

# Computational Typology

Gerhard Jäger

University of Tübingen, Seminar für Sprachwissenschaft, Keplerstr. 2, 72074 Tübingen, Germany

## ARTICLE HISTORY

Compiled April 28, 2025

## ABSTRACT

Typology is a subfield of linguistics that focuses on the study and classification of languages based on their structural features. Unlike genealogical classification, which examines the historical relationships between languages, typology seeks to understand the diversity of human languages by identifying common properties and patterns, known as universals. In recent years, computational methods have played an increasingly important role in typological research, enabling the analysis of large-scale linguistic data and the testing of hypotheses about language structure and evolution. This article provides an illustration of the benefits of computational statistical modeling in typology.

## KEYWORDS

typology, statistics, phylogenetics, language universals

## 1. Introduction

Typology is the subfield of linguistics studying and classifying languages according to their structural features (as opposed, e.g. according to their genealogical classification). Its aims are to delineate the diversity of human languages, and to identify common properties of all (or most) languages, so-called *universals*.

The aim of *statistical typology* is to identify robust correlations between typological features of languages or between linguistic and non-linguistic properties of populations and their languages, and to do so in a statistically sound way.

Researchers now have access to extensive datasets that enable large-scale quantitative analyses of linguistic features. These databases, such as the World Atlas of Language Structures (WALS; Dryer and Haspelmath 2013), Grambank (Skirgård et al., 2023), Glottolog (Hammarström, Forkel, Haspelmath, & Bank, 2020), and PHOIBLE (Moran & McCloy, 2019), provide detailed information on typologically relevant properties of languages. This facilitates the use of modern statistical software when conducting such analyses. In this chapter, these new developments will be reviewed and illustrated.

### 1.1. Historical Context

The origins of linguistic typology can be traced back to the early 19th century with the pioneering work of August Wilhelm von Schlegel. In 1818, Schlegel proposed a tripartite classification of languages based on morphological characteristics: fusional, agglutinative, and

isolating languages. (See Bynon, 2004 for details on the historical background of this classification.) Later, polysynthetic languages were added to this classification. Schlegel's work laid the groundwork for understanding how languages can be categorized based on their structural properties, such as how words are formed and how meaning is conveyed through grammar.

Schlegel's morphological types illustrated how languages could evolve from one type to another. For instance, "a fusional language can develop into one of the isolating type, an isolating language can become agglutinative, an agglutinative language may move towards a fusional profile, and so on" (Dixon, 1994, 182-183). Schlegel's insights highlighted the dynamic nature of language structure and set the stage for further exploration into the mechanisms of language evolution.

Georg von der Gabelentz (1840-1893) further advanced the field by emphasizing the interconnectedness of linguistic features within a language system. He argued that languages are organic systems where all parts are interdependent, and changes in one part can affect the whole. In fact, his program for linguistic typology sounds distinctly modern, as can be seen from this quotation:

"But how gainful would it be if we could straightforwardly say to a language: you have this characteristic, consequently, you have those further characteristics, and that general character! – if, like the bold botanists have tried to do, we could construct the lime tree from the lime leaf. If I were allowed to baptize an unborn child, I would choose the name *typology*." (von der Gabelentz, 1901, 481, translation quoted from Elffers et al., 2008, 194)

Von der Gabelentz's insights highlighted the need for a holistic approach to studying language typology, setting the stage for more systematic and statistically sound investigations. His vision of typology as a means of understanding the underlying principles of language structure was encapsulated in his metaphor of reconstructing the entire lime tree from a single leaf.

The modern era of linguistic typology began with Joseph Greenberg's seminal work in the mid-20th century. In his 1963 paper, "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements," Greenberg (1963) identified 45 universally or near-universally valid statements about language structure. His approach involved analyzing a diverse sample of 30 languages to uncover patterns that hold across different language families and geographical regions.

