Phylogenetic analyses for the origin of sortal classifiers in Mongolic, Tungusic, and Turkic languages

[蒙古語、通古斯語和突厥語中分類詞起源的親緣演化分析]

Marc Allassonnière-Tang [唐威洋],[1]
Zhong-Liang Gao [高仲良],[2] Shen-An Chen [陳甚安][2] and One-Soon Her [何萬順][2,3]
[1] CNRS/MNHN/University Paris City [法國國家科學研究中心、法國國立自然史博物館及法國巴黎城市大學聯合實驗室] | [2] National Chengchi University [政治大學] | [3] Tunghai University [東海大學]

Numeral classifiers are one of the most common types of nominal classification systems. Their geographical distribution worldwide is concentrated in Asia, which infers a scheme of diffusion from a linguistic innovation. This study investigates the origin of classifier systems in the Mongolic, Tungusic, and Turkic languages in the Altaic region with a phylogenetic analysis based on data from 55 languages. The Single Origin Hypothesis suggests that Sinitic is the most probable original source of classifier systems found in Asia (Her & Li 2023). Under this hypothesis, classifiers are unlikely to be an indigenous feature of the Altaic region, and indeed their phylogenetic signal turns out to be weak. We also conduct a qualitative analysis on the classifier inventory of the studied languages to assess the robustness of phylogenetic methods. The results also indicate that classifiers are most likely a borrowed feature in the Mongolic, Tungusic, and Turkic languages.

Keywords: sortal classifier, Single Origin Hypothesis, Altaic, Mongolic, Tungusic, Turkic

關鍵詞： 分類詞、單一起源假設、阿爾泰語、蒙古語、通古斯語、突厥語

1.   Introduction

Numeral classifiers, along with gender/noun class, are the two essential types of nominal classification systems (Corbett 1991, Aikhenvald 2003, Seifart 2010, Grinevald 2015, Audring 2016), which constitute some of the most common and most important linguistic mechanisms to fulfil the need of the human mind to perceive and organize elements and experiences in a categorized scheme (Lakoff & Johnson 2003:162–163, Clahsen 2016:599). On one hand, an example of the gender/noun class is the masculine/feminine system found in languages such as Spanish, where nouns are assigned to one of the genders found in the language. Such a system has grammatical agreement with other elements of a sentence, e.g., adjectives. On the other hand, examples of numeral classifiers are the Mandarin sortal classifier *tiáo* for entities with a long shape, as in *sān tiáo yú* (three CLF.LONG fish) 'three fish' and the mensural classifier *bàng* 'pound', as in *sān bàng yú* (three pound fish) 'three pounds of fish'. In the current study, we only consider numeral sortal classifiers, which are defined as independent morphemes or affixes that categorize nouns according to the inherent features of their referents based on criteria such as shape, consistency and animacy (Allan 1977, Grinevald 2000:71, Her & Hsieh 2010, Kilarski & Allassonnière-Tang 2021).

In the following text, we thus use the term 'numeral classifier' to refer to 'numeral sortal classifier'. In terms of geographical distribution, the worldwide prevalence of such systems is indicated in two surveys in the World Atlas of Language Structures Online (WALS, Dryer & Haspelmath 2013), i.e., gender/noun class systems: 43.6%, 112/257 (Corbett 2013) and classifier systems: 35%, 140/400 (Gil 2013).

While numeral classifiers have been extensively studied from various perspectives, such as syntax, semantics, discourse, and cognition, the origin of their geographical concentration has yet to be confirmed. One existing hypothesis suggests that numeral classifiers in Asia emerged first in the Sinitic language group and spread to nearby languages, e.g., Tai-Kadai, Austroasiatic, Tibeto-Burman, etc. (Her & Li 2023). The details of this Single Origin Hypothesis are discussed in Section 2. However, for Altaic classifier languages, only 6 are included and largely left for future research in Her & Li (2023). If this hypothesis is on the right track, then classifiers in the Altaic region are unlikely to be indigenous either and are probably a result of diffusion. Based on Hölzl & Cathcart (2019) and Chen, Allassonnière-Tang & Her (to appear), this paper further investigates the phylogenetic signals of classifiers in three well-established individual language groups, Mongolic, Tungusic, and Turkic (abbreviated as MTT hereafter). We also consider the phylogenetic tree regrouping these three language groups as if the three share a common root (Proto-MTT). Our purpose is to determine whether

classifiers can be justified at Proto-MTT. In Section 3 we present the 55 MTT languages covered in this study. Section 4 then discusses the phylogenetic comparative methods employed and the results of the analyses. In Section 5 we review a qualitative analysis on the classifier inventory of the Mongolic, Tungusic, and Turkic languages to further assess the robustness of the phylogenetic methods. In Section 6 we conclude that the two analyses suggest that classifiers are most likely not present at Proto-MTT.

2.   The single origin hypothesis

As mentioned, an important motivation behind this study was to examine whether classifiers in languages of the Altaic region are consistent with the Single Origin Hypothesis proposed in Her & Li (2023) based on data of 490 numeral classifier languages worldwide, as shown in Figure 1.
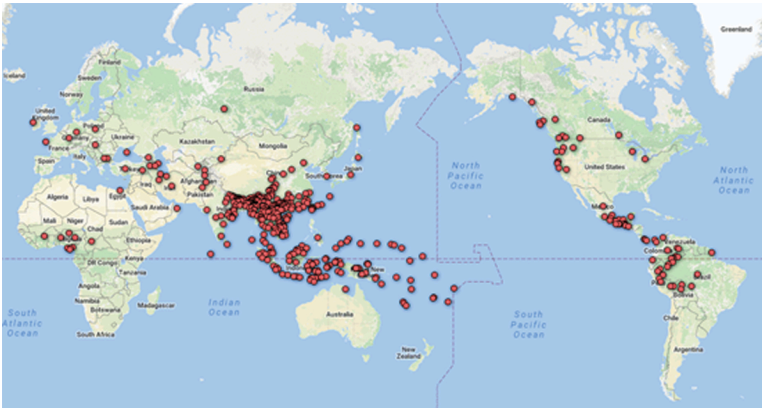

