stephen C Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PloS One* 7:e45198.

Dunn, Michael J., Angela Terrill, Ger P. Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072 – 2075.

Dybo, Anna V, and George S Starostin. 2008. In defense of the comparative method, or the end of the Vovin controversy (in memory of Sergei Starostin). In *Aspekty komparativistiki [aspects of the comparative studies]*, volume 3, 119–258. Moscow: Rossijskij gosudarstvennyj gumanitarnyj institut [Russian state institute of humanities].

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.

Forkel, Robert, Johann-Mattis List, Simon J Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A Kaiping, and Russell D Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data* 5:1–10.

François, Alexandre. 2015. Trees, waves and linkages: models of language diversification. In *The Routledge Handbook of Historical Linguistics*, 179–207. Routledge.

Fritz, Susanne A, and Andy Purvis. 2010. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24:1042–1051.

Gray, Russell. 2005. Pushing the time barrier in the quest for language roots. *Science* 309:2007–2008.

Gray, Russell D., David Bryant, and Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365:3923–3933.

Gray, Russell D, Alexei J Drummond, and Simon J Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.

Greenberg, Joseph H. 1966. *Universals of language*. The MIT Press, second edition edition.

Greenberg, Joseph H. 1978. Diachrony, synchrony and language universals. In *Universals of Human Language*, ed. Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, volume III: Word Structure, 47–82. Stanford University Press.

Greenhill, Simon J, Quentin D Atkinson, Andrew Meade, and Russell D Gray. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences* 277:2443–2450.

Greenhill, Simon J, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C Levinson, and Russell D Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114:E8822–E8829.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* 112:13296–13301.

Guy, Jacques BM. 1995. The incidence of chance resemblances on language comparison. *Anthropos* 223–228.

Hale, Mark. 2003. Neogrammarian sound change. *The Handbook of Historical Linguistics* 19:343–368.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog 4.2.1. Max Planck Institute for the Science of Human History. URL `https://glottolog.org/accessed2020-06-03`.

Hammarström, Harald, Hedvig Skirgård, Jeremy Collins, Hannah Haynie, Alena Witzlack, Stephen C. Levinson, Russell Gray, Jakob Lesage, Richard Kowalik, Robert Forkel, Linda Raabe, Suzanne van der Meer, Jana Winkler, Ger Reesink, Tessa Yuditha, Patience Epps, Luise Dorenbusch, Hilário de Sousa, Cheryl Akinyi Oluoch, Claire Bowern, Giada Falcone, Eloisa Ruppert, Martin Haspelmath, Nataliia Hübler, Karolin Abbas, Jesse Peacock, Hugo de Vos, Olga Krasnoukhova, Robert Borges, Stephanie Petit, Michael Dunn, Carolina Kipf, Jay Latarche, Nancy Bakker, Roberto Herrera, Johanna Nickel, Giulia Barbos, Kristin Sverredal, Tim

Witte, Ruth Singer, Michael Dunn, Janina Klingenberg, Sören Danielsen, Swintha Pieper, and Damian Blasi. 2017. *Grambank: A world-wide typological database. Electronic database under development*. Max Planck Institute for the Science of Human History.

Heggarty, Paul, Warren Maguire, and April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:3829–3843.

Heinrich, Patrick, Shinsho Miyara, and Michinori Shimoji, ed. 2015. *Handbook of the Ryukyuan languages: History, structure, and use*, volume 11. Walter de Gruyter GmbH & Co KG.

Huson, Daniel H., and David Bryant. 2010. Splitstree4.

Hübler, Nataliia, and Robert Forkel. 2022. hueblerstability. URL `https://doi.org/10.5281/zenodo.7135936`.

Janhunen, Juha, ed. 2003. *The Mongolic Languages*. Language Family Series. London and New York: Routledge.

Johanson, Lars. 2020. The classification of the Turkic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 104–114. Oxford University Press.

Johanson, Lars, and Martine Irma Robbeets. 2010. *Transeurasian verbal morphology in a comparative perspective: genealogy, contact, chance*, volume 78. Otto Harrassowitz Verlag.

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *The Royal Society Open Science* 5:171504.

Litsios, Glenn, and Nicolas Salamin. 2012. Effects of phylogenetic signal on ancestral state reconstruction. *Systematic Biology* 61:533–538. URL `https://doi.org/10.1093/sysbio/syr124`.

Macklin-Cordes, Jayden L, Claire Bowern, and Erich R Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38:210–258.

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.

Nichols, Johanna. 1993. Diachronically stable structural features. *Amsterdam studies in the theory and history of linguistic science series* 4:337–337.

Nichols, Johanna. 2003. Diversity and stability in language. In *The handbook of historical linguistics*, ed. Brian D. Joseph and Richard D. Janda, 283–310. Backwell Publishing.

Norvik, Miina, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen, and Outi Vesakoski. 2022. Uralic typology in the light of a new comprehensive dataset. *Journal of Uralic linguistics* 1:1:4–42.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Phillips, Joshua, and Claire Bowern. 2022. Bayesian methods for ancestral state reconstruction in morphosyntax: Exploring the history of argument marking strategies in a large language family. *Journal of Language Evolution* URL `https://doi.org/10.1093/jole/lzac002`.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Reesink, Ger, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biol* 7:e1000241.

Revell, Liam J, Luke J Harmon, and David C Collar. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic biology* 57:591–601.

Ringe, Don. 1999. How hard is it to match CVC-Roots? *Transactions of the Philological Society* 97:213–244.

Ringe, Don. 2013. Syntactic change. In *Historical linguistics: Toward a twenty-first century reintegration*, ed. Don Ringe and Joseph F. Eska, 212–227. Cambridge University Press.

Robbeets, Martine. 2017. The Transeurasian languages. In *The Cambridge Handbook of Areal Linguistics*, 586–626. Cambridge University Press.

Robbeets, Martine. 2020. The classification of the Transeurasian languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 31–39. Oxford University Press.

Robbeets, Martine, and Remco Bouckaert. 2018. Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. *Journal of Language Evolution* 3:145–162.

Robbeets, Martine, Remco Bouckaert, Matthew Conte, Alexander Savelyev, Tao Li, Deog-Im An, Ken-ichi Shinoda, Yinqiu Cui, Takamune Kawashima, Geonyoung Kim, et al. 2021. Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 599:616–621.

Robbeets, Martine Irma. 2005. *Is Japanese related to Korean, Tungusic, Mongolic and Turkic?*, volume 64. Otto Harrassowitz Verlag.

Ronquist, Fredrik. 2004. Bayesian inference of character evolution. *Trends in Ecology & Evolution* 19:475–481.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur* 3:786–787.

Schmidt, Johannes. 1872. *Die Verwandschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Böhlau.

Schwarz, Michal, Ondřej Srba, and Václav Blažek. 2020. A comparative approach to the pronominal system in Transeurasian. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 554–584. Oxford University Press.

Shimoji, Michinori, and Thomas Pellard, ed. 2010. *An introduction to Ryukyuan languages*. Research Institute for Languages and Cultures of Asia and Africa.

Starostin, Sergei A, Anna Dybo, Oleg Mudrak, and Ilya Gruntov. 2003. *Etymological dictionary of the Altaic languages*. Leiden: Brill.

Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino, and Outi Vesakoski. 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, Yuxin Tao, Kongyang Zhu, Guillaume Jacques, Robin J. Ryder, José Andrés Alonso de la Fuente, Anton Antonov, Ziyang Xia, Yuxuan Zhang, Xiaoyan Ji, Xiaoying Ren, Guanglin He, Jianxin Guo, Rui Wang, Xiaomin Yang, Jing Zhao, Dan Xu, Russell D. Gray, Menghan Zhang, Shaoqing Wen, Chuan-Chao Wang, and Thomas Pellard. 2022. Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative. *bioRxiv* URL `https://www.biorxiv.org/content/early/2022/06/12/2022.06.09.495471`.

Verkerk, Annemarie. 2014. The evolutionary dynamics of motion event encoding. Doctoral Dissertation, Radboud University, Nijmegen.

Vovin, Alexander. 2005. The end of the Altaic controversy. In memory of Gerhard Doerfer. *Central Asiatic Journal* 49:71–132.

Weiss, Michael. 2015. The comparative method. In *The Routledge Handbook of Historical Linguistics*, ed. Claire Bowern and Bethwyn Evans, 127–145. London & New York: Routledge.

Whaley, Lindsay J, and Sofia Oskolskaya. 2020. The classification of the Tungusic languages. In *The Oxford Guide to the Transeurasian Languages*, ed. Martine Robbeets and Alexander Savelyev, 80–91. Oxford University Press.

Wichmann, Søren, and Eric W Holman. 2009. *Temporal stability of linguistic typological features*. Lincom Europa.

Wolfram, Walt, and Natalie Schilling-Estes. 2003. Dialectology and linguistic diffusion. In *The Handbook of Historical Linguistics*, ed. Brian D. Joseph and Richard D. Janda, 713–735. Backwell Publishing.

Annexes

Electronic Supplementary Materials for the article "Phylogenetic signal and rate of evolutionary change in language structures"

Nataliia Hübler

Max Planck Institute for the Science of Human History, Jena, Germany
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Table S1: Feature set with feature ID, a short feature name, original feature formulation, part of speech, functional and level (language domain) categorisation.

ID	Feature (short)	Description	PoS	Function	Level
GB020	DefArt	Are there definite or specific articles?	article	deixis	NP
GB021	IndfArt	Do indefinite nominals commonly have indefinite articles?	article	deixis	NP
GB022	PreNArt	Are there prenominal articles?	article	deixis	NP
GB023	PostNArt	Are there postnominal articles?	article	deixis	NP
GB026	AdjDisc	Can adnominal property words occur discontinuously?	not assignable	word order	NP
GB027	ConjCom	Are nominal conjunction and comitative expressed by different elements?	noun/pronoun	argument marking (non-core)	NP
GB028	InclExcl	Is there an inclusive/exclusive distinction?	pronoun	deixis	word
GB030	3PPGender	Is there a gender distinction in independent 3rd person pronouns?	pronoun	deixis	word
GB031	PronDual	Is there a dual or unit augmented form (in addition to plural or augmented) for all person categories in the pronoun system?	pronoun	deixis	word
GB035	Dist3Dem	Are there three or more distance contrasts in demonstratives?	demonstrative	deixis	word
GB037	VisDem	Do demonstratives show a visible-nonvisible distinction?	demonstrative	deixis	word
GB039	AllomNNum	Is there nonphonological allomorphy of noun number markers?	noun	quantification	word
GB041	SupplNPlu	Are there several nouns (more than three) which are suppletive for number?	noun	quantification	word
GB042	SGN	Is there productive overt morphological singular marking on nouns?	noun	quantification	word
GB043	DualN	Is there productive morphological dual marking on nouns?	noun	quantification	word
GB044	PluN	Is there productive morphological plural marking on nouns?	noun	quantification	word
GB046	AssN	Is there an associative plural marker for nouns?	noun	quantification	word
GB047	ActionDer	Is there a productive morphological pattern for deriving an action/state noun from a verb?	noun	derivation	word
GB048	AgentDer	Is there a productive morphological pattern for deriving an agent noun from a verb?	noun	derivation	word

GB049	ObjDer	Is there a productive morphological pattern for deriving an object noun from a verb?	noun	derivation	word
GB057	NumClass	Are there numeral classifiers?	noun	quantification	word
GB059	AlienPoss	Is the adnominal possessive construction different for alienable and inalienable nouns?	noun	possession	word
GB068	AdjPredV	Do core adjectives (defined semantically as property concepts such as value, shape, age, dimension) act like verbs in predicative position?	adjective	modification	word
GB069	AdjAttrV	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?	adjective	modification	word
GB070	NCoreCase	Are there morphological cases for non-pronominal core arguments (i.e. S/A/P)?	noun	argument marking (core)	word
GB071	PronCoreCase	Are there morphological cases for pronominal core arguments (i.e. S/A/P)?	pronoun	argument marking (core)	word
GB072	NOblCase	Are there morphological cases for oblique non-pronominal NPs (i.e. not S/A/P)?	noun	argument marking (non-core)	word
GB073	PronOblCase	Are there morphological cases for oblique independent personal pronouns (i.e. not S/A/P)?	pronoun	argument marking (non-core)	word
GB074	Prep	Are there prepositions?	other	argument marking (non-core)	NP
GB075	Postp	Are there postpositions?	other	argument marking (non-core)	NP
GB079	PrefVerb	Do verbs have prefixes/proclitics, other than those that ONLY mark A, S or O (do include portmanteau: A and S + TAM)?	verb	TAME+	word
GB080	SuffVerb	Do verbs have suffixes/enclitics, other than those that ONLY mark A, S or O (do include portmanteau: A and S + TAM)?	verb	TAME+	word
GB082	PrsVerb	Is there overt morphological marking of present tense on verbs?	verb	TAME+	word
GB083	PstVerb	Is there overt morphological marking on the verb dedicated to past tense?	verb	TAME+	word
GB084	FutVerb	Is there overt morphological marking on the verb dedicated to future tense?	verb	TAME+	word