Greenberg's work introduced the concept of *linguistic universals*, which can be either unconditional or conditional. Unconditional universals are features that are present in all languages, such as the tendency for subjects to precede objects in declarative sentences. Conditional universals, on the other hand, describe patterns that occur with more than chance frequency when certain conditions are met, such as the tendency for languages with Verb-Subject-Object (VSO) order to place adjectives after nouns.

## 1.2. Language universals

Greenberg (1963) identified four types of linguistic universals based on their absolute or statistical nature and their conditional or unconditional status. These types are summarized in Table 1.2.

Linguistic universals can be categorized into two main types: *unconditional* and *conditional* universals. Below are examples of each type, taken from (Greenberg, 1963):

### Unconditional Universals

- **Universal 1:** In declarative sentences with nominal subject and object, the dominant

	Absolute (exceptionless)	Statistical (tendencies)
<b>Unconditional (unrestricted)</b>	Type 1. “Unrestricted absolute universals” <i>All languages have property X</i>	Type 2. “Unrestricted tendencies” <i>Most languages have property X</i>
<b>Conditional (restricted)</b>	Type 3. “Exceptionless implicational universals” <i>If a language has property X, it also has property Y</i>	Type 4. “Statistical implicational universals” <i>If a language has property X, it will tend to have property Y</i>

**Table 1.** Typology of linguistic universals (from Evans and Levinson, 2009), Table 1

order is almost always one in which the subject precedes the object.

- **Universal 14:** In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.
- **Universal 35:** There is no language in which the plural does not have some nonzero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero.
- ...

### Conditional Universals

- **Universal 2:** In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes it.
- **Universal 3:** Languages with dominant VSO order are always prepositional.
- **Universal 4:** With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.
- **Universal 13:** If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.
- ...

Furthermore, Greenberg distinguishes between *absolute* and *statistical universals*. An absolute universal applies to all languages without exception, while a statistical universal holds true for the majority of languages but may have exceptions. Here are some examples.

### Absolute Universals

- **Universal 14:** In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.
- **Universal 44:** If a language has gender distinctions in the first person, it always has gender distinctions in the second or third person, or in both.
- **Universal 45:** If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also.
- ...

### Statistical Universals

- **Universal 17:** With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.
- **Universal 18:** When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise.
- **Universal 41:** If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.
- ...

Since Greenberg's seminal work, a plethora of similar studies with larger language samples have been conducted.

### *1.3. Correlation between linguistic and non-linguistic traits*

Another major focus of research in statistical typology is the correlation between linguistic and non-linguistic traits. The idea is that languages are shaped by the social and ecological environment in which they are spoken. This is a very old idea, going back at least to the 19th century, but it has been given a new lease of life by the availability of large-scale databases that allow us to test these ideas in a more systematic way.

For instance, Lupyán and Dale (2010) make a strong case that languages spoken by larger populations tend to have simpler morphological structures and rely more on lexical strategies, while those spoken by smaller groups are more morphologically complex, suggesting that language structures adapt to the social environments in which they are used. Atkinson (2011) observes that the size of the phoneme inventory of a language is inversely correlated with the distance from Africa along the likely migration routes of early humans. Hay and Bauer (2007) observe a positive correlation between population size and sound inventory size, while Everett, Blasi, and Roberts (2015) argue for a connection between climate and tonality patterns in languages.

This list can be expanded considerably. Roberts and Winters (2013), however, urge caution when digging for correlations, lest we fall into the trap of “just-so stories” that are not backed up by scientifically plausible causal mechanisms. To bring this point home, they list correlations like those between the inflectional synthesis of verbs in a language with the habit of its speakers to hold siestas, or the presence of tone in a language with the presence of acacia trees.

## **2. Electronic Resources**

The availability of large-scale databases has revolutionized the field of typology by providing researchers with access to extensive linguistic data. These databases contain information on a wide range of typological features, such as word order, case marking, and phonological systems, for hundreds of languages. Some of the most widely used databases include:

- **World Atlas of Language Structures (WALS):** This database is a key resource for understanding the structural properties of languages, including word order, case marking, and phonological systems. It is based on extensive fieldwork by typologists and is freely accessible online through the WALS website <https://wals.info/>, (Dryer & Haspelmath, 2013).
- **Grambank:** This database focuses on the morphosyntactic properties of languages, such as word order, case marking, and agreement systems. Like WALS, it is built on fieldwork by typologists and is available online through the Grambank website <https://grambank.clld.org/>, (Skirgård et al., 2023).