Figure 1. 490 classifier languages in the world (Her & Li 2023: 22)
(Map from WLMS 16 www.gmi.org/wlms. ©2016 Google – Map data ©2016)

Observing the pattern of distribution, where the classifier feature seems to have a clear center of clustering in East and Southeast Asian languages and radiates outward in all directions with classifier languages gradually thinning out and with less intensive use of classifiers (Gil 2013, Her, Hammarström & Allassonnière-Tang 2022), Her & Li (2023) contend that this distribution pattern is by no means accidental and is more likely due to the diffusion of the this feature from the center, which leads to the Single Origin Hypothesis. This hypothesis concerns only classifier languages in Asia and the Pacific and thus excludes classifier languages in the rest of the world, e.g., Europe, Africa, Papua New Guinea,

and the Americas. It investigates exactly which language group is the alleged original source of numeral classifiers in Asia and the Pacific. The current results show that Sinitic and Tai are the most likely candidates, and that the available evidence favors Sinitic as the innovator and Tai as a borrower.

According to the Single Origin Hypothesis in its current form, the classifier feature first emerged in Sinitic languages in northern China and all the other classifier languages in Asia and the Pacific acquired this feature via language contact. While fully acknowledging the highly speculative nature of this ambitious hypothesis, the proposers have provided some tentative and circumstantial evidence for a number of language families or groups, including Miao-Yao, Austroasiatic, Tai-Kadai, Tibeto-Burman, Indo-Arya, Dravidian, and Austronesian. However, for languages in the Altaic region (i.e., the MTT languages), they only conjecture that, given a sparse and sporadic number of classifier languages in this region, they have likely acquired classifiers from other language families. No further quantitative analysis has been conducted to investigate the origin of classifiers in MTT languages, which inspires our study to investigate the development of classifiers in MTT languages from a quantitative and qualitative perspective.

On one hand, if quantitative methods show that classifiers are ancestral to MTT languages, then additional analyses are required to further investigate how the time frame matches with the development of classifiers in surrounding language families and how the further findings are consistent with or contradict the Single Origin Hypothesis. On the other hand, if quantitative methods show that classifiers were not present at the root of MTT languages or that classifiers are likely to have been imported in MTT languages, then the findings are consistent with the Single Origin Hypothesis. In both scenarios, the quantitative analysis also needs to be supported by comprehensive surveys that consider classifiers found in individual languages from a linguistic perspective.

3.   Data

Our primary goal is to demonstrate that a classifier system is unlikely to be found at the root of the MTT languages. The root of the MTT family, or the proto-MTT language, though controversial, enjoys the ardent support of a number of researchers, which suggest a common heritage based on lexical, morphological and structural comparisons (Robbeets 2015, Robbeets & Savelyev 2020, Robbeets et al. 2021). Most notably, in a recent study, Robbeets & Bouckaert (2018) use lexical etymologies with Bayesian phylogenetic inference to Mongolic, Tungusic, Turkic, Japonic, and Koreanic languages. They find strong support for Tungusic, Mongolic, and Turkic as a unity and Japano-Koreanic as a separate unity. Note

that Robbeets & Bouckaert (2018) call this Mongolic-Tungusic-Turkic group 'Altaic'. However, to avoid unnecessary confusion, we shall use the more transparent term 'Proto-MTT' to refer to the common ancestral language of the three MTT language groups.

The current study does not include Japonic and Koreanic languages for two reasons. First, Robbeets & Bouckaert's (2018) phylogenetic support for Japon-Koreanic languages as a separate unity is even stronger than that for the MTT group. Second, the existing literature clearly shows that both Japanese and Korean have borrowed substantially from the Chinese numeral system and numeral classifiers; however, indigenous classifiers and an indigenous numeral system also exist alongside the Sino-Japanese and Sino-Korean numeral systems and classifiers. Research on the origin of classifiers in the two languages is rather scarce; there is thus no consensus whatsoever as to whether indigenous classifiers were already in use prior to the loans from their Sinitic neighbors. This issue thus warrants a separate in-depth study and is beyond the scope of the current paper. Nevertheless, we still provide an overview of the classifier systems in Japanese and Korean in Supplementary material 1,[1] to allow readers a better understanding of this theoretical choice.

Our data consists of a convenience sample of 55 languages, drawn from WACL (The World Atlas of Classifier Languages) (Her, Hammarström & Allassonnière-Tang 2022), including 15 Mongolic languages, 9 Tungusic languages, and 31 Turkic languages. This ratio is thought to be an acceptable reflection of the total size of the language families which is 15, 13, and 44 languages (Hammarström, Forkel & Haspelmath 2019). Languages not included in the data had to be removed due to a lack of available data. For each language, grammars, sketches, published chapters, and/or articles were consulted to identify the presence/absence of classifiers. As a soft recall, the term 'classifier' as used in this paper refers to numeral sortal classifiers as mentioned in Section 1. As an example, Kirghiz (Turkic, Glottocode: kirg1245) is considered to be a classifier language since available sources provide examples of classifiers that match our definition, as shown in (1). For example, the classifier *tujaq* highlights the 'head' feature of a horse and is used in an enumeration context. As an additional note, the sources we consulted also mention that classifiers in Kirghiz are optional. Nevertheless, in our data, we do not differentiate between obligatory and optional classifiers (Gil 2013). If a language uses classifiers, whether one or several classifiers and whether obligatory or optional, the language is annotated as having classifiers.

---

1. The link to the supplementary materials is provided at the end of the paper in the Data Availability Statements section, before the References.