GB086	PfvIpfv	Is a morphological distinction between perfective and imperfective aspect available on verbs?	verb	TAME+	word
GB089	SSuffVerb	Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?	verb	argument marking (core)	word
GB091	ASuffVerb	Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?	verb	argument marking (core)	word
GB103	BenApplVerb	Is there a benefactive applicative marker on the verb (including indexing)?	verb	valency	word
GB105	RecPatient	Can the recipient in a ditransitive construction be marked like the monotransitive patient?	verb	argument marking (non-core)	word
GB107	NegVerb	Can standard negation be marked by an affix, clitic or modification of the verb?	verb	negation	word
GB108	DirLocVerb	Is there directional or locative morphological marking on verbs?	verb	valency	word
GB110	SupplTAVerb	Is there verb suppletion for tense or aspect?	verb	TAME+	word
GB111	ConjClass	Are there conjugation classes?	verb	TAME+	word
GB113	Transit	Are there verbal affixes or clitics that turn intransitive verbs into transitive ones?	verb	valency	word
GB114	ReflVerb	Is there a phonologically bound reflexive marker on the verb?	verb	valency	word
GB115	RecipVerb	Is there a phonologically bound reciprocal marker on the verb?	verb	valency	word
GB117	CopPredNom	Is there a copula for predicate nominals?	verb	other	clause
GB118	SVC	Are there serial verb constructions?	verb	other	clause
GB119	MoodAux	Can mood be marked by an inflecting word (“auxiliary verb”)?	verb	TAME+	clause
GB120	AspectAux	Can aspect be marked by an inflecting word (“auxiliary verb”)?	verb	TAME+	clause
GB121	TenseAux	Can tense be marked by an inflecting word (“auxiliary verb”)?	verb	TAME+	clause
GB122	VerbComp	Is verb compounding a regular process?	verb	other	word
GB123	LightVerbs	Are there verb-adjunct (aka light-verb) constructions?	verb	other	clause
GB126	ExistVerb	Is there an existential verb?	verb	other	clause
GB127	PostureVerb	Are different posture verbs used obligatorily depending on an inanimate locatum’s shape or position (e.g. ‘to lie’ vs. ‘to stand’)?	verb	other	clause
GB132	OrderVMed	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?	verb	word order	clause
GB133	OrderVFin	Is a pragmatically unmarked constituent order verb-final for transitive clauses?	verb	word order	clause

GB134	OrderMainSub	Is the order of constituents the same in main and subordinate clauses?	not assignable	word order	clause
GB135	ClausObjNObj	Do clausal objects usually occur in the same position as nominal objects?	not assignable	word order	clause
GB136	OrderCorArgFix	Is the order of core argument (i.e. S/A/P) constituents fixed?	not assignable	word order	clause
GB137	NegFin	Can standard negation be marked clause-finally?	verb	negation	clause
GB138	NegInit	Can standard negation be marked clause-initially?	verb	negation	clause
GB139	ProhDeclNeg	Is there a difference between imperative (prohibitive) and declarative negation constructions?	verb	negation	clause
GB140	NegLocExNom	Is verbal predication marked by the same negator as all of the following types of predication: locational, existential and nominal?	verb	negation	clause
GB146	CtrlEvents	Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?	noun	argument marking (core)	other
GB147	Passive	Is there a morphological passive marked on the lexical verb?	verb	valency	word
GB150	ClauseChain	Is there clause chaining?	verb	other	clause
GB151	CorefVerb	Is there an overt verb marker dedicated to signalling coreference or noncoreference between the subject of one clause and an argument of an adjacent clause (“switch reference”)?	verb	argument marking (core)	clause
GB152	SimSeqClaus	Is there a morphologically marked distinction between simultaneous and sequential clauses?	verb	TAME+	clause
GB155	CausAffix	Are causatives formed by affixes or clitics on verbs?	verb	valency	word
GB156	CausSay	Is there a causative construction involving an element that is unmistakably grammaticalized from a verb for ‘to say’?	verb	valency	word
GB158	RdplVerb	Are verbs reduplicated?	verb	TAME+	word
GB159	RdplNoun	Are nouns reduplicated?	noun	quantification	word
GB160	RdplOther	Are elements apart from verbs or nouns reduplicated?	adjective	other	word
GB166	PaucNumN	Is there productive morphological paucal marking on nouns?	noun	quantification	word
GB167	LogophPro	Is there a logophoric pronoun?	pronoun	deixis	clause
GB184	AdjNumAgr	Can an adnominal property word agree with the noun in number?	adjective	quantification	NP
GB185	DemNumAgr	Can an adnominal demonstrative agree with the noun in number?	demonstrative	quantification	NP

GB187	DimN	Is there any productive diminutive marking on the noun (exclude marking by system of nominal classification only)?	noun	derivation	word
GB188	AugN	Is there any productive augmentative marking on the noun (exclude marking by system of nominal classification only)?	noun	derivation	word
GB196	PPron2	Is there a male/female distinction in 2nd person independent pronouns?	pronoun	deixis	word
GB197	fmPron1	Is there a male/female distinction in 1st person independent pronouns?	pronoun	deixis	word
GB204	AllEvery	Do collective ('all') and distributive ('every') universal quantifiers differ in their forms or their syntactic positions?	not assignable	quantification	other
GB250	PredPossHab	Can predicative possession be expressed with a transitive 'habeo' verb?	verb	possession	clause
GB252	PredPossLoc	Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?	noun/pronoun	possession	clause
GB253	PredPossDat	Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?	noun/pronoun	possession	clause
GB254	PredPossAdn	Can predicative possession be expressed with an S-like possessum and a possessor that is coded like an adnominal possessor?	noun/pronoun	possession	clause
GB256	PredPossCom	Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?	noun/pronoun	possession	clause
GB257	QIntonation	Can polar interrogation be marked by intonation only?	not assignable	interrogation	clause
GB263	InterPtclFin	Is there a clause-final polar interrogative particle?	particle	interrogation	clause
GB264	InterPtclMid	Is there a polar interrogative particle that most commonly occurs neither clause-initially nor clause-finally?	particle	interrogation	clause
GB265	CompSurpass	Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?	verb	modification	clause
GB266	CompLoc	Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?	noun/pronoun	modification	clause
GB273	CompThan	Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass/exceed' meaning?	particle	modification	clause

GB275	CompMarkAdj	Is there a bound comparative degree marker on the property word in a comparative construction?	adjective	modification	word
GB276	CompMarkFree	Is there a non-bound comparative degree marker modifying the property word in a comparative construction?	particle	possession	clause
GB285	InterPtclVerbM	Can polar interrogation be marked by a question particle and verbal morphology?	not assignable	interrogation	clause
GB286	InterVerbM	Can polar interrogation be indicated by overt verbal morphology only?	verb	interrogation	clause
GB296	Ideophones	Is there a phonologically or morphosyntactically definable class of ideophones that includes ideophones depicting imagery beyond sound?	other	other	word
GB297	InterVnotV	Can polar interrogation be indicated by a V-not-V construction?	verb	interrogation	clause
GB298	NegAuxV	Can standard negation be marked by an inflecting word (“auxiliary verb”)?	verb	negation	clause
GB299	NegAuxPtcl	Can standard negation be marked by a non-inflecting word (“auxiliary particle”)?	particle	negation	clause
GB301	InclusoryConstr	Is there an inclusory construction?	noun/pronoun	deixis	clause
GB302	PassPtcl	Is there a phonologically free passive marker (“particle” or “auxiliary”)?	particle	valency	clause
GB304	PassAgentOvert	Can the agent be expressed overtly in a passive clause?	not assignable	other	clause
GB305	ReflPron	Is there a phonologically independent reflexive pronoun?	pronoun	other	word
GB306	Non2PRecipPron	Is there a phonologically independent non-bipartite reciprocal pronoun?	pronoun	other	word
GB309	MultiplePstFut	Are there multiple past or multiple future tenses, distinguishing distance from Time of Reference?	verb	TAME+	clause
GB312	MoodV	Is there overt morphological marking on the verb dedicated to mood?	verb	TAME+	word
GB313	PossPron	Are there special adnominal possessive pronouns that are not formed by an otherwise regular process?	pronoun	possession	word
GB316	SGFree	Is singular number regularly marked in the noun phrase by a dedicated phonologically free element?	particle	quantification	NP

GB318	PluralFree	Is plural number regularly marked in the noun phrase by a dedicated phonologically free element?	particle	quantification	NP
GB322	DirEvid	Is there grammatical marking of direct evidence (perceived with the senses)?	verb	TAME+	other
GB323	IndirEvid	Is there grammatical marking of indirect evidence (hearsay, inference, etc.)?	verb	TAME+	other
GB324	DoWhat	Is there an interrogative verb for content interrogatives (who?, what?, etc.)?	verb	interrogation	clause
GB325	HowMuchMany	Is there a count/mass distinction in interrogative quantifiers?	particle	quantification	word
GB326	ContInterInSitu	Do (nominal) content interrogatives normally or frequently occur in situ?	not assignable	word order	clause
GB327	NRelatCl	Can the relative clause follow the noun?	not assignable	word order	clause
GB328	RelatClNoun	Can the relative clause precede the noun?	not assignable	word order	clause
GB329	InterHeadRelCl	Are there internally-headed relative clauses?	not assignable	word order	clause
GB330	CorrelatRelCl	Are there correlative relative clauses?	not assignable	word order	clause
GB333	DecimalNS	Is there a decimal numeral system?	not assignable	quantification	word
GB400	PersonNeutral	Are all person categories neutralized in some voice, tense, aspect, mood and/or negation?	verb	TAME+	clause
GB401	PatientlabV	Is there a class of patient-labile verbs?	verb	valency	clause
GB403	SupplCome	Does the verb for ‘come’ have suppletive verb forms?	verb	TAME+	word
GB408	AccAlignment	Is there any accusative alignment of flagging?	noun/pronoun	argument marking (core)	clause
GB409	ErgAlignment	Is there any ergative alignment of flagging?	noun/pronoun	argument marking (core)	clause
GB410	NeutAlignment	Is there any neutral alignment of flagging?	noun/pronoun	argument marking (core)	clause
GB415	Polite2Prs	Is there a politeness distinction in 2nd person forms?	pronoun	deixis	word
GB421	ComplPrep	Is there a preposed complementizer in complements of verbs of thinking and/or knowing?	not assignable	word order	clause
GB422	ComplPostp	Is there a postposed complementizer in complements of verbs of thinking and/or knowing?	not assignable	word order	clause
GB431	PrfPossessed	Can adnominal possession be marked by a prefix on the possessed noun?	noun/pronoun	possession	word

GB432	SuffPossessor	Can adnominal possession be marked by a suffix on the possessor?	noun/pronoun	possession	word
GB433	SuffPossessed	Can adnominal possession be marked by a suffix on the possessed noun?	noun/pronoun	possession	word
GB519	AuxPtclMood	Can mood be marked by a non-inflecting word ("auxiliary particle")?	particle	TAME+	clause
GB520	AuxPtclAspect	Can aspect be marked by a non-inflecting word ("auxiliary particle")?	particle	TAME+	clause
GB521	AuxPtclTense	Can tense be marked by a non-inflecting word ("auxiliary particle")?	particle	TAME+	clause
GB522	Prodrop	Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context ("pro-drop" or "null anaphora")?	not assignable	other	clause
TE003	VHPalat	Is there palatal vowel harmony?	not assignable	phonological distinctiveness	phonological shape
TE004	VHTongueRoot	Is there tongue root vowel harmony?	not assignable	phonological distinctiveness	phonological shape
TE005	InitVelarNasal	Is there an initial velar nasal?	not assignable	phonological distinctiveness	phonological shape
TE006	InitR	Is there an initial r-?	not assignable	phonological distinctiveness	phonological shape
TE007	ConsClusters	Are there initial consonant clusters in native words?	not assignable	phonological distinctiveness	phonological shape
TE008	VoicDistStops	Is there a voicing distinction in stops?	not assignable	phonological distinctiveness	phonological shape
TE018	MiTiDistPrsPron	Is there a mi-ti opposition in 1 vs. 2 SG personal pronouns?	pronoun	other	word
TE019	OblStemWithN	Is the secondary oblique stem of personal pronouns formed with a dental nasal?	pronoun	argument marking (non-core)	word
TE027	PlPlusCollectPl	Can 1PL marker be augmented by a collective plural marker?	pronoun	deixis	word
TE031	NonInitVelarNasal	Is there a non-initial velar nasal?	not assignable	phonological distinctiveness	phonological shape
TE037	VoicDistFricat	Is there a voicing distinction in fricatives?	not assignable	phonological distinctiveness	phonological shape
TE038	DistinctionLR	Are there two separate liquid phonemes (r/l)?	not assignable	phonological distinctiveness	phonological shape