- **Glottolog**: Specializing in the genealogical classification of languages, Glottolog provides detailed information on language families, subfamilies, and isolates. It is based on the work of historical linguists who have reconstructed language relationships and is accessible online at <https://glottolog.org/>, (Hammarström et al., 2020).
- **PHOIBLE**: This database offers comprehensive data on the phonological properties of languages, including segment inventories, syllable structures, and tone systems. It is the result of phonological analyses by experts in the field and is available online at <https://phoible.org/>, (Moran & McCloy, 2019).
- **APiCS**: Focusing on pidgin and creole languages, APiCS provides insights into their grammatical properties, such as word order, case marking, and tense-aspect systems. It is based on research by creolists and is freely available online at <https://apics-online.info/>, (Michaelis, Maurer, Haspelmath, & Huber, 2013).
- **AUTOTYP**: Similar to WALS, AUTOTYP provides information on the typological properties of languages, including word order, case marking, and phonological systems. It is also based on fieldwork by typologists and is accessible online at <https://autotyp.uzh.ch/>, (Bickel et al., 2018).
- **Lexibank**: This database contains information on the lexical properties of languages, such as word lists, cognate sets, and semantic domains. It is based on the work of historical linguists who have reconstructed language relationships and is available online at <https://lexibank.cld.org/>, (List et al., 2022). Lexibank currently (March 2025) comprises about 3.5 million lexical entries from ca. 7,500 languages. It incorporates the data from earlier lexical data collection efforts such as the Automatic Similarity Judgment Program (Wichmann, Holman, & Brown, 2022) and NorthEuraLex (Dellert et al., 2020).

These databases are crucial tools for linguistic research, offering a wealth of data that can be used to explore linguistic diversity and typology.

### 3. Statistical non-independence

#### 3.1. Two case studies

A central methodological challenge in statistical typology is that languages do not constitute independent samples. This issue, commonly referred to as *Galton's Problem* (see Naroll, 1961), arises because similarities between languages may result from shared ancestry or contact rather than independent development. When related languages exhibit the same feature, it may have been inherited from a common ancestor. Similarly, when geographically proximate languages share features, the cause might be areal diffusion through language contact. As a result, standard statistical tests that assume independent observations are not applicable without adjustments.

To illustrate this point, I will use two running examples – a putative language universal, and a seeming correlation between linguistic and non-linguistic traits:

- **Greenberg's Universal 27**: If a language is exclusively suffixing, it is postpositional; if it is exclusively prefixing, it is prepositional. Greenberg (1963)
- **Phoneme inventory size and population size**: Hay and Bauer (2007) observe a positive correlation between population size and sound inventory size.

**Greenberg's Universal 27** To assess the empirical support of the statement, I accessed the WALS database and extracted information on feature 26 (Prefixing vs. Suffixing in Inflectional

Morphology; Dryer, 2013b) and feature 85 (Order of Adposition and Noun Phrase, Dryer, 2013a). For 589 languages, values for both features were available. The results are summarized in Table 2.