(1)   Examples of classifier in Kirghiz, Turkic   (Hu 1986: 51)
    a.   bir daana   qalem
        one CLF.GEN pencil
        'one pencil'
    b.   bir tujaq   at
        one CLF.HEAD horse
        'one horse'

As an opposite example, Bonan (Mongolic, Glottocode: bona1250) is annotated as not having classifiers in our data, since a scan of available sources indicated that it does not have sortal numeral classifiers and only has structures similar to mensural classifiers, which are based on containers such as cup and bowl (Fried 2010: 141–142). As shown in (2), the classifier-like structures found in Bonan are not sortal classifiers since they convey an information of quantity and cannot be removed from a clause without affecting its meaning, cf. *five noodles* vs. *five bowls of noodles*. The numerals just indicate the number of containers but not the number of nouns, which makes them different from sortal classifiers.[2]

(2)   Examples of mensural classifiers in Bonan, Mongolic   (Fried 2010: 141–142)
    a.   tʰəmtʰoχ dzəjal tʰaun
        mianpian bowl five
        'five bowls of mianpian noodles'
    b.   rakə   dampe ʁuraŋ
        alcohol bottle three
        'three bottles of beer'

A scan for each of the 55 languages resulted in 35 languages without, and 20 languages with, classifiers. Amongst the classifier languages, we found three Mongolic languages (20%, 3/15), three Tungusic languages (33%, 3/9), and 14 (45%, 14/31) Turkic languages. A full list of the included languages and their metadata is available in the supplementary materials. The language name and its affiliated Glottocode are both included to facilitate comparative analyses. Then, the presence/absence of classifiers is annotated, with 1s representing the presence of classifiers and 0s indicating the absence of classifiers. Finally, the phylogenetic affiliation of the language and its geographical coordinates are included (The raw table is included in Supplementary Material 2). A geographical overview of the

---

2. This definition of classifier languages is widely adopted in the literature, e.g., Gil (2013) and Her, Hammarström & Allassonnière-Tang (2022), and is thus without controversy. Furthermore, for the purpose of the current paper, i.e., to identify the origin of sortal classifiers in MTT languages, it is more important to identify sortal classifiers in these languages than to determine what qualifies a language to be a real classifier language.

55 languages is provided in Figure 2. Each point in the map is extracted from Glottolog and indicates the geographical center point of the geographical area in which the speakers of a language are located. This point may also consider the historical location and/or the demographic center point of language speakers depending on their currently reported number.
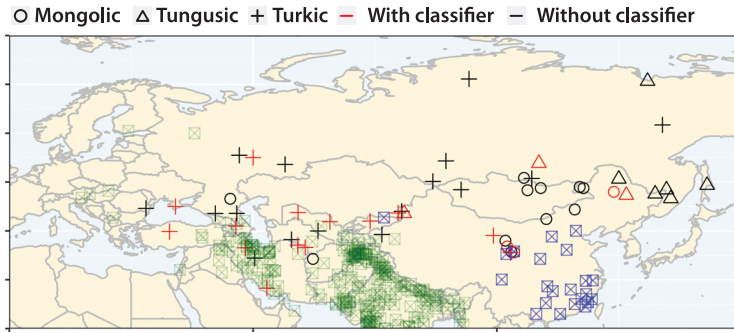

Figure 2. A geographical overview of the 55 languages included in the data. Neighboring Indo-Iranian languages (green) and Sinitic languages (blue) are shown in shaded colors

The geographical distribution of MTT languages with numeral classifiers matches findings reported in previous studies (e.g., Chen, Allassonnière-Tang & Her (to appear)). Classifier languages are not rare in MTT languages, however, they are not the majority, representing one third of the data (36%, 20/55). Most of these classifier languages are located in proximity to Sinitic or Indo-Iranian languages, which are considered to be the two most likely sources of classifier systems in the area.

4.   Phylogenetic analysis

To provide a quantitative diachronic analysis of classifiers in MTT languages, we used phylogenetic comparative methods. These methods originate from evolutionary biology and have been imported by historical linguistics because they can address the non-independence of linguistic features during evolutionary processes (Galton's problem, Mace & Holden 2005). Such an analysis typically consists of three main steps (Allassonnière-Tang & Dunn 2020). First, the analysis involves the use of a phylogeny, i.e., a family tree of languages. Such a tree or sample of trees are created from historical linguistic knowledge and Bayesian phylogenetic inference. A language family tree can then be used to infer the evolutionary process of a linguistic feature within a language family. Second, the phylogenetic

signal of a selected linguistic feature is measured on the tree. A strong signal shows that the trait's evolution is predictable under Brownian motion. A weak signal can point to different scenarios. It could indicate that the diversity of the feature has been affected by non-evolutionary factors such as contact with languages of other families. However, it could also mean that the selected feature is evolving extremely quickly, or that a hidden state probably exists, among others. Third, if the phylogenetic signal is strong, a further analysis can be conducted to infer the evolutionary dynamics of the feature. Additional explanations of this methodology are explained in the following sections. The Bayesian phylogenetic inference was conducted using MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003), and the results assessed with Tracer (Rambaut et al. 2018), while the phylogenetic comparative analyses are carried out with the Phangorn (Schliep 2011), Phytools (Revell 2012), and Ape (Popescu, Huber & Paradis 2012) packages in R 4.1.2 (R Core Team 2021). The used code is provided in Supplementary Material 3.

4.1   Tree inference

Two samples of trees were considered for the phylogenetic analysis. First, we considered the tree sample of trees generated by Robbeets & Bouckaert (2018). The trees of the sample were pruned to keep only the languages included both in both the tree sample and our data set. This filter resulted in a smaller sample of 29 languages (7 Mongolic, 6 Tungusic, and 16 Turkic languages). As a way to visualize the overall shape of the tree, the sample of trees can be summarized with a *Maximum Clade Credibility tree*. To obtain such a tree, we went through the sample of trees and evaluate the likelihood of each branch based on how often a branch occurred. Then, each tree was assigned a credibility score based on the product of the likelihoods of each branch in each tree. The tree with the highest credibility score was the maximum credibility tree and considered to be the representative tree for the entire sample. The maximum clade credibility tree of the sample of trees from Robbeets & Bouckaert (2018) is shown in Figure 3.