TE039	VowelLength	Is there vowel length distinction?	not assignable	phonological distinctiveness	phonological shape
TE050	PlNounAfterNum	Do cardinal numerals require agreement on noun phrases?	noun	quantification	NP
TE052	AccObjSpecific	Are accusative-marked objects specific while unmarked objects are non-specific?	noun/pronoun	argument marking (core)	word
TE053	RecipLocSame	Are recipient and location expressed by the same marker?	noun/pronoun	argument marking (non-core)	word
TE054	PolarContInterr	Is there a distinction between marking of interrogation in polar and content questions?	not assignable	interrogation	clause
TE059	CausPassSame	Are causative and passive expressed by a formally identical marker?	verb	valency	word
TE066	VRoot2SGImper	Is imperative form for 2SG identical to bare verb root?	verb	TAME+	word
TE078	Stops3LarContr	Are there three laryngeal contrasts for stops?	not assignable	phonological distinctiveness	phonological shape
TS001	NumNoun	Can numeral precede the noun?	not assignable	word order	NP
TS002	NounNum	Can numeral follow the noun?	not assignable	word order	NP
TS003	DemN	Can demonstrative precede the noun?	not assignable	word order	NP
TS005	PossPossessed	Can adnominal possessor precede the possessed noun?	not assignable	word order	NP
TS006	PossessedPoss	Can adnominal possessor follow the possessed noun?	not assignable	word order	NP
TS007	AdjNoun	Can adnominal property word precede the noun?	not assignable	word order	NP
TS009	AllNoun	Can adnominal collective universal quantifier ('all') precede the noun?	not assignable	word order	NP
TS010	NounAll	Can adnominal collective universal quantifier ('all') follow the noun?	not assignable	word order	NP
TS079	VHheight	Is there height vowel harmony?	not assignable	phonological distinctiveness	phonological shape
TS080	VHLabial	Is there labial vowel harmony?	not assignable	phonological distinctiveness	phonological shape
TS086	SV	Is the pragmatically unmarked order of SV in intransitive clauses?	not assignable	word order	clause
TS088	AspiratedStops	Is there an aspiration distinction in stops?	not assignable	phonological distinctiveness	phonological shape

Table S2: Number of languages coded per feature, number of "Present" values and median D. The "Values" include all coding that is not a "?" ("not enough information"), i.e. a feature absence or a feature presence. A

D value between 0 and 0.5 means the feature has a phylogenetic signal, between 0.5 and 1: the feature has no phylogenetic signal, below 0: the feature is overclumped, above 1: the feature is overly dispersed. Abbreviations: D = phylogenetic signal measured as D (median), SD(D) = standard deviation from the median D.

Feature	Values	Present	D	SD(D)
GB020	60	3	0.97	0.15
GB021	60	2	0.79	0.26
GB022	59	15	0.56	0.08
GB023	60	4	0.11	0.14
GB026	29	8	-0.08	0.19
GB027	34	17	0.74	0.13
GB028	57	16	-0.15	0.12
GB030	57	1	-0.41	1.09
GB031	56	3	-0.2	0.53
GB035	55	19	0.15	0.09
GB037	55	2	1.52	0.64
GB039	57	15	0.15	0.09
GB041	53	1	-1.4	2.33
GB042	60	1	1.72	0.64
GB043	60	1	1.69	0.66
GB044	59	45	0.3	0.09
GB046	31	27	-0.01	0.25
GB047	41	39	-0.22	0.38
GB048	37	32	0.65	0.12
GB049	40	36	0.23	0.24
GB057	55	17	-0.01	0.11
GB059	56	9	-0.5	0.13
GB068	54	14	-0.09	0.13
GB069	54	9	-0.6	0.17
GB070	60	59	-0.89	0.28
GB071	59	58	-0.8	0.22
GB072	60	59	-0.89	0.28
GB073	58	58	-4.78	1.01
GB074	57	1	1.09	2.7
GB075	56	56	-4.77	1
GB079	60	1	0.06	0.54
GB080	60	60	-4.96	0.84
GB082	59	51	1.19	0.1
GB083	60	58	0.45	0.1
GB084	57	30	0.11	0.09
GB086	57	44	0.6	0.09
GB089	60	38	-0.35	0.11
GB091	60	38	-0.36	0.11
GB103	59	2	0.2	0.25
GB105	54	5	0.52	0.09
GB107	57	42	-0.25	0.12
GB108	56	9	-0.33	0.14
GB110	46	1	2.31	1.13
GB111	51	22	-0.88	0.16
GB113	50	50	-4.88	1.09
GB114	46	25	-0.39	0.14
GB115	48	39	0.25	0.13
GB117	50	46	0.65	0.2
GB118	37	8	0.84	0.12
GB119	56	37	0.34	0.09
GB120	56	41	0.78	0.07

GB121	57	29	0.61	0.07
GB122	40	8	-0.18	0.16
GB123	28	23	-1.08	0.32
GB126	41	30	0.59	0.07
GB127	24	2	-1.58	1.35
GB132	60	4	1.06	0.15
GB133	60	60	-4.96	0.81
GB134	46	44	-0.05	0.43
GB135	38	33	0.59	0.2
GB136	41	19	0.02	0.11
GB137	58	44	0.11	0.09
GB138	56	2	0.89	0.7
GB139	51	35	0.08	0.1
GB140	49	7	0.58	0.15
GB146	23	9	-0.47	0.22
GB147	58	52	-0.29	0.19
GB150	52	49	0.86	0.24
GB151	59	1	-1.3	2.06
GB152	53	45	1.03	0.12
GB155	56	56	-4.87	1.09
GB156	58	1	-0.77	0.9
GB158	40	4	1.01	0.27
GB159	40	6	1.22	0.17
GB160	39	30	1.01	0.14
GB166	60	1	1.67	0.66
GB167	56	2	1.5	0.61
GB184	53	5	0.73	0.16
GB185	40	12	0.56	0.12
GB187	45	41	-0.05	0.26
GB188	44	5	0.16	0.27
GB196	57	1	-0.4	1.04
GB197	57	1	-0.4	1.04
GB204	28	24	0.96	0.24
GB250	41	4	-0.27	0.31
GB252	38	30	0.52	0.15
GB253	38	19	-0.41	0.18
GB254	42	23	0.22	0.11
GB256	35	10	-0.2	0.24
GB257	50	10	0.9	0.12
GB263	55	49	0.78	0.11
GB264	54	3	0.92	0.35
GB265	49	1	-0.41	1.19
GB266	48	39	-0.48	0.15
GB273	51	16	-0.4	0.15
GB275	45	13	0.39	0.12
GB276	46	4	0.49	0.4
GB285	54	1	-1.7	1.79
GB286	55	5	1.07	0.21
GB296	41	14	0.51	0.11
GB297	56	1	-2	1.68
GB298	58	13	-0.37	0.13
GB299	55	16	-0.17	0.12
GB301	16	7	-0.09	0.3
GB302	59	1	2.17	2.38
GB304	41	36	0.09	0.17
GB305	50	49	-1.01	0.62

GB306	34	4	-0.02	0.24
GB309	56	23	-0.01	0.1
GB312	60	59	0.04	0.18
GB313	55	8	0.88	0.16
GB316	60	1	1.73	0.65
GB318	60	1	1.7	0.65
GB322	55	12	0.6	0.09
GB323	54	15	1.04	0.08
GB324	44	21	1.26	0.08
GB325	41	12	1.12	0.1
GB326	39	35	0.87	0.18
GB327	52	9	0.68	0.13
GB328	52	51	0.01	0.65
GB329	41	5	0.57	0.2
GB330	45	1	-0.94	2.66
GB333	50	50	-4.48	0.94
GB400	54	5	0.83	0.23
GB401	14	5	0.91	0.29
GB403	40	4	-1.31	0.39
GB408	60	59	-0.31	0.17
GB409	55	1	-0.16	1.04
GB410	55	32	-0.18	0.11
GB415	57	18	0.56	0.09
GB421	34	6	-0.26	0.28
GB422	32	8	-0.58	0.25
GB431	60	1	1.06	0.55
GB432	60	50	0.04	0.09
GB433	59	44	-0.9	0.16
GB519	57	11	0.81	0.1
GB520	59	2	1.22	0.8
GB521	59	4	0.95	0.28
GB522	50	49	-0.59	0.57
TE003	58	31	-0.48	0.12
TE004	58	8	-0.44	0.14
TE005	49	11	-0.29	0.13
TE006	44	11	-0.45	0.17
TE007	56	8	-0.3	0.17
TE008	59	48	0.29	0.11
TE018	56	40	-0.27	0.13
TE019	57	43	-0.38	0.14
TE027	54	22	0.6	0.1
TE031	59	44	-0.12	0.11
TE037	59	35	0.66	0.06
TE038	59	42	-0.53	0.13
TE039	57	38	0.36	0.08
TE050	54	2	0.79	0.6
TE052	55	30	-0.58	0.14
TE053	57	35	0.23	0.08
TE054	44	39	0.61	0.14
TE059	52	7	0.55	0.13
TE066	60	36	-0.87	0.15
TE078	58	4	0.03	0.23
TS001	55	53	0.32	0.16
TS002	53	13	-0.07	0.11
TS003	59	59	-4.68	1.06
TS005	59	59	-4.6	1.01

TS006	59	43	-0.66	0.14
TS007	58	58	-4.85	1.01
TS009	30	23	1.09	0.12
TS010	28	10	0.7	0.16
TS079	54	5	-0.26	0.23
TS080	54	24	0.08	0.1
TS086	60	60	-4.94	0.81
TS088	59	12	-0.19	0.13

Table S3: Evolutionary rate for two models: ER (“equal rates”) and ARD (“All Rates Differ”). All the median rates and standard deviations are log10-transformed. Prior to the log10-transformation, median rates equal to 0 were replaced by 0.0000000001. Since the SD values for these features would receive the value -10 upon the log10-transformation (which would be misleading), these were not replaced by 0.0000000001 and thus have the value -Inf after the log10-transformation. A positive rate means the feature evolves relatively fast, a negative rate means the feature evolves relatively slow. Abbreviations: Rate(ER) = median rate according to the ER model, SD(ER) = standard deviation from the median rate (ER model), q10 = median rate of feature loss according to the ARD model, SD(q10) = standard deviation from the median rate of feature loss (ARD model), q01 = median rate of feature gain according to the ARD model, SD(q01) = standard deviation from the median rate of feature gain (ARD model).

Feature	Rate(ER)	SD(ER)	q10	SD(q10)	q01	SD(q01)
GB020	-1.4	-2	1.1	1.54	-0.19	0.25
GB021	-1.7	-Inf	0.75	1.48	-0.72	-0.03
GB022	-0.15	1.62	1.32	1.64	0.84	1.16
GB023	-1.4	-2	0.27	1	-0.89	-0.17
GB026	-0.77	-0.59	-0.41	1.08	-0.89	0.65
GB027	2	1.66	2	1.66	1.97	1.64
GB028	-0.68	0.46	-0.42	1.23	-10	0.81
GB030	-10	-Inf	-10	-Inf	-10	-Inf
GB031	-1.52	-2	0	0.57	-1.3	-0.72
GB035	-0.4	1.41	-0.23	1.49	-0.51	1.2
GB037	-1.7	-Inf	1.16	1.56	-0.3	0.09
GB039	-0.55	0.85	-0.12	1.3	-0.6	0.84
GB041	-10	-Inf	-10	-Inf	-10	-Inf
GB042	-10	-Inf	-10	-Inf	-10	-Inf
GB043	-10	-Inf	-10	-Inf	-10	-Inf
GB044	-0.49	1.22	-0.44	1.09	-0.07	1.57
GB046	-1	0.13	-1	0.61	-10	1.34
GB047	-1.3	-2	-1.3	-2	-10	-Inf
GB048	-0.85	0.78	0.06	0.88	0.77	1.6
GB049	-1.05	-1.7	-1.05	-0.24	-10	0.62
GB057	-0.64	1.18	-0.34	1.51	-0.64	1.15
GB059	-1.05	-1.7	-0.38	-0.82	-1.4	-1.7
GB068	-0.82	0.35	-0.48	0.9	-0.85	0.44
GB069	-1.1	-1.7	-0.49	-0.96	-2	-1.52
GB070	-1.52	-2	-1.7	-Inf	-10	-Inf
GB071	-1.52	-2	-1.7	-1	-10	0.5
GB072	-1.52	-2	-1.7	-Inf	-10	-Inf
GB073	-2	-Inf	-10	-Inf	-10	-2
GB074	-10	-Inf	-10	-Inf	-10	-Inf
GB075	-2	-Inf	-10	-Inf	-10	-2
GB079	-10	-Inf	-10	-Inf	-10	-Inf
GB080	-2	-Inf	-10	-Inf	-10	-1.7
GB082	-0.72	1.12	0.96	0.83	1.7	1.58
GB083	-1.3	-2	-1.4	-0.02	-10	1.28
GB084	-0.47	1.46	-0.42	1.52	-0.51	1.55