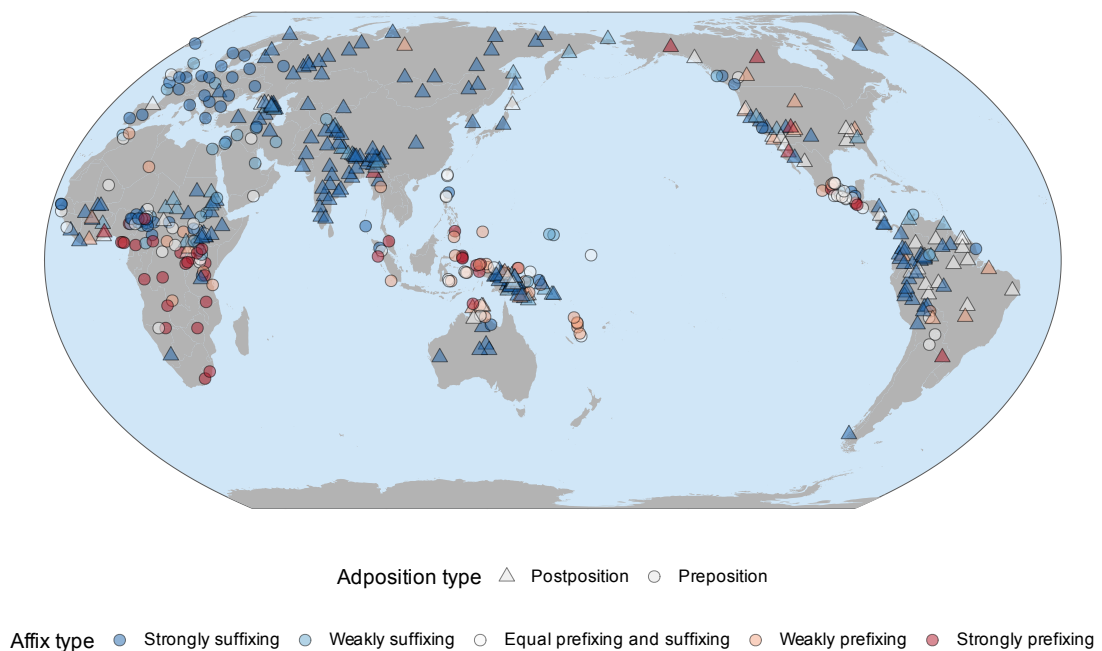
Affix Type	Postposition	Preposition
Strongly suffixing	214	63
Weakly suffixing	50	33
Equal prefixing and suffixing	52	61
Weakly prefixing	24	43
Strong prefixing	9	40

**Table 2.** Distribution of affix types across postpositions and prepositions, with preposition percentages.

These numbers, strictly speaking, do neither confirm nor disconfirm Greenberg’s statement, since exclusively prefixing or suffixing languages are not listed. However, the table shows something very similar in spirit: the more prefixing a language is, the more likely it is to be prepositional. The same holds for suffixing languages and postpositions.

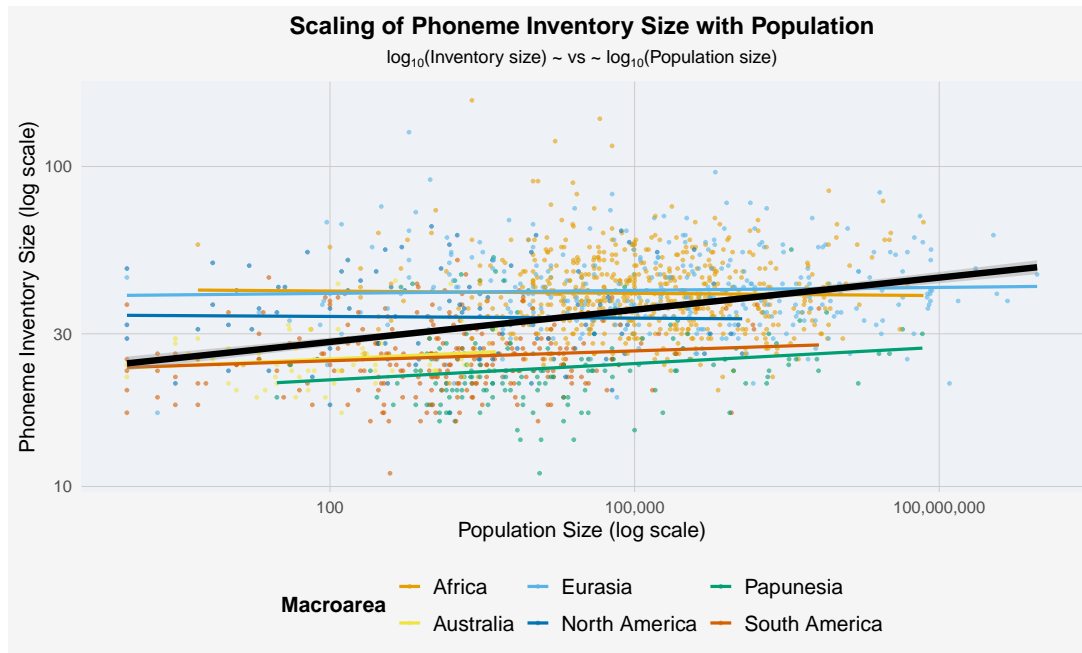
Figure 1 shows the geographic distribution of affix and adposition types. It is clearly visible