Second, to avoid a coincidental bias resulting from the filter based on the language sample from Robbeets & Bouckaert (2018), we used Bayesian Monte Carlo Markov Chain (MCMC) phylogenetic inference to create a sample of trees to match with the genealogical constraints derived from the Glottolog reference phylogeny and the existing literature (A full list of constraints is provided in the Supplementary Material 4). For the branching structure between the Mongolic, Tungusic, and Turkic language groups, we followed the findings reported by Robbeets & Bouckaert (2018), who show that the Tungusic group is expected to be an outgroup, followed by a binary branch that consists of the Mongolic and
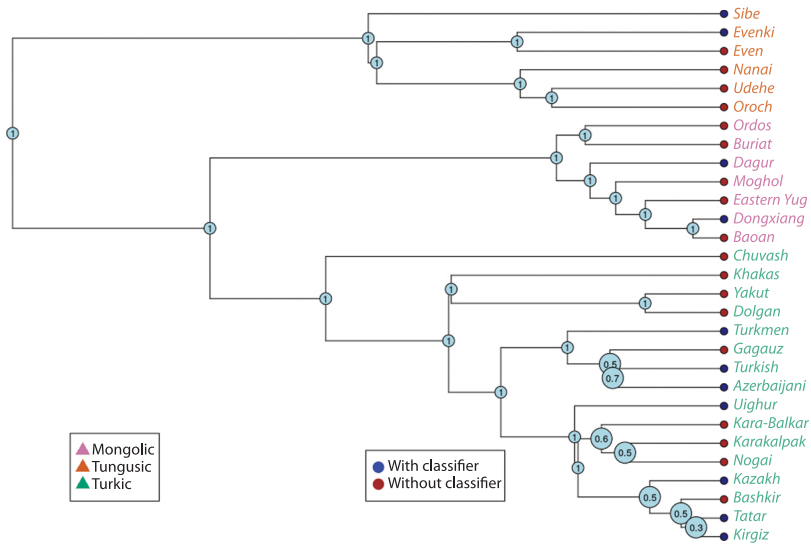
Figure 3. The maximum clade credibility tree extracted from the sample of trees from Robbeets & Bouckaert (2018). The numbers on the nodes of the tree indicate the posterior probabilities

Turkic groups. At the same time, the internal branching structure of the Mongolic, Tungusic, and Turkic groups was consistent with the branching structure as specified in Glottolog. The constraints from the Glottolog tree and from Robbeets & Bouckaert (2018) were the only priors used when generating the sample of trees. Based on these constraints, we generated a sample of trees in such a way that each tree in the sample represents a historical hypothesis consistent with the established historical linguistic understanding of how these languages are related. The sample as a whole also gives a measure of uncertainty where the Glottolog tree is unresolved (and where existing tree data diverge). As an example, the three Mongolic languages Halh Mongolian, Oirad-Kalmyk, and Peripheral Mongolian follow the same branch in the Glottolog Mongolic tree, but the sub-branch does not specify their hierarchy. That is to say, Glottolog does not specify if Halh Mongolian and Oirad-Kalmyk form a binary branch first, with Peripheral Mongolian as an outgroup, or other combinations. This uncertainty was captured by the sample of the trees, as trees with most possible combinations were included in the sample. Proportional branch lengths were computed using the Grafen method, in the function *compute.brlen* in the *ape* package of R (R Core Team 2021).

The set of theoretically possible trees was generated via Markov Chain Monte Carlo (MCMC) methods (Metropolis et al. 1953), applying the following principles: First, the sampling starts with an arbitrarily chosen tree. Then, in each gen-

eration of the algorithm, this tree is modified and compared with the original tree in terms of likelihood. The MCMC starts from a randomly selected tree with a low likelihood; then, the MCMC changes one small component of the tree at a time and generates trees with a higher likelihood toward the data. This process of generations commonly starts with a burn-in period during which the stationarity is not reached yet. The generations of this period are thus excluded from the dataset. The burn-in phase is typically indicated by the rapid increase in likelihood at the beginning of the MCMC search, and the cut-off point is decided when the likelihood reaches a plateau. Thus, the more generations performed, the more likely it is to simulate evolution under each possible tree in proportion to the posterior probability of the particular scenario. An adequate amount of generations is commonly evaluated via the average standard deviation of split frequencies, which shows the convergence of the sample. In terms of parameters in MrBayes, one run with four chains is conducted. The MCMC run following the constraints from Glottolog resulted in 3000000 generations, among which 1500000 were considered as burn-in and are thus removed. Then, one tree was extracted for each 1500 trees, resulting in a final sample of 1000 trees (mean effective sample size = 1000). The final sample of trees is available in Supplementary Material 5. The maximum credibility tree of the tree sample is shown in Figure 4.
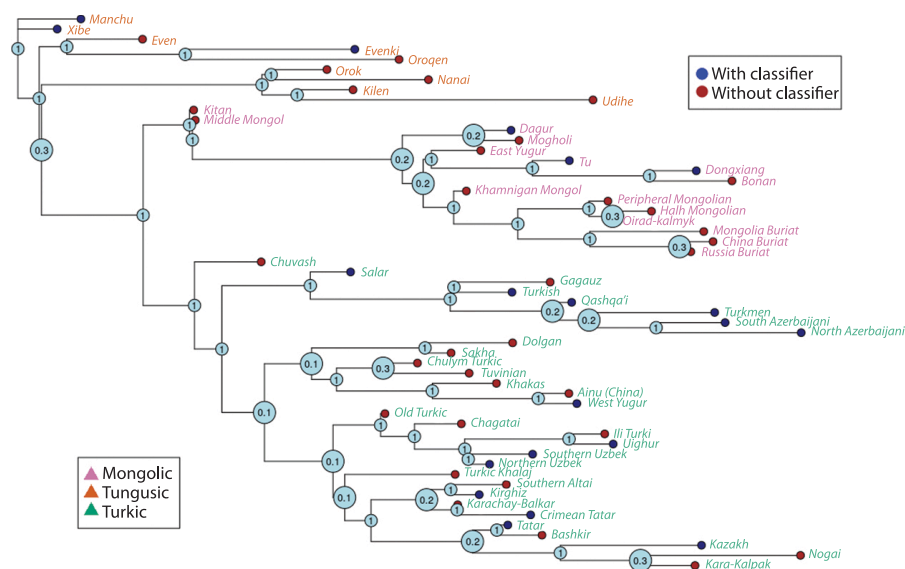