GB086	-0.48	1.38	0.02	1.08	0.49	1.57
GB089	-0.8	0.21	-0.92	0.63	-0.68	0.85
GB091	-0.8	-1.4	-0.92	0.59	-0.68	0.8
GB103	-1.52	-Inf	0.66	1.47	-0.8	-0.03
GB105	-1.05	-1.7	1.32	1.59	0.34	0.59
GB107	-0.85	0.32	-0.89	0.08	-0.6	0.5
GB108	-1.1	-1.7	-0.33	-0.6	-1.05	-1.7
GB110	-10	-Inf	-10	-Inf	-10	-Inf
GB111	-1.15	-1.7	-1.4	-1.22	-1.22	-1.4
GB113	-2	-Inf	-10	-Inf	-10	-2
GB114	-0.8	0.61	-1.4	1.25	-0.8	1.3
GB115	-0.77	0.53	-0.72	0.65	-0.4	1.24
GB117	-1.05	-1.7	-1.1	-1.7	-10	-0.8
GB118	-0.46	1.51	1.06	1.6	0.49	1.03
GB119	-0.43	1.26	-0.38	1.12	-0.21	1.39
GB120	-0.09	1.56	0.32	1.17	0.71	1.58
GB121	1.07	1.64	1.09	1.64	1.08	1.64
GB122	-0.92	0.16	-0.39	1.19	-1.05	0.58
GB123	-1.15	-1.7	-1.1	-1.7	-10	-1.3
GB126	1.28	1.67	1.58	1.23	1.98	1.63
GB127	-10	-Inf	-10	-1	-10	-2
GB132	-1.22	-2	1.21	1.56	0.06	0.41
GB133	-2	-Inf	-10	-Inf	-10	-1.7
GB134	-1.3	-2	-1.3	-2	-10	-Inf
GB135	-0.85	0.5	0.46	0.85	1.19	1.59
GB136	-0.6	1.44	-0.48	1.57	-0.66	1.49
GB137	-0.66	0.84	-0.59	0.94	-0.13	1.4
GB138	-1.7	-Inf	0.88	1.5	-0.57	0.03
GB139	-0.6	1.12	-0.72	1.13	-0.33	1.44
GB140	-0.89	0.47	0.47	1.51	-0.33	0.72
GB146	-0.66	-1.22	-0.48	0.97	-0.8	0.75
GB147	-1.15	-2	-1.15	-1.7	-10	-0.82
GB150	-1.15	-2	-1.15	0.26	-10	1.35
GB151	-10	-Inf	-10	-Inf	-10	-Inf
GB152	-0.72	1	0.72	0.9	1.41	1.6
GB155	-2	-Inf	-10	-Inf	-10	-2
GB156	-10	-Inf	-10	-Inf	-10	-Inf
GB158	-1.1	-1.7	1.21	1.56	0.26	0.6
GB159	-0.82	0.03	0.85	1.43	0.1	0.67
GB160	1.39	1.67	1.37	1.16	1.85	1.62
GB166	-10	-Inf	-10	-Inf	-10	-Inf
GB167	-1.7	-Inf	1.11	1.54	-0.36	0.07
GB184	-1.05	-1.7	1.23	1.58	0.25	0.59
GB185	-0.35	1.59	0.61	1.62	0.23	1.24
GB187	-1.1	-1.7	-1.1	0.03	-10	0.94
GB188	-1.1	-2	-0.04	1.18	-0.92	0.28
GB196	-10	-Inf	-10	-Inf	-10	-Inf
GB197	-10	-Inf	-10	-Inf	-10	-Inf
GB204	-0.85	1.3	0.14	0.87	0.81	1.55
GB250	-1.15	-1.7	-0.18	0.96	-1.4	-0.03
GB252	-0.64	1.56	0.79	1.13	1.31	1.65
GB253	-0.72	0.58	-0.72	1.08	-0.74	1.06
GB254	-0.33	1.13	-0.34	1.15	-0.3	1.21
GB256	-0.89	0.78	-0.42	1.29	-0.82	0.88
GB257	-0.52	1.55	1.4	1.62	0.79	1.01
GB263	-0.85	0.52	0.76	0.76	1.6	1.6

GB264	-1.3	-2	1.19	1.55	-0.05	0.31
GB265	-10	-Inf	-10	-Inf	-10	-Inf
GB266	-1	-1.52	-1.15	-1.4	-0.74	-1.22
GB273	-0.85	-1.4	-1.52	-0.92	-0.89	-1.52
GB275	-0.44	1.49	0.08	1.58	-0.35	1.18
GB276	-1.3	-2	0.49	1.39	-0.6	0.36
GB285	-10	-Inf	-10	-Inf	-10	-Inf
GB286	-1.05	-1.7	1.66	1.59	0.66	0.58
GB296	0.54	1.66	1.24	1.66	0.94	1.36
GB297	-10	-Inf	-10	-Inf	-10	-Inf
GB298	-1.05	-1.7	-0.59	-0.7	-1.05	-1.7
GB299	-0.82	0.18	-0.49	0.92	-0.85	0.53
GB301	-0.35	0.98	-0.27	1.35	-0.52	1.19
GB302	-10	-Inf	-10	-Inf	-10	-Inf
GB304	-1	-1.7	-0.85	0.05	-0.12	0.84
GB305	-1.52	-2	-1.52	-2	-10	-2
GB306	-1.05	-1.7	-0.15	1.36	-0.85	0.47
GB309	-0.59	0.79	-0.46	1.33	-0.72	1.17
GB312	-1.52	-2	-1.7	-Inf	-10	-1.7
GB313	-0.85	0.28	0.99	1.56	0.21	0.79
GB316	-10	-Inf	-10	-Inf	-10	-Inf
GB318	-10	-Inf	-10	-Inf	-10	-Inf
GB322	-0.62	1.02	0.36	1.47	-0.22	0.91
GB323	2	1.63	2	1.59	1.57	1.16
GB324	2	1.56	2	1.55	1.94	1.5
GB325	1.53	1.66	1.93	1.62	1.54	1.22
GB326	-0.96	0.44	0	0.75	0.82	1.59
GB327	-0.66	0.84	0.94	1.55	0.27	0.87
GB328	-1.52	-2	-1.52	-2	-10	-1.52
GB329	-0.92	0.29	1.24	1.6	0.38	0.73
GB330	-10	-Inf	-10	-Inf	-10	-Inf
GB333	-2	-Inf	-10	-Inf	-10	-Inf
GB400	-1.05	-1.7	0.84	1.51	-0.14	0.51
GB401	2	1.63	1.97	1.6	1.67	1.3
GB403	-1.52	-2	-0.66	-0.92	-1.7	-2
GB408	-1.52	-2	-1.7	-Inf	-10	-2
GB409	-10	-Inf	-10	-Inf	-10	-Inf
GB410	-0.72	-0.01	-0.77	1.05	-0.64	1.17
GB415	-0.21	1.62	0.64	1.63	0.28	1.29
GB421	-1.05	0.15	-0.21	1.18	-1.3	0.5
GB422	-1	0.21	-0.55	-1	-1.05	-1.52
GB431	-10	-Inf	-10	-Inf	-10	-Inf
GB432	-0.89	-1.52	-0.96	-0.01	-0.34	0.65
GB433	-1.22	-1.7	-1.15	-1.52	-10	-1.52
GB519	-0.4	1.58	1.6	1.61	0.98	0.98
GB520	-1.7	-Inf	0.92	1.52	-0.55	0.03
GB521	-1.22	-2	1.26	1.56	0.12	0.41
GB522	-1.52	-2	-1.52	-2	-10	-Inf
TE003	-0.89	0.12	-1	-1.15	-0.89	-1.4
TE004	-1.1	-1.7	-0.35	0.5	-1.3	-0.3
TE005	-1	-1.7	-0.38	0.65	-1.05	0.1
TE006	-0.96	-1.52	-0.62	-1.05	-1.05	-1.52
TE007	-1.15	-1.7	-0.18	-0.59	-1.1	-1.7
TE008	-0.7	0.51	-0.55	0.54	-0.05	1.14
TE018	-0.89	0.1	-0.85	0.3	-10	0.67
TE019	-1	-1.7	-0.92	-1.52	-10	-1.7

TE027	1.27	1.65	1.37	1.65	1.19	1.47
TE031	-0.82	0.44	-0.85	0.5	-0.54	0.93
TE037	1.15	1.64	1.11	1.5	1.25	1.64
TE038	-1	-1.7	-0.92	-1.52	-10	-1.7
TE039	-0.24	1.57	-0.17	1.31	0.09	1.59
TE050	-1.52	-Inf	0.58	1.42	-0.82	-0.04
TE052	-0.92	-1.52	-1.22	0.44	-0.89	0.5
TE053	-0.43	1.47	-0.48	1.41	-0.21	1.59
TE054	-0.89	-1.7	-0.07	0.68	0.71	1.49
TE059	-0.92	0.48	1.21	1.6	0.39	0.79
TE066	-1.3	-1.7	-10	-1.7	-1.1	-1.52
TE078	-1.3	-2	-0.2	1.36	-1.15	0.22
TS001	-1.3	-2	-1.3	-0.02	-10	1.23
TS002	-0.77	0.9	-0.39	1.3	-0.74	0.81
TS003	-2	-Inf	-10	-Inf	-10	-1.7
TS005	-2	-Inf	-10	-Inf	-10	-1.7
TS006	-1.1	-1.7	-1.05	-1.7	-10	-1.7
TS007	-2	-Inf	-10	-Inf	-10	-2
TS009	0.63	1.67	0.94	1.18	1.39	1.64
TS010	-0.06	1.55	0.27	1.54	0	1.26
TS079	-1.3	-2	-0.33	-0.36	-1.22	-1.7
TS080	-0.57	1.3	-0.43	1.53	-0.66	1.43
TS086	-2	-Inf	-10	-Inf	-10	-1.7
TS088	-0.92	-1.52	-0.28	0.5	-1	-0.09

Table S4: Reconstructed states for each of the five language families. The closer the value is to 0, the more likely it is that the feature was not present in the proto-language. A value close to 1 means that the feature was most likely present in the proto-language. A value around 0.5 means that the feature can be reconstructed neither as absent nor as present in the proto-language. Abbreviations: Trk = probability of 1 for Proto-Turkic, Mng = probability of 1 for Proto-Mongolic, Tng = probability of 1 for Proto-Tungusic, Krn = probability of 1 for Proto-Koreanic, Jpn = probability of 1 for Proto-Japonic.

Feature	Trk	Mng	Tng	Krn	Jpn
GB020	0.05	0.05	0.05	0.04	0.05
GB021	0.03	0.03	0.03	0.02	0.03
GB022	0.28	0.24	0.25	0.2	0.23
GB023	0.05	0.1	0.06	0.03	0.05
GB026	0.53	0.15	0.83	0.32	0.12
GB027	0.49	0.49	0.49	0.42	0.48
GB028	0.26	0.96	0.96	0.1	0.2
GB030	0	0	0	0	0
GB031	0.03	0.02	0.02	0.01	0.03
GB035	0.6	0.21	0.2	0.8	0.67
GB037	0.03	0.03	0.03	0.03	0.03
GB039	0.17	0.59	0.19	0.65	0.38
GB041	0	0	0	0	0
GB042	0	0	0	0	0
GB043	0	0	0	0	0
GB044	0.8	0.75	0.64	0.6	0.47
GB046	0.96	0.92	0.96	0.87	0.95
GB047	1	1	1	1	1
GB048	0.91	0.91	0.91	0.89	0.9
GB049	0.99	0.99	0.99	1	0.9
GB057	0.16	0.18	0.39	0.71	0.78
GB059	0.39	0.07	0.88	0.04	0.08
GB068	0.1	0.07	0.06	0.92	0.85

GB069	0.17	0.13	0.12	0.99	0.96
GB070	1	1	1	1	1
GB071	1	1	1	1	1
GB072	1	1	1	1	1
GB073	1	1	1	1	1
GB074	0	0	0	0	0
GB075	1	1	1	1	1
GB079	0	0	0	0	0
GB080	1	1	1	1	1
GB082	0.85	0.85	0.85	0.84	0.85
GB083	1	1	1	1	1
GB084	0.67	0.29	0.81	0.51	0.29
GB086	0.73	0.77	0.76	0.69	0.77
GB089	0.87	0.45	0.69	0.03	0.06
GB091	0.87	0.45	0.69	0.03	0.06
GB103	0.03	0.03	0.03	0.16	0.05
GB105	0.1	0.1	0.09	0.14	0.1
GB107	0.95	0.47	0.78	0.94	0.96
GB108	0.05	0.05	0.63	0.03	0.15
GB110	0	0	0	0	0
GB111	0.37	0.05	0.97	0.49	0.96
GB113	1	1	1	1	1
GB114	0.77	0.04	0.1	0.19	0.04
GB115	0.91	0.91	0.85	0.61	0.41
GB117	1	1	1	1	1
GB118	0.22	0.22	0.21	0.25	0.23
GB119	0.76	0.76	0.67	0.2	0.39
GB120	0.71	0.71	0.7	0.67	0.71
GB121	0.5	0.5	0.5	0.41	0.48
GB122	0.11	0.1	0.14	0.92	0.83
GB123	1	0.19	0.97	0.99	1
GB126	0.71	0.7	0.7	0.74	0.71
GB127	0	0	0	0	0
GB132	0.07	0.07	0.07	0.06	0.07
GB133	1	1	1	1	1
GB134	1	1	1	1	1
GB135	0.85	0.85	0.84	0.87	0.85
GB136	0.76	0.52	0.17	0.45	0.19
GB137	0.85	0.49	0.73	0.38	0.86
GB138	0.03	0.03	0.03	0.03	0.03
GB139	0.41	0.89	0.56	0.87	0.89
GB140	0.13	0.13	0.14	0.35	0.2
GB146	0.2	0.65	0.63	0.85	0.64
GB147	1	0.99	1	1	1
GB150	1	1	1	1	1
GB151	0	0	0	0	0
GB152	0.84	0.84	0.84	0.84	0.84
GB155	1	1	1	1	1
GB156	0	0	0	0	0
GB158	0.1	0.1	0.1	0.09	0.1
GB159	0.15	0.15	0.15	0.27	0.15
GB160	0.75	0.75	0.75	0.76	0.75
GB166	0	0	0	0	0
GB167	0.03	0.03	0.03	0.03	0.03
GB184	0.1	0.1	0.1	0.08	0.09
GB185	0.28	0.35	0.33	0.2	0.28