### Affixing type by adposition



**Figure 1.** Distribution of affix types across postpositions and prepositions.

that the distribution of affix types is not random, but rather shows a clear geographical pattern. This is a clear indication that the distribution of affix types is not independent, and that we need to take this into account when testing for correlations.



**Figure 2.** Scatterplot of population size and phoneme inventory size.

**Population size and phoneme inventory size** As mentioned above, Hay and Bauer (2007) observe a positive correlation between population size and sound inventory size. To explore this effect, sound inventory size data were extracted from the PHOIBLE database (Moran & McCloy, 2019), and population size data were obtained from Ethnologue (Lewis, 2009) via ASJP (Wichmann et al., 2022).

The scatterplot in Figure 2 shows the relationship between log-transformed population size and, also log-transformed phoneme inventory size. At a first glance, there seems to be a very strong positive correlation between the two variables, as indicated by the black trend line. However, Moran, McCloy, and Wright (2012) argue convincingly that this association is in fact a statistical artifact.

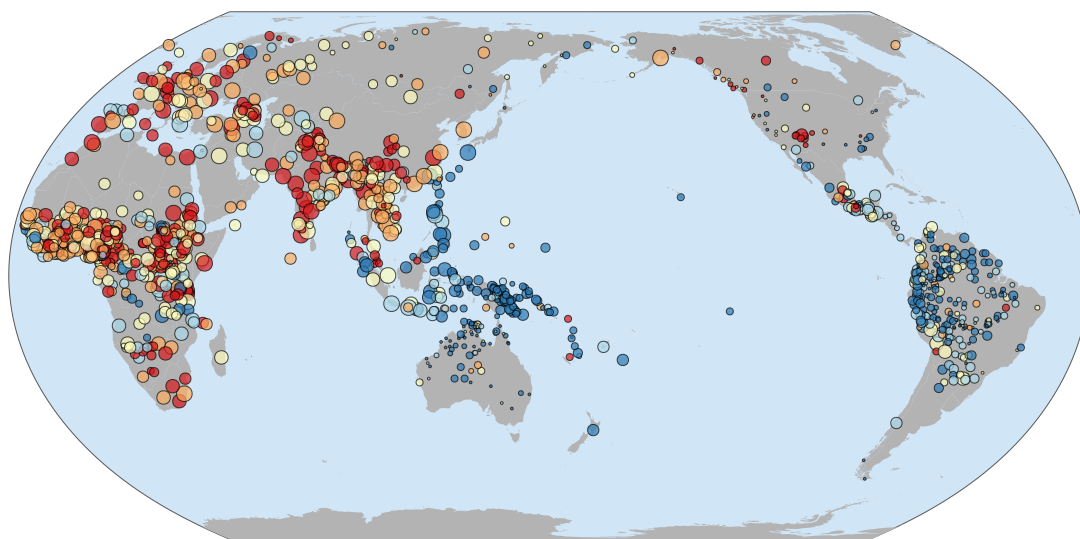
To get an intuition why this is so, have a look at the map in Figure 3. It shows the population size and phoneme inventory size of the languages in the sample. The map suggests that both languages with a small sound inventory and languages with a small number of speakers are strongly concentrated in Australia/Oceania and in the Americas.