Figure 4. The maximum clade credibility tree extracted from the sample of 1000 trees generated based on Glottolog constraints. The numbers on the nodes of the tree indicate the posterior probabilities (a probability of 1 means a Glottolog constraint)

The tree conformed to our expectations in two ways. One one hand, languages from the Mongolic, Tungusic, and Turkic groups are not mixed together. Furthermore, the hierarchy of the groups is as we specified when generating the sample of trees: The Tungusic group is an outgroup, followed by the Mongolic and Turkic groups forming a binary branch. On the other hand, the probabilities of the branches match with our settings. As an example, the branching between Even, Evenki, and Oroqen is fixed in Glottolog, i.e., Evenki and Oroqen are first under the same branch, while Even is an outgroup. Each tree in the sample of 1000 trees thus has these two branches, resulting in a probability of 1 (100%) for their occurrence. With regard to the non-constrained nodes, we observe that the uncertainty is rather high, as the posterior probability is low on most of the nodes, being below 0.5. This infers that the probability of ancestral state estimations for the presence/absence of numeral classifiers is bound to be low. Nevertheless, since checking the MCC tree showed that the specified constraints had been applied correctly, we proceeded to measure the phylogenetic signal of classifiers in the sample of trees.

4.2 Testing for phylogenetic signal

We first assess the strength of the phylogenetic signal indicating the presence/absence of classifiers, to provide a statistical measure of how this presence/absence matches with the phylogeny proposed by historical linguistics and Glottolog. As an example, Figure 4 shows the MTT tree of our study with tip labels colored according to the presence/absence of classifiers in each language. The phylogenetic signal is considered to be strong if the tendency of related languages to resemble each other is stronger than languages drawn at random from the same tree. If languages on the same branch mostly share the same value, it means that the presence/absence of classifiers is easily predictable, which in turn infers a strong phylogenetic signal of classifiers in the tree. However, if languages on the same branch tend to have different values, it means that the presence/absence of classifiers is not easily predictable, which infers a weak phylogenetic signal of classifiers in the tree. A weak signal can be interpreted as the impact of an external influence which has disrupted the inheritance of the trait in the tree.

The measurement of the phylogenetic signal and its testing for statistical significance were conducted with the D statistics (Fritz & Purvis 2010). This measure is typically used to measure the phylogenetic signal of binary traits in a tree, which fits our data and purpose. A D close to 1 indicates that the targeted feature has a distribution that is random compared with the phylogeney. A D close to 0 indicates that the targeted feature matches the Brownian motion model of evolution compared with the phylogeny. In cases where the measured trait is more dis-

persed than randomly expected, the D statistics are greater than 1. Vice-versa, if the measured trait is more phylogenetically conserved than as expected under the Brownian motion model of evolution, the D statistics are smaller than 0. Since we have a sample of trees for both trees from Robbeets & Bouckaert (2018) and the trees generated based on the Glottolog constraints, the D statistic and its statistical significance for classifiers is measured on each tree of the samples. The results are shown in Figure 5.
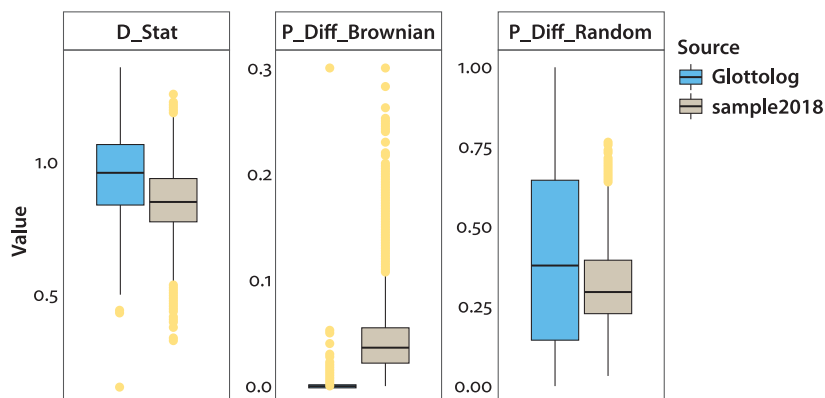
Figure 5. The distribution of the D statistic and its departure from random association (P_Diff_Random) and the clumping expected under a Brownian evolution threshold model (P_Diff_Brownian) based on the two samples of trees from the Mongolic, Tungusic, and Turkic languages. The dashed lines indicate the significance level of .05 for *p* values

The measured D statistic is close to one for the majority of the trees in both samples. Taking the sample from Robbeets & Bouckaert (2018) as an example, the mean and the median of the measured D statistic are 0.85 (The first quartile is 0.77 and the third quartile is 0.94). This shows that within all trees of the sample, the phylogenetic signal is generally weak. That is to say, even when considering different trees, the presence/absence of classifiers in MTT languages is not predictable from the phylogeny. This can be interpreted in different ways. One the one hand, a D statistic close to 1 is consistent with a scenario of substantial horizontal transmission. In other words, classifiers are likely to have been borrowed in MTT languages. This distribution of weak signals is significantly different from the expected clumping under a Brownian evolution, and not significantly different from a random association, which confirms that the observed phylogenetic signal is weak in a statistically significant way. On the other hand, borrowing is not the only possible explanation for this result. For example, classifiers could have

evolved in a tree-like fashion in MTT languages, just not according to a Brownian motion. Furthermore, a low signal can also be evidence of parallel independent innovation. As an attempt to cover the latter possibilities, we conducted a preliminary analysis inferring the ancestral states of classifiers at the root of MTT languages using both tree samples (included in Supplementary Material 3). The results match with the observed phylogenetic signal. The current tree samples do not provide sufficient information to distinguish if the root of MTT languages had classifiers or not, as the probability of having classifiers at the root of MTT languages is 50%, while the probability of not having them is also 50%.