GB187	0.98	0.97	0.98	0.99	0.98
GB188	0.07	0.1	0.27	0.06	0.09
GB196	0	0	0	0	0
GB197	0	0	0	0	0
GB204	0.88	0.88	0.88	0.9	0.88
GB250	0.07	0.08	0.06	0.06	0.72
GB252	0.76	0.78	0.77	0.8	0.78
GB253	0.19	0.88	0.85	0.17	0.34
GB254	0.71	0.37	0.5	0.86	0.38
GB256	0.09	0.44	0.25	0.08	0.08
GB257	0.2	0.2	0.2	0.18	0.2
GB263	0.87	0.87	0.87	0.85	0.88
GB264	0.06	0.06	0.06	0.05	0.06
GB265	0	0	0	0	0
GB266	0.92	0.93	0.72	0.03	0.08
GB273	0.05	0.04	0.25	0.95	0.87
GB275	0.36	0.24	0.26	0.15	0.22
GB276	0.08	0.07	0.07	0.2	0.07
GB285	0	0	0	0	0
GB286	0.09	0.09	0.09	0.08	0.09
GB296	0.32	0.31	0.41	0.38	0.39
GB297	0	0	0	0	0
GB298	0.03	0.03	0.68	0.05	0.06
GB299	0.09	0.81	0.07	0.91	0.06
GB301	0.55	0.3	0.44	0.17	0.24
GB302	0	0	0	0	0
GB304	0.89	0.82	0.92	0.94	0.94
GB305	1	1	1	1	1
GB306	0.1	0.1	0.22	0.59	0.37
GB309	0.84	0.19	0.6	0.5	0.13
GB312	1	1	1	1	1
GB313	0.15	0.14	0.14	0.12	0.14
GB316	0	0	0	0	0
GB318	0	0	0	0	0
GB322	0.21	0.23	0.19	0.41	0.2
GB323	0.27	0.27	0.27	0.26	0.27
GB324	0.47	0.47	0.47	0.46	0.47
GB325	0.29	0.29	0.29	0.36	0.29
GB326	0.92	0.92	0.92	0.93	0.92
GB327	0.2	0.16	0.19	0.13	0.16
GB328	1	1	1	1	1
GB329	0.12	0.12	0.18	0.1	0.13
GB330	0	0	0	0	0
GB333	1	1	1	1	1
GB400	0.1	0.1	0.1	0.07	0.09
GB401	0.33	0.34	0.33	0.37	0.34
GB403	0.03	0.03	0.02	0.05	0.83
GB408	1	1	1	1	1
GB409	0	0	0	0	0
GB410	0.88	0.86	0.21	0.47	0.1
GB415	0.32	0.32	0.29	0.53	0.34
GB421	0.45	0.07	0.05	0.04	0.06
GB422	0.12	0.09	0.06	0.79	0.86
GB431	0	0	0	0	0
GB432	0.93	0.95	0.33	0.97	0.95
GB433	0.99	1	0.95	0.05	0.09

GB519	0.19	0.19	0.19	0.17	0.19
GB520	0.03	0.03	0.03	0.02	0.03
GB521	0.07	0.07	0.07	0.06	0.07
GB522	1	1	1	1	1
TE003	0.85	0.49	0.03	0.31	0.03
TE004	0.1	0.08	0.87	0.55	0.08
TE005	0.06	0.09	0.78	0.07	0.14
TE006	0.1	0.36	0.06	0.82	0.89
TE007	0.06	0.26	0.05	0.4	0.17
TE008	0.82	0.68	0.87	0.26	0.86
TE018	1	1	1	0.08	0.16
TE019	1	1	1	0.07	0.14
TE027	0.4	0.41	0.39	0.49	0.39
TE031	0.83	0.96	0.97	0.97	0.31
TE037	0.58	0.58	0.57	0.57	0.58
TE038	1	1	1	0.07	0.14
TE039	0.61	0.67	0.71	0.6	0.69
TE050	0.03	0.03	0.04	0.02	0.04
TE052	0.84	0.52	0.1	0.01	0.02
TE053	0.5	0.81	0.74	0.85	0.66
TE054	0.89	0.89	0.87	0.82	0.88
TE059	0.13	0.13	0.14	0.38	0.13
TE066	0.85	0.91	0.07	0	0
TE078	0.03	0.03	0.03	0.84	0.04
TS001	1	1	1	0.99	1
TS002	0.09	0.17	0.19	0.39	0.79
TS003	1	1	1	1	1
TS005	1	1	1	1	1
TS006	1	1	0.98	0.06	0.12
TS007	1	1	1	1	1
TS009	0.74	0.74	0.74	0.72	0.74
TS010	0.33	0.37	0.39	0.4	0.36
TS079	0.03	0.05	0.27	0.01	0.02
TS080	0.75	0.42	0.2	0.11	0.15
TS086	1	1	1	1	1
TS088	0.1	0.7	0.09	0.92	0.13

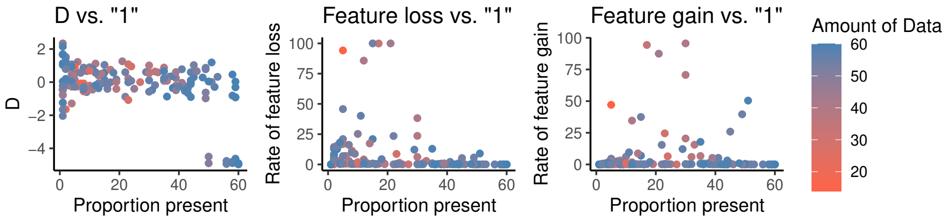

Figure S1: Relationship between features coded as "present", D and evolutionary rate: There is only a weak negative correlation between number of features coded as "present" and any of the measures: D ($\tau = -0.21$), rate of loss ($\tau = -0.34$) and rate of gain ($\tau = -0.34$).

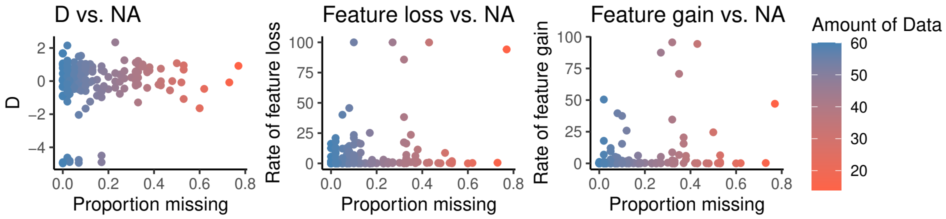

Figure S2: Relationship between missing data, D and evolutionary rate: There is only a weak positive correlation between missing data and any of the measures: D ($\tau = 0.06$), rate of loss ($\tau = 0.22$) and rate of gain ($\tau = 0.22$).

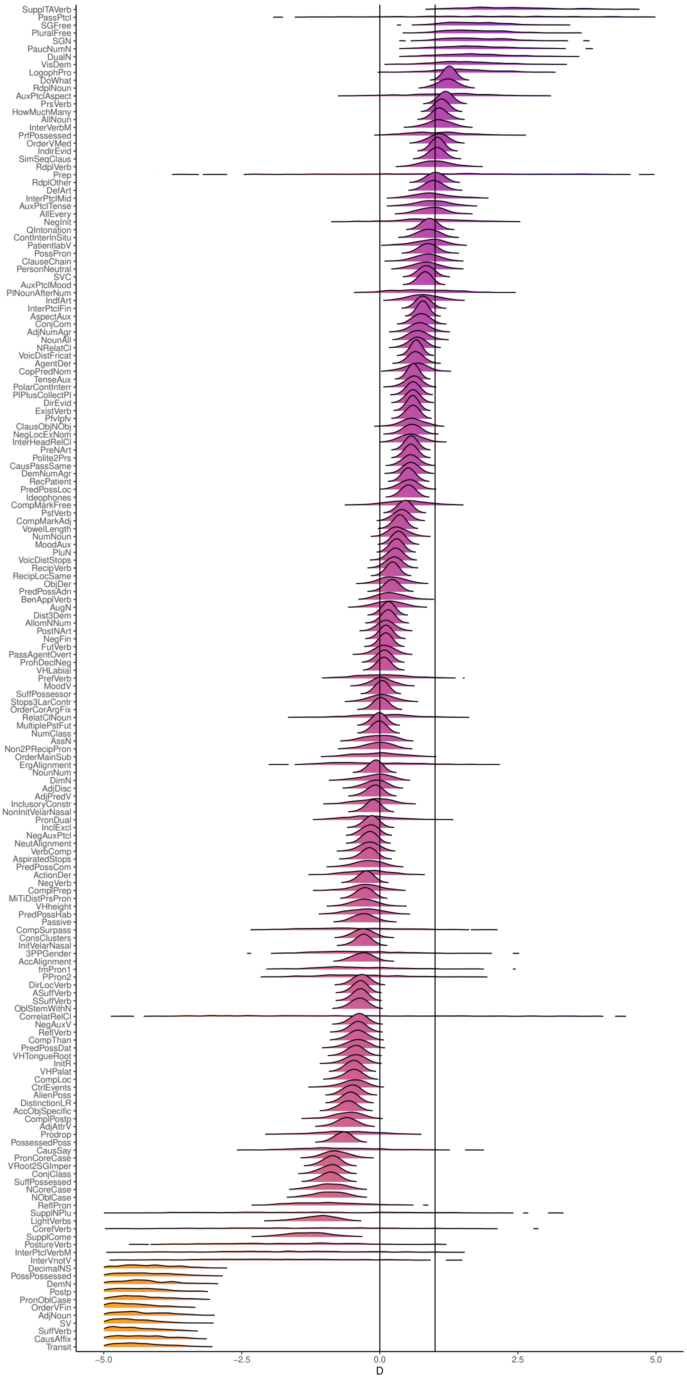

Figure S3: Distribution of D values across features.

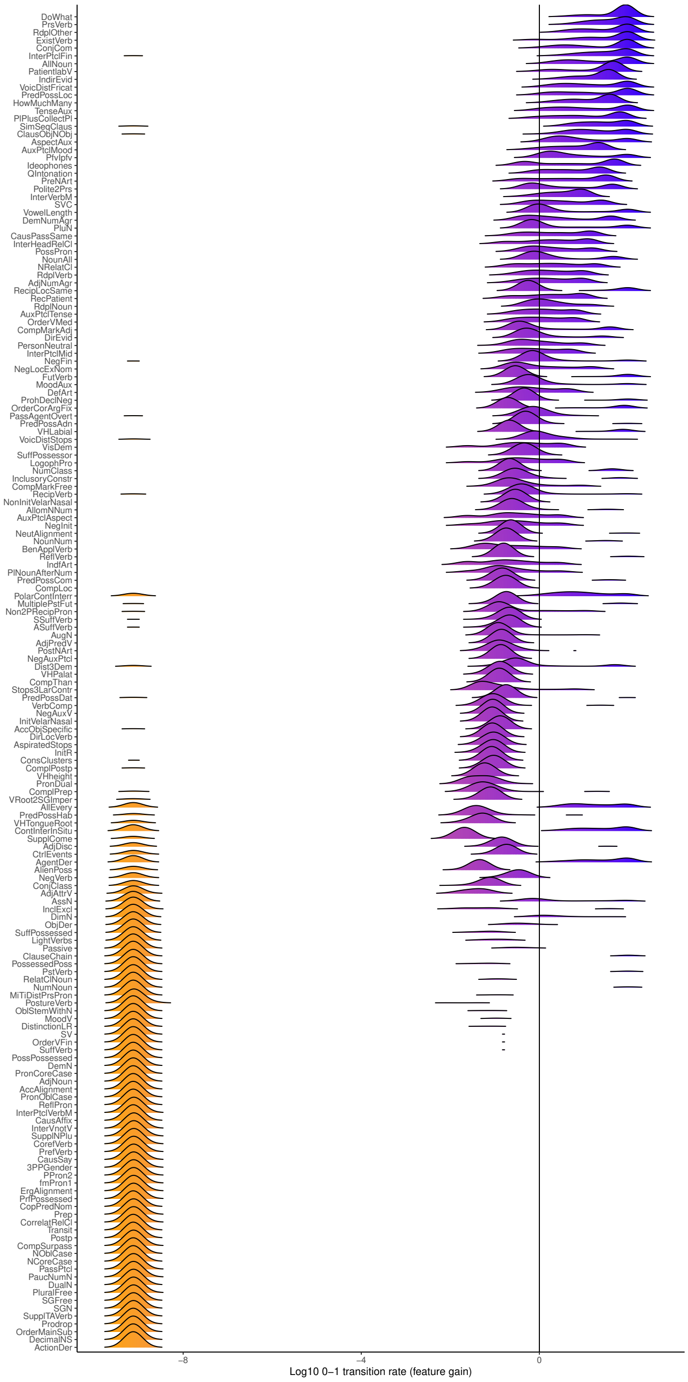

Figure S4: Distribution of rate of gain across features.

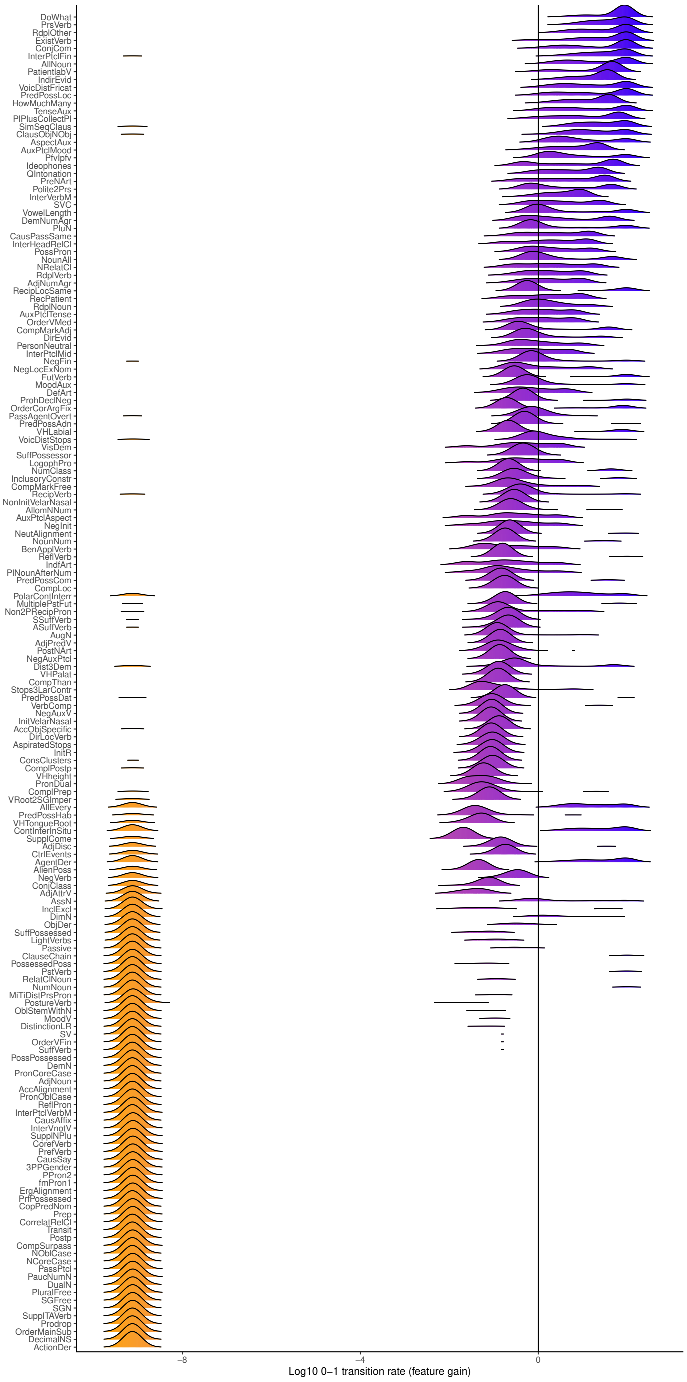

Figure S5: Distribution of rate of loss across features.

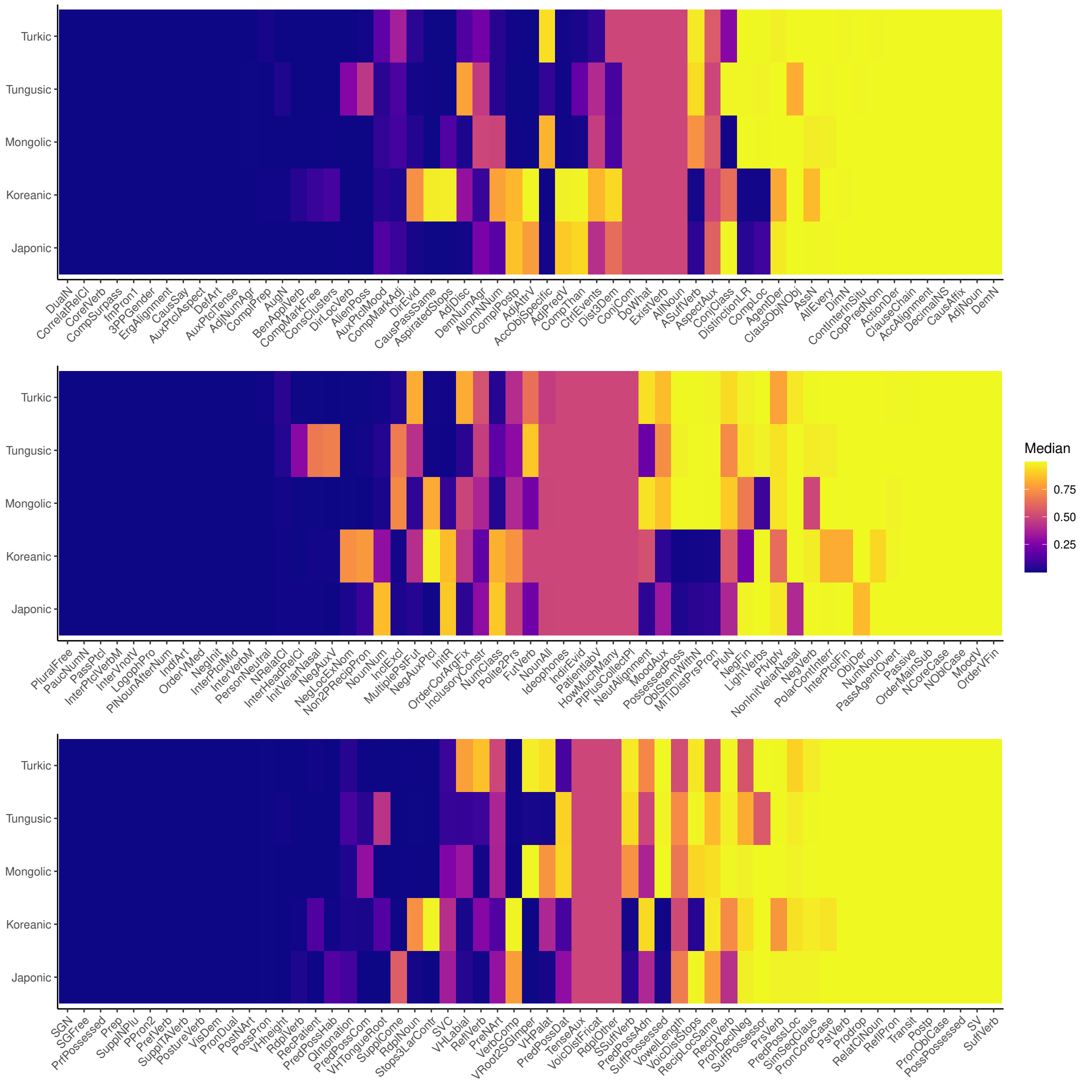

Figure S6: Reconstructed ancestral states: dark blue colouring for “absent”, yellow for “present”, shades of pink for an intermediate state, therefore a poor reconstruction.

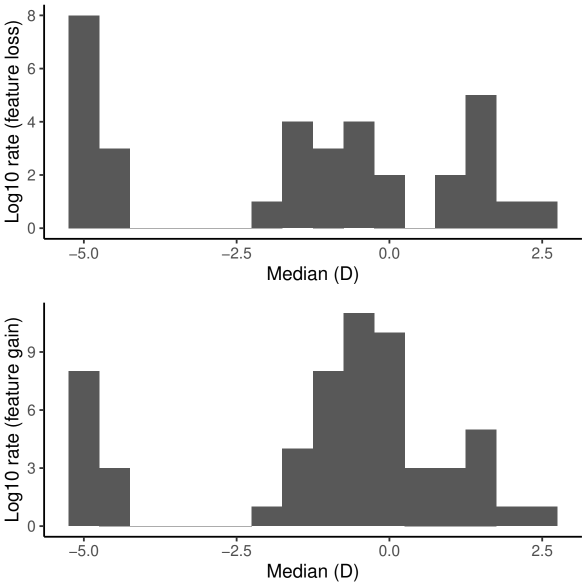

Figure S7: Distribution of D values for rates around zero

Electronic supplementary materials for the article "Modelling admixture across language levels to evaluate deep history claims"

1 Population structure across K2–K10 and language levels

1.1 Phonology

At $K = 2$, there is a split between Tungusic/Koreanic/Japonic and Turkic languages (see Fig. S1). In the Mongolic languages, there is a split in Mongolic languages between Southern Periphery Mongolic languages, sharing their ancestry with the Tungusic/Koreanic/Japonic languages, and the other Mongolic languages sharing their ancestry with Turkic. Some languages taking in an intermediate position between the two groups.

At $K = 3$, Japonic languages surface as a separate group. The rest of the ancestries remain roughly the same. Old Japanese shares most of its ancestry with Tungusic languages.

At $K = 4$, Tungusic languages stand out as a separate group.

At $K = 5$, there is more admixture in all languages, but the structure is roughly the same. Manchu starts having the ancestry similar to Koreanic languages, Old Japanese and Southern Periphery Mongolic languages.

At $K = 6$, Manchu, Southern Periphery Mongolic languages, Old Japanese, Yonaguni, some Turkic languages and Koreanic languages show roughly the same profile with increased levels of admixture. Modern Japonic and modern Tungusic languages have the lowest levels of admixture among the language families.

From $K = 7$ to $K = 10$, the admixture in all language families, apart from Japonic and Tungusic languages, continues to increase.

1.2 Morphology

At $K = 2$, there are two homogeneous groups: the so-called Altaic languages (Turkic, Mongolic, Tungusic) and Japonic/Koreanic (see Fig. S2).

At $K = 3$, Turkic languages surface as a separate group.

At $K = 4$, Tungusic languages surface as a separate group.

Starting from $K = 5$, there are no major changes in the grouping of languages, only the levels of admixture continue to increase until $K = 10$.

1.3 Syntax

At $K = 2$, there are two major groups: Altaic and Japonic/Koreanic. Tungusic languages show a high level of admixture and take in an intermediate position between these two groups (see Fig. S3).

At $K = 3$, there are three major groups: Turkic, Mongolic/Tungusic and Japonic/Koreanic.

At $K = 4$, a slight split between Mongolic and Tungusic languages becomes visible, with most Mongolic and Tungusic languages sharing both ancestries.

From $K = 5$ onward, the levels of admixture continue to increase in all languages, especially in Mongolic and Tungusic languages. At all the higher levels, only Turkic languages and Japonic/Koreanic languages form two homogeneous groups (with a few exceptions among Turkic languages).

1.4 Summary

Across all language levels, Japonic languages retained their homogeneous admixture profile (they are almost indistinguishable from Koreanic languages on the morphological and syntactic levels).

Among all levels, morphology showed most consistent separation of languages into four groups at all K's: Turkic, Mongolic, Tungusic and Japono-Koreanic.

2 Population structure with the whole feature set

The results of our analysis based on the whole dataset are most similar to those in morphology only (see Fig. S4).

3 Sampling

STRUCTURE is reported to sometimes fail to distinguish smaller groups, if the sample sizes of the populations are disproportionate. Mostly this effect is seen if there are too few loci (in our case features). 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 datasets increase in size, and our dataset 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.

In order to test the robustness of the findings and account for the possible affect of small samples sizes on our results, we ran the same analysis without the division of features into language levels for 10 random samples with two languages per family.