As a summary, since the phylogenetic signal of classifiers is weak in the MTT languages, we cannot infer the ancestral state nor the transition rates of classifiers in these languages.[3] In our case, two possibilities can be considered: either classifiers were present at the roots of the trees, but were lost in some languages due to contact with non-classifier languages, or classifiers were not present at the roots of the trees, but were acquired in some languages due to contact with classifier languages. As we will show in the following section, the latter hypothesis is more likely.

5.   Qualitative analysis

Since the phylogenetic signal of classifiers was found to be weak in the MTT languages, two main hypotheses need to be considered. On one hand, it is possible that classifiers were present at the root of MTT languages and were lost due to contact. On the other hand, it is also possible that classifiers were not present at the root of MTT languages and were acquired due to contact. We assess the probability of both hypotheses based on the existing literature and a deeper analysis of our data.

A search of the existing literature favors the second hypothesis that classifiers were not present at the respective root of the MTT languages. First, in our research we have not come across any published claims that the classifier feature is indigenous in any of the MTT languages. Furthermore, analyses conducted separately on Mongolic, Tungusic, and Turkic languages show that classifiers in those languages are often borrowed from languages of neighboring languages with classifier systems. For instance, the sortal classifiers found in Tungusic languages such as Manchu are likely to have been collective numeral suffixes, which experienced

---

3. We also considered conducting phylogenetic analyses on Mongolic, Tungusic, and Turkic languages separately. However, the small sample size of the Mongolic and Tungusic languages would make the results less robust for such an analysis.

reinterpretation as classifiers under Mandarin Chinese influence (Alonso de la Fuente 2017: 76). As another example, most of the MTT classifier languages have optional classifiers (Robbeets 2017: 11), as shown with Bonan in (2) and other languages, which is usually the case when classifier systems are not firmly established.

More importantly, a recent qualitative survey by Chen, Allassonnière-Tang & Her (to appear) based on examples from 66 MTT languages identified 14 classifiers languages: two Mongolic languages (Daur and Monguor), three Tungusic languages (Manchu, Evenki, and Xibe), and nine Turkic languages (Azerbaijani, Crimean Tatar, Kazakh, Salar, Tatar, Uyghur, Uzbek, Turkish, West Yugur). A close comparison of etymology and word form has shown that the origin of classifiers in MTT languages has two main possible sources: Persian and Mandarin Chinese.

With regard to classifier languages in Mongolic, taking Monguor as an example, this language has two numeral systems (Slater 2003: 94). The Monguor numeral system has almost been entirely replaced by the Chinese system. Nowadays, only the numbers 1 *nige* and 2 *ghu* are still being used. In contexts of enumeration, Monguor numerals are used without numeral classifiers (3a), while borrowed Chinese numerals are used with classifiers (3b). The Monguor classifiers are few and mostly borrowed from Chinese (Slater 2003: 95–96). For example, the most common classifiers are the general classifier *ge* and the classifier for long objects *tiao*, which are identical with their Chinese counterparts.

(3) Examples of classifier in Monguor, Mongolic   (Slater 2003: 106,184)
    a.   nige loahan
        one oldman
        'an oldman'
    b.   yi-ge   laohan
        one-CLF.GEN oldman
        'an oldman'

As an additional evidence to the borrowing of classifiers from Chinese to Monguor, such borrowings are also found in other linguistic domains. For instance, the original word order of Monguor was [Noun Numeral], with the noun preceding the numeral. However, with the borrowing of classifiers, the order [Numeral CLF Noun] was also borrowed, which is gradually changing the word order of nouns and numerals in Monguor. Furthermore, the Monguor language has also borrowed a lot of vocabulary from Chinese, including nouns, adjectives, and function words (Junastu 1981). Such frequent borrowings show that classifiers in Monguor were most likely borrowed from Chinese.

A similar situation regarding the Chinese influence on classifiers is found in Manchu, Evenki, and Xibe, the three Tungusic classifier languages identified in

Chen, Allassonnière-Tang & Her (to appear). Manchu, with a geographical area closest to China and the longest history of interaction with Chinese among MTT languages, has the most classifiers among the MTT languages, as expected. The example in (4a) shows the Manchu classifier *giyan* for houses and rooms, a clear borrowing of the Chinese classifier *jian*,. The Xibe example in (5) and the Manchu example in (4b) in turn show the Manchu influence on Xibe.

(4) Examples of classifiers in Manchu, Tungusic   (Wang 2005: 133, 289)

    a.   əm-giyan   bo
            one-CLF.HOUSE house
            'a house'

    b.   əm-da   gauliang
            one-CLF.PLANT sorghum
            'a sorghum plant'

(5) Example of classifiers in Xibe, Tungusic   (Li & Zhong 1986: 55)
    əm-da   xɛlin
    one-CLF.PLANT tree
    'a tree'

As for classifiers in Turkic languages, taking Turkish as an example, it has five optional classifiers: *tane* (CLF.GEN), *kişi* (CLF.HUMAN), *baş* (CLF.ANIMAL), *parça* (CLF.2D), and *aded* (CLF.GEN.FORMAL) (Underhill 1980, Kornfilt 1997). On one hand, the optionality and the small inventory size of these classifiers already hint at a borrowed system. On the other hand, these classifiers share similarities with their counterparts in Persian, which also points in the direction of borrowing from Persian. As an example, the most commonly used classifier is the general classifier *tane*, which has a similar pronunciation and function to the Persian general classifier *dane*. Other examples of similarities in terms of pronunciation and function can also be found in other classifiers, for example, the classifier *adet*, used as for *tane* but only in formal speech, is similar to its Persian counterpart [ʕadad] (van Schaaik 1996). As an additional evidence to the borrowing of classifiers from Persian to Turkish, the two languages are found in a close geographical proximity and vocabulary borrowings from Persian to Turkish are also frequent.

In general, classifiers are not native to any of the Mongolic, Tungusic, and Turkish language groups and are thus not likely to be present at the root of the MTT family. These facts also match with our findings from the phylogenetic analysis, i.e., since classifiers have most likely been introduced in the MTT languages through language contact with either Persian or Chinese, it is expected that the phylogenetic signal of classifiers is rather weak.

We also conducted a closer analysis on languages within our data. Since some languages have already been discussed in Chen, Allassonnière-Tang & Her (to

appear), we focus on languages that were not included in the 14 classifiers languages mentioned in that study, i.e., Dongxiang (Mongolic), Kirghiz (Turkic), Qashqa'I (Turkic) and Turkmen (Turkmen). The two other different languages refer to a variation of North and South Uzbek and Azerbaijani, which we do not discuss here. First, Dongxiang is spoken in the western part of China, and is attested to have borrowed classifiers (and also lexical items) from Mandarin Chinese (Field 1997: 240, 247, 264). Second, classifiers in Kirghiz generally have a pronunciation similar to Persian. As an example, in *bir daana qalem* (one CLF.GEN pen) 'one pencil', the general classifier *daana* is phonologically similar to the Persian general classifier *dane*. Third, in Qashqa'i, which is spoken by a minority ethnic group in Iran, classifiers are also similar to Persian classifiers (Soper 1987). For instance, the human classifier *tān*, which literally means 'body', is phonologically similar to its Persian equivalent *ta*. Fourth, the general classifier of Turkmen *da:ne* is also phonologically similar to its Persian counterpart *dane*. In short, our qualitative analysis of available matches the literature. Classifiers in languages of the Altaic region tend to have been borrowed from neighboring classifier languages, more specifically from Persian or Mandarin Chinese. Furthermore, the effect of contact is also found in non-classifier languages. For example, Bonan is a non-classifier language. Nevertheless, its numeral phrase word order has been influenced by Amdo Tibetan, which is a classifier language (Duojie 2005). This could suggest that the mensural classifiers found in Bonan could in fact be the starting point of a new classifier system which will eventually develop sortal classifiers.

6.   Conclusion

This study aimed at assessing the likelihood of classifier systems at the root of the MTT family being a unity, our hypothesis being that classifiers in MTT languages are all from a borrowing origin. To accomplish this, we conducted two analyses. First, we used Bayesian phylogenetic inference to measure the phylogenetic signal of classifiers in MTT languages. Our results show that this phylogenetic signal is weak in MTT languages. This indicates that two main possibilities are likely: either classifiers were present at the root of MTT languages and were lost due to contact or classifiers were not present at the root of MTT languages and were acquired due to contact. Then, we provide a qualitative analysis based on a review of the literature to assess which of the two options is the most likely. Our results show that since classifiers found in MTT languages can mostly be traced to an origin of borrowing from either Chinese or Persian, the probability of classifiers at the root of MTT languages is low. These results are thus compatible with the find-

ings of Hölzl & Cathcart (2019) and Chen, Allassonnière-Tang & Her (to appear) and consistent with the Single Origin Hypothesis.

In terms of further research, regarding the method, the current paper used a binary coding of classifier systems in terms of presence/absence. A more fine-grained analysis could be expected to shed more light on the development of classifier systems in MTT languages. For example, future studies could consider the contrasts of obligatoriness/optionality and sortal/mensural nature of classifiers in their quantitative models. Ideally, the phylogenetic signal could also be measured for each classifier individually to evaluate if it is stronger for some classifiers than others. If that is indeed the case, phylogenetic methods of correlated evolution could also be used to infer in which order classifiers developed in MTT languages. Furthermore, as Japonic and Koreanic languages were not included in the current paper, separate in-depth qualitative analyses could be conducted for these languages. Additional research is also needed to investigate the origin of classifiers in Persian. The same analysis could also be replicated on language groups and families found in neighboring geographical areas. For example, classifiers are also found in Austronesian languages. It would thus be relevant to assess the phylogenetic signal and the probability at the root of classifiers in these languages.

Funding

Data availability statements

Supplementary materials including the data and code underlying this article are available in OSF (Open Science Framework) at https://osf.io/u6wfg/?view_only=None, and can be accessed directly.

Acknowledgements

List of abbreviations

CLF	classifier	CLF.LONG	long classifier
CLF.GEN	general classifier	CLF.PLANT	plant classifier

References

Aikhenvald, Alexandra. 2003. Mechanisms of change in areal diffusion: New morphology and language contact. *Journal of Linguistics* 39:1–29.

Allan, Keith. 1977. Classifiers. *Language* 53.2:285–311.

Allassonnière-Tang, Marc, and Michael Dunn. 2020. The evolutionary trends of grammatical gender in Indo-Aryan languages. *Language Dynamics and Change* 11.2:211–240.

Alonso de la Fuente, José Andrés. 2017. From converb to classifier? On the etymology of Literary Manchu nofi. *Essays in the History of Languages and Linguistics. Dedicated to Marek Stachowski on the Occasion of his 60th Birthday*, ed. by Michał Németh, Barbara Podolak and Mateusz Urban, 57–80. Cracow: Księgarnia Akademicka.

Audring, Jenny. 2016. Gender. *Oxford Research Encyclopedia of Linguistics*, ed. by M. Aronoff. Oxford: Oxford University Press. Retrieved August 1, 2022, from

Chen, Shen-An, Marc Allassonnière-Tang, and One-Soon Her. (to appear). On the distribution and origin of sortal classifiers in Altaic languages. *Journal of Chinese Linguistics*.

Clahsen, Harald. 2016. Contributions of linguistic typology to psycholinguistics. *Linguistic Typology* 20.3:599–614.

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2013. Number of genders. *The World Atlas of Language Structures Online*, ed. by Matthew Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved August 15, 2022, from http://wals.info/chapter/30

Dryer, Matthew, and Martin Haspelmath (eds.) 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Duojie, Dongzhi. 2005. Zangyu Anduo fangyan zhong de liangci [The classifiers in Amdo Tibetan]. *Hanzangyuxi Liangci Yanjiu* [Studies of classifiers in Sino-Tibetan], ed. by Jinfang Li and Suhua Hu, 47–54. Beijing: Central University for Nationalities Press.

Field, Kenneth. 1997. A Grammatical Overview of Santa Mongolian. Doctoral dissertation, University of California, Santa Barbara, CA.