In four samples, there was not enough structure in the data, so that all languages have approximately the same contribution to each of the ancestries (samples 3, 4, 5, 9). In three samples, Koreanic and Japonic languages share exactly the same ancestry (samples 1, 2, 6, 10). In sample 7 Koreanic shares most of its ancestry with Japonic, in sample 8 it shares approximately the same amount of Japonic, Turkic and Tunguso-Mongolic ancestry. In most cases, even with so little data, STRUCTURE was able to attribute languages to their respective language families correctly. In some cases (sample 2 and 8), Mongolic and Tungusic were grouped together, which was also hypothesized in Altaic literature, in other cases (samples 6 and 10), Turkic and Mongolic were grouped together, which was also hypothesized in Altaic literature (see Fig. S5). The support of the one or the other hypothesis on the higher-level groupings inside of Altaic depends on the amount of data one has, but Japonic and Koreanic do not seem to be heavily affected by the low amount of data and are reliably clustered together even in small data sets.

References

Hubisz, Melissa J, Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources* 9:1322–1332.

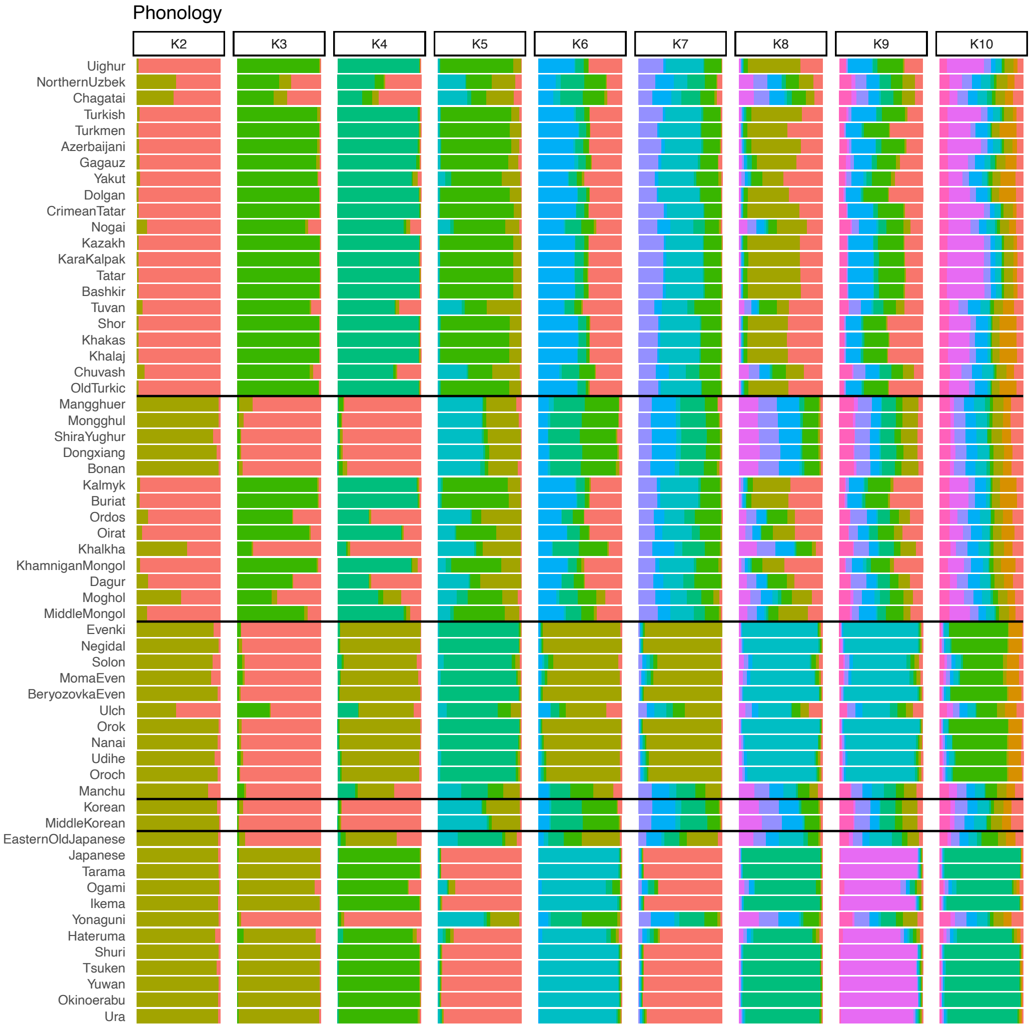

Figure S1: Admixture at the phonological level, K = 2 to K = 10.

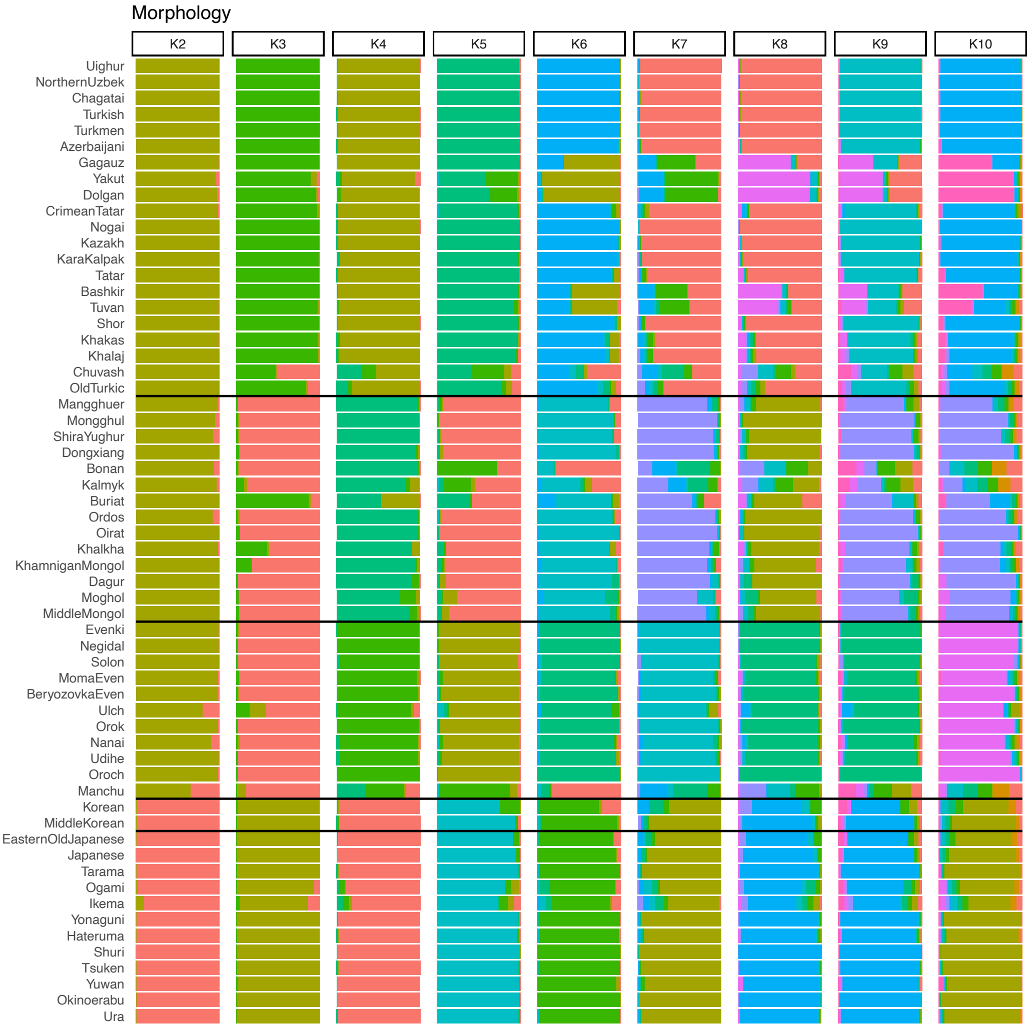

Figure S2: Admixture at the morphological level, K = 2 to K = 10.

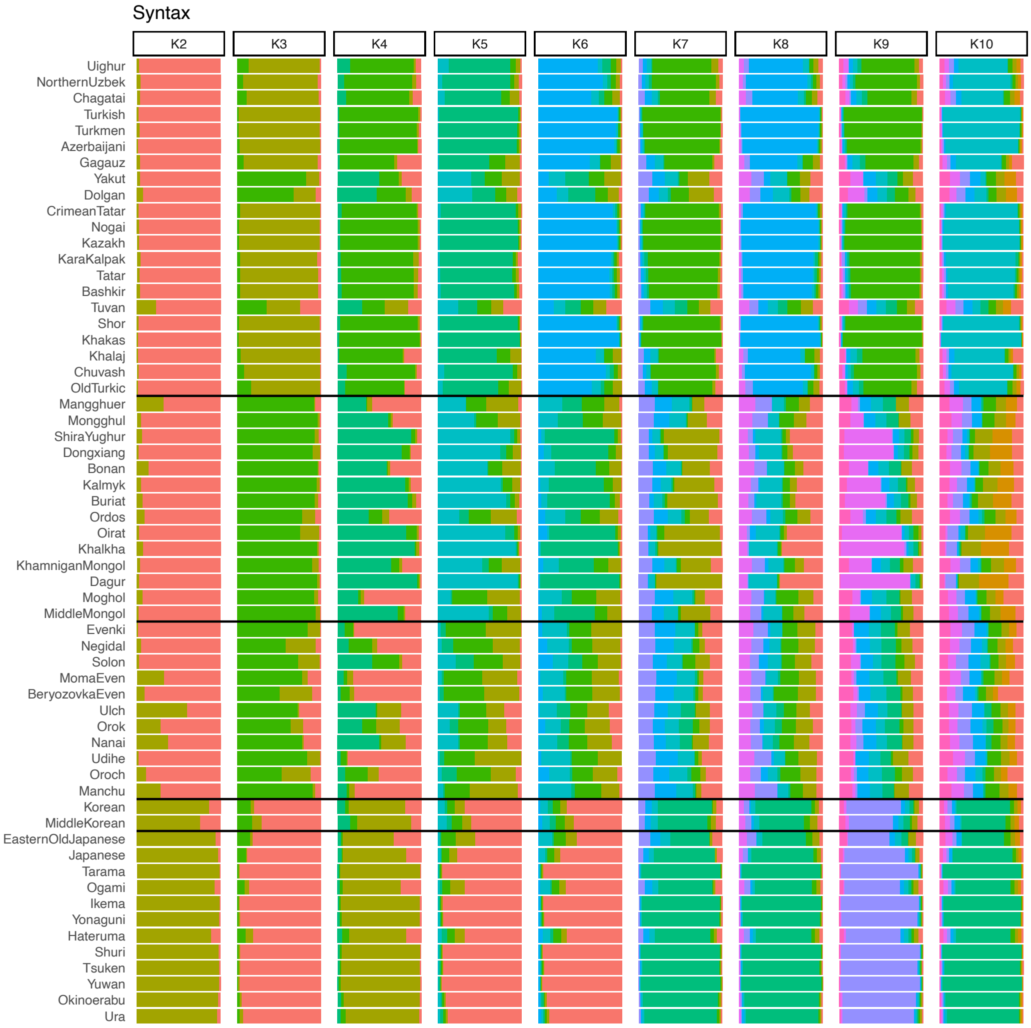

Figure S3: Admixture at the syntactic level, K = 2 to K = 10.

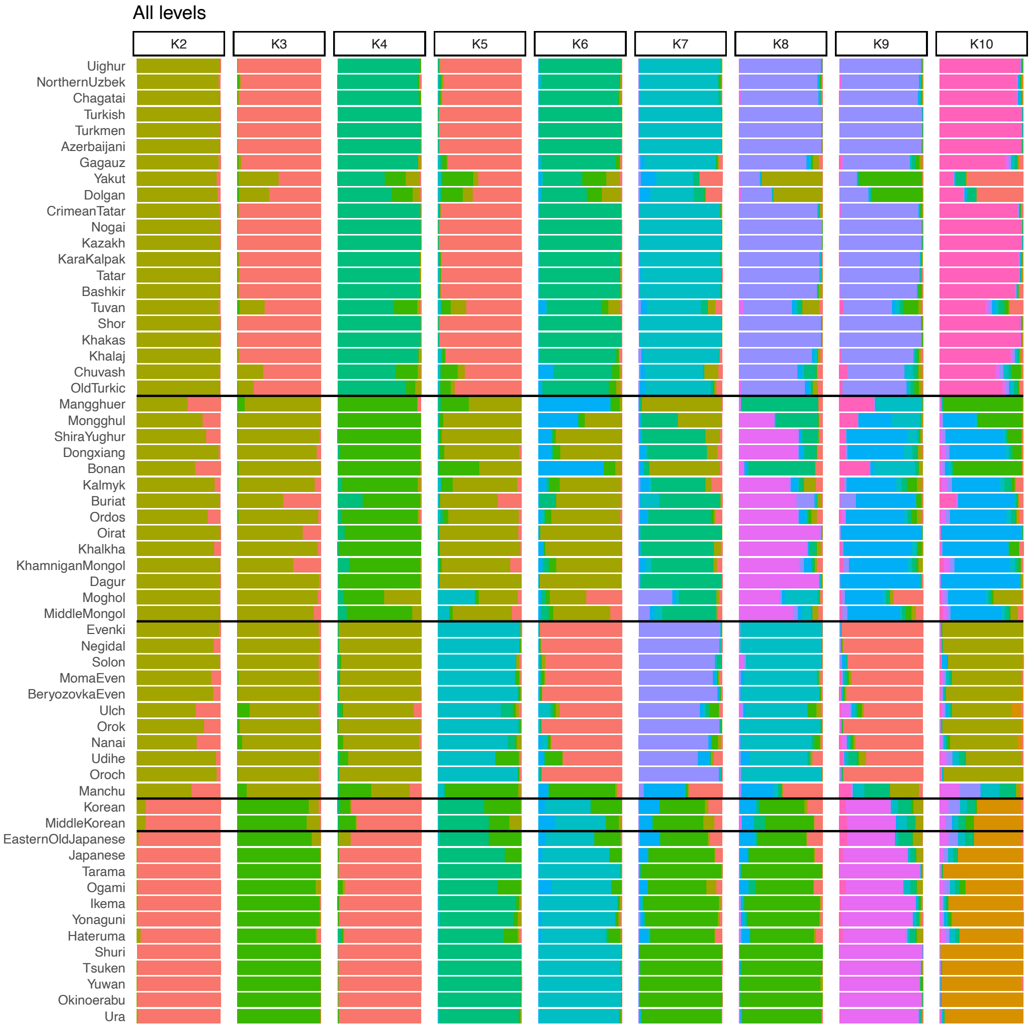

Figure S4: Admixture without a split into levels, K = 2 to K = 10.

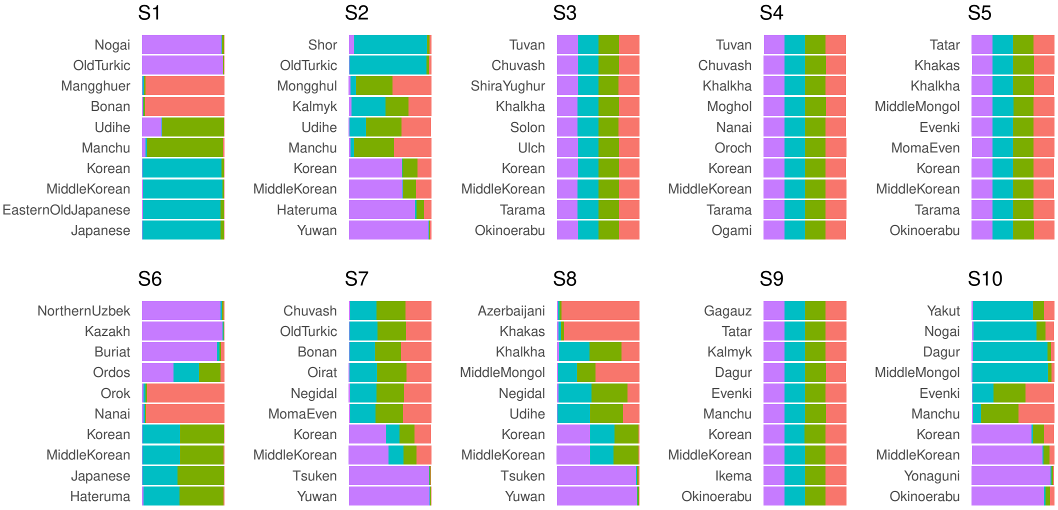

Figure S5: Admixture without a split into levels for subsamples of data, K = 2 to K = 10. Each sample contains 2 languages of each of the language families. The order of the language families is bottom to upper: Japonic, Koreanic, Tungusic, Mongolic, Turkic.

Summary/Zusammenfassung

Summary

Structural features have the potential to push the time barrier, after which we cannot test hypotheses about relatedness of languages, back in time. However, we have to know the stability of structural features in order to be able to apply them for such purposes. In this thesis I describe the typological profile of the Transeurasian languages, which serve later on as my sample for the analysis of stability, build a phylogenetic tree with these languages (Chapter 2), measure the stability of structural features as phylogenetic signal and evolutionary rate, reconstruct ancestral states of structural features (Chapter 3) and apply an admixture model from population genetics to test the performance of phonological, morphological and syntactic features in assigning languages to their respective language families and to investigate the level of diffusion in these three feature sets (Chapter 4).

In the first manuscript, I give a broad idea of what the main typological similarities between the so-called Transeurasian (Turkic, Mongolic, Tungusic, Koreanic, Japonic) languages are, supported by examples from the relevant languages. Building upon the debated assumption about the relatedness of these languages, I construct a phylogenetic tree of these five language families using Bayesian tree-building methods. In the resulting sample of high probability trees, there is a clear division between the Altaic and Japono-Koreanic languages, but most of the internal branches have a rather low resolution.

In the second manuscript, I tackle the question of stability of structural features and use two measures to investigate it: phylogenetic signal and evolutionary rate. More than half of structural features appear to have a high phylogenetic signal and evolve at a slow rate. I compare the stability across functional categories, parts of speech and language levels and come to a conclusion that argument marking (flagging and indexing), derivation and valency are the most stable functional categories, pronouns and nouns the most stable parts of speech and phonology and morphology the most stable

language levels. Furthermore, about one fifth of the features can be reconstructed as “Present” or “Absent” in the proto-language with 95% certainty.

Since we cannot exclude horizontal transmission in structural features, we have to use a method that deals with it appropriately and still provides valid results. In the third manuscript, I apply a method from population genetics (STRUCTURE) to the sets of structural features covering phonology, morphology and syntax. I compare the level of admixture and the precision of assignment of languages to language families across the three language levels and find that features targeting the level of morphology perform best in both aspects and can be used in future research to test hypotheses about genealogical relationships between languages or language families. The method is able to identify Turkic, Mongolic and Tungusic language families at the levels of morphology and syntax, whereas Japonic and Koreanic languages are not distinct enough in terms of their morphosyntax and are assigned to the same ancestry.

In summary we can state that the investigation of the stability of structural features is at the core of this dissertation. Phylogenetic comparative methods allow us to quantify the concept of stability and to compare the differences in stability across different sets of structural features. The methods from population genetics have proven helpful, especially when the genealogical tree, established as a method in historical linguistics, came to its limits. Even though only five families were used as a data basis for the analysis, the methodology can be transferred onto other language families, owing to the publicly available code and the utilization of a standard data format (cldf). One of the most important insights is that morphological features carry the most genealogical information, and these features could be used in the future to test relationships above the language family level. One possible improvement in the future studies would be an extension of the phonological features and a reassessment of their stability compared to other language domains.

Zusammenfassung

Strukturelle Merkmale haben das Potenzial, die Zeitschranke, nach der wir keine Hypothesen mehr über die Verwandschaft der Sprachen testen können, nach hinten zu verschieben. Allerdings müssen wir die Stabilität der strukturellen Merkmale kennen, um sie für solche Zwecke einsetzen zu können. In dieser Dissertation beschreibe ich das typologische Profil der Transeurasischen Sprachen, die später als meine Stichprobe für die Analyse der Stabilität dienen, konstruiere einen phylogenetischen Baum mit diesen Sprachen (Kapitel 2), messe die Stabilität der strukturellen Merkmale als phylogenetisches Signal und Evolutionsrate, rekonstruiere die Urzustände der strukturellen Merkmale (Kapitel 3) und verwende das Admixture Model aus der Populationsgenetik, um die Leistung phonologischer, morphologischer und syntaktischer Merkmale im Zuordnen der Sprachen zu ihren jeweiligen Sprachfamilien zu testen und das Niveau der Diffusion in diesen drei Sets zu untersuchen (Kapitel 4).

Im ersten Manuskript gebe ich eine allgemeine Vorstellung von den wichtigsten typologischen Gemeinsamkeiten zwischen den sogenannten Transeurasischen (Türkischen, Mongolischen, Tungusischen, Koreanischen, Japanischen) Sprachen und unterstütze diese mit Beispielen aus den entsprechenden Sprachen. Basierend auf der debattierten Annahme über die Verwandschaft dieser Sprachen baue ich einen phylogenetischen Baum von diesen fünf Sprachfamilien mithilfe der Bayes'schen Methoden der Baumkonstruktion. In der daraus resultierenden Stichprobe der Bäume, die eine große Wahrscheinlichkeit haben, gibt es eine klare Aufteilung in Altaische und Japanisch-Koreanische Sprachen, aber die meisten internen Zweige haben eine ziemlich niedrige Auflösung.

Im zweiten Manuskript gehe ich der Frage der Stabilität der strukturellen Merkmale auf den Grund und verwende zwei Maße, um diese zu untersuchen: phylogenetisches Signal und Evolutionsrate. Mehr als die Hälfte der strukturellen Merkmale weisen ein hohes phylogenetisches Signal auf und entwickeln sich im langsamen Tempo. Ich vergleiche die Stabilität über funktionale Kategorien, Wortarten und Sprachebenen hinweg und komme zum Schluss, dass die Markierung von Argumenten (Flagging und Indexing), Derivation und Valenz die stabilsten funktionalen Kategorien sind, die Pronomen und Nomen die stabilsten Wortarten, und Phonologie und Morphologie die stabilsten Sprachebenen. Zudem lässt sich ein Fünftel der Merkmale mit einer 95% Wahrscheinlichkeit als "Vorhanden" oder "Nicht vorhanden" in der Proto-Sprache rekonstruieren.

Da wir eine horizontale Übertragung bei den strukturellen Merkmalen nicht ausschließen können, müssen wir eine Methode verwenden, die in geeig-

neter Weise damit umgeht und immerhin gültige Ergebnisse liefert. Im dritten Manuskript wende ich eine Methode aus der Populationsgenetik (STRUCTURE) auf die Sets von strukturellen Merkmalen an, die Phonologie, Morphologie und Syntax umfassen. Ich vergleiche das Niveau der Beimischung und die Genauigkeit der Zuweisung der Sprachen zu den Sprachfamilien über die drei Sprachebenen hinweg und stelle fest, dass die Merkmale, die auf das Niveau der Morphologie abzielen, die beste Leistung in beiden Aspekten bringen und in der künftigen Forschung angewandt werden können, um die Hypothesen über genealogische Beziehung zwischen den Sprachen oder Sprachfamilien zu testen. Die Türkische, die Mongolische und die Tungusiche Sprachfamilien lassen sich auf den Ebenen der Morphologie und der Syntax identifizieren, wobei die Japanischen und die Koreanischen Sprachen aus morphosyntaktischer Sicht nicht ausreichend verschieden sind und der gleichen Abstammung zugeordnet werden.

Zusammenfassend lässt sich feststellen, dass die Untersuchung der Stabilität der strukturellen Merkmale im Mittelpunkt dieser Dissertation steht. Phylogenetische komparative Methoden erlauben es, das Konzept der Stabilität zu quantifizieren und die Unterschiede in der Stabilität über verschiedene Gruppen von strukturellen Merkmalen hinweg zu vergleichen. Die Methoden aus der Populationsgenetik haben sich besonders dann als hilfreich erwiesen, wo der in der historischen Linguistik etablierte Stammbaum als Methode an seine Grenzen kam. Obwohl nur fünf Sprachfamilien als Datengrundlage verwendet wurden, können die gleichen Methoden auf andere Sprachfamilien übertragen werden, dank der Veröffentlichung vom Code und der Verwendung von einem standardisierten Datenformat (cldf). Zu den wichtigsten Erkenntnissen gehört die Tatsache, dass morphologische Merkmale die meiste genealogische Information liefern, und diese Merkmale können künftig verwendet werden, um die Beziehungen über dem Niveau der Sprachfamilie zu testen. Eine mögliche Verbessrung in den künftigen Studien wäre eine Ausweitung der phonologischen Merkmale und eine Überarbeitung der Einschätzung ihrer Stabilität im Vergleich zu den anderen Sprachdomänen.

Erklärung

Ich erkläre,

(a) dass mir die geltende Promotionsordnung der Fakultät bekannt ist,

(b) dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe,

(c) dass mich ausschließlich die folgenden Personen bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben:

- Simon Greenhill - Betreuung, Beratung bei der Datenanalyse,
- Volker Gast - Betreuung, Korrekturlesen,
- Bettina Bock - Korrekturlesen, allgemeine Beratung,
- Ron Hüber - Korrekturlesen,

(d) dass die Hilfe einer kommerziellen Promotionsvermittlung nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,

(e) dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe,

(f) dass ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Jena, den 27. Oktober 2022

Nataliia Hübler

Acknowledgements

I would like to thank Alena Witzlack-Makarevich for motivating me to start my PhD journey, Alexander Savelyev for being my personal and professional advisor during the first years of my PhD, Ezequiel Koile for his support with Bayesian statistics, Simon Greenhill for his helpful methodological advice during the last years of my PhD, Volker Gast for agreeing to become my supervisor on a short notice, Michael Dunn for kindly agreeing to be my reviewer, anonymous reviewers of the articles included in the thesis for a constructive discussion, Ron and Naomi Hübler for their patience. The support by Russell Gray and Iren Hartmann was crucial to finishing this dissertation.