Fried, Robert. 2010. A Grammar of Bao'an Tu, a Mongolic Language of Northwest China. Doctoral dissertation, State University of New York at Buffalo, New York, NY.

Fritz, Susanne, and Andy Purvis. 2010. Selectivity in Mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24.4:1042–1051.

Gil, David. 2013. Numeral classifiers. *The World Atlas of Language Structures Online*, ed. by Matthew Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved August 15, 2022, from http://wals.info/chapter/55

Grinevald, Colette. 2000. A morphosyntactic typology of classifiers. *Systems of Nominal Classification*, ed. by Gunter Senft, 50–92. Cambridge: Cambridge University Press.

Grinevald, Colette. 2015. Linguistics of classifiers. *International Encyclopedia of the Social & Behavioral Sciences*, ed. by James Wright, 811–818. Oxford: Elsevier.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 4.1.* Jena: Max Planck Institute for the Science of Human History. Retrieved August 15, 2022, from https://glottolog.org/

Her, One-Soon, Harald Hammarström, and Marc Allassonnière-Tang. 2022. Defining numeral classifiers and identifying classifier languages of the world. *Linguistics Vanguard* 8.1:151–164.

Her, One-Soon, and Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11.3:527–550.

Her, One-Soon, and Bing-Tsiong Li. 2023. A single origin of numeral classifiers in Asia and Pacific: A hypothesis. *Nominal Classification in Asia: Functional and Diachronic Perspectives*, ed. by Marc Allassonnière-Tang and Marcin Kilarski, 113–160. Amsterdam: John Benjamins.

Hölzl, Andreas, and Chundra Cathcart. 2019. Sortal numeral classifiers in Central Asia. Paper presented at the 13th Conference of the Association for Linguistic Typology, University of Pavia, Pavia. Retrieved August 1, 2022, from https://www.academia.edu/40273362/Sortal_numeral_classifiers_in_Central_Asia

Hu, Zhen-Hua. 1986. *Keerkeziyu Jianzhi* [*A Grammar Sketch of Kirghiz*]. Beijing: Publishing House of Minority Nationalities.

Huelsenbeck, John, and Fredrik Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17:754–755.

Junastu. 1981. *Tuzuyu Jianzhi* [*A Grammar Sketch of Tuzu*]. Beijing: Publishing House of Minority Nationalities.

Kilarski, Marcin, and Marc Allassonnière-Tang. 2021. Classifiers in morphology. *Oxford Research Encyclopedia of Linguistics*, ed. by Mark Aronoff, 1–28. Oxford: Oxford University Press.

Kornfilt, Jaklin. 1997. *Turkish Descriptive Grammar.* London: Routledge.

Lakoff, George, and Mark Johnson. 2003. *Metaphors We Live by.* London: University of Chicago Press.

Li, Shulan, and Qian Zhong. 1986. *Xiboyu Jianzhi* [*A Brief Description of Xibe*]. Beijing: Minzu Chubanshe.

Mace, Ruth, and Clare Holden. 2005. A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution* 20.3:116–121.

Metropolis, Nicholas, Arianna Rosenbluth, Marshall Rosenbluth, and Augusta Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21:1087–1091.

Popescu, Andrei, Katharina Huber, and Emmanuel Paradis. 2012. Ape 3.0: New tools for distance based phylogenetics and evolutionary analysis in R. *Bioinformatics* 28:1536–1537.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing* (Version 4.1.2) [Computer Software]. Vienna: R Foundation for Statistical Computing. Retrieved March 18, 2022, from https://www.R-project.org/

Rambaut, Andrew, Alexei Drummond, Dong Xie, Guy Baele, and Marc Suchard. 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67.5:901–904.

doi Revell, Liam. 2012. Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3.2:217–223.

Robbeets, Martine. 2015. *Diachrony and verb morphology: Japanese and the Transeurasian languages*. Berlin: Mouton De Gruyter.

doi Robbeets, Martine. 2017. The Turkic languages. *The Turkic Languages* 21.1:3–35.

Robbeets, Martine & Alexander Savelyev. 2020. *The Oxford guide to the Transeurasian languages*. Oxford: Oxford University Press.

doi Robbeets, Martine, and Remco Bouckaert. 2018. Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. *Journal of Language Evolution* 3.2:145–162.

Robbeets, M., Bouckaert, R., Conte, M. et al. 2021. Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 599, 616–621.

doi Ronquist, Fredrik, and John Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)* 19.12:1572–1574.

doi Schliep, Klaus. 2011. Phangorn: Phylogenetic analysis in R. *Bioinformatics* 27.4:592–593.

doi Seifart, Frank. 2010. Nominal classification. *Language and Linguistics Compass* 4.8:719–736.

Slater, Keith. 2003. *A Grammar of Mangghuer*. London: Routledge.

Soper, John. 1987. Loan Syntax in Turkic and Iranian: The Verb Systems of Tajik, Uzbek, and Qashqai. Doctoral dissertation, University of California, Los Angeles, CA.

Underhill, Robert. 1980. *Turkish Grammar*. Cambridge, MA: MIT Press.

van Schaaik, Gervan. 1996. *Studies in Turkish Grammar Turcologica*. Wiesbaden: Harrassowitz Verlag.

Wang, Qingfeng. 2005. *Manyu Yanjiu* [*A Study of Manchu*]. Beijing: Minzu Chubanshe.

Address for correspondence

One-Soon Her
Department of Foreign Languages and Literature
Tunghai University
Taichung, TAIWAN

hero@thu.edu.tw
Graduate Institute of Linguistics
National Chengchi University
Taipei, Taiwan

onesoon@gmail.com

Co-author information

Marc Allassonnière-Tang
The UMR 7206 "Ecological Anthropology"
Joint Lab
CNRS/MNHN/University Paris City
Paris, FRANCE

marc.allassonniere-tang@mnhn.fr

Zhong-Liang Gao
Graduate Institute of Linguistics
National Chengchi University
Taipei, TAIWAN

taibeigao@gmail.com

Shen-An Chen
Graduate Institute of Linguistics
National Chengchi University
Taipei, TAIWAN

q19940220@gmail.com

Publication history