It stands to reason that Hay and Bauer’s association predominantly obtains between continents rather than within continents. This impression is reinforced by the slope of the per-macroarea trend lines shown in Figure 2. It can be seen that there are only weak trends within macroareas, and they can be positive or negative.

This observation does not invalidate Hay and Bauer’s claim, but it raises the question whether the correlation is statistically significant once non-independence of languages due to geographic proximity is taken into account.

In the remainder of this section, I will compare analyses of these running examples first under the assumption of independence, and then with two more sophisticated methods taking the genealogical interdependence between languages into account. It will be demonstrated that the results are very different, and that the assumption of independence is not tenable. Similar statistical controls are possible for spatial non-independence due to contact. For reasons of space, I omit a discussion of this issue here, the interested reader is referred to (Guzmán Naranjo &

## Phoneme inventory (color) and population size (size)



Segment number range   ● 11-25   ● 26-30   ● 31-37   ● 38-45   ● 46-161

Population range  
● <100   ● 1k-10k   ● 100k-1M   ● >10M  
● 100-1k   ● 10k-100k   ● 1M-10M

**Figure 3.** Map of population size and phoneme inventory size.

Becker, 2021) for more information.

### 3.2. Statistical evaluation under the assumption of independence

As starting point, I will analyze the data in the two running examples under the assumption that the languages in the sample are independent. This is a very strong assumption, and it is very likely that it is not true. However, it is a useful starting point to get an idea of the strength of the correlation.

All analyses were performed using the R programming language (R Core Team, 2021) in combination with the statistics software *Stan* (Stan Development Team, 2022) and the R-package *rstan* Stan Development Team (2024).

**Affix and adposition types** Dependencies between two variables are often modeled via linear or logistic regression, where you have an independent and a dependent variable. The statistical model describes the influence that the independent variable has on the dependent one. This means that only the dependent variable is actually modeled; the independent variable is taken as given. While this is appropriate for experimental studies, it is not ideal for observational data such as those we are dealing with here. Rather, we want to model the distributions of both variables, and the dependency between them. This can be achieved via a *bivariate* model.

It is important to note that both variables are discrete. The affix variable is **ordinal**. Its levels are:

- (1) Strongly suffixing
- (2) Weakly suffixing
- (3) Equal prefixing and suffixing
- (4) Weakly prefixing
- (5) Strongly prefixing

The values are ordered, but it is not assumed that the distance between the levels is equal. The adposition variable is **binary**, and its levels are:

- (1) Postposition
- (2) Preposition

For both variables, the model assumes a latent continuous variable. The latent affix variable  $z_{\text{affix}}$  is linked to the affix level via an *ordered logistic* link function. This means that the model additionally assumes four cutpoints, splitting the real line into five intervals. The observed affix level is assumed to correspond to the interval in which the latent variable falls. The model takes the cutpoints to be ordered, and the distance between them not necessarily being equal. Rather, the positions of the cutoff points are estimated from the data. The cutpoints are denoted by  $c_1, c_2, c_3, c_4$ .

The latent adposition variable  $z_{\text{adposition}}$  is linked to the observed level – postposition or preposition – via a *logistic* link function. This means that the model assumes that the probability of observing a preposition equals the logistic transformation of  $z_{\text{adposition}}$ .

Taken together, the model can be summarized as shown in Model 1 ( $z_i$  consists of the two latent variables  $z_{\text{affix}}$  and  $z_{\text{adposition}}$  for language  $i$ ). The two variables  $z_{\text{affix}}$  and  $z_{\text{adposition}}$  follow a bivariate normal distribution with mean 0, standard deviations  $\sigma_1$  and  $\sigma_2$ , and a correlation coefficient  $\rho$ , which are estimated from the data.

The model was fitted using the *rstan* package in R. The results are shown in Table 3. The model converged successfully, as indicated by the Rhat values close to 1 and the large effective sample size (n\_eff). The crucial outcome is the estimated value of  $\rho$  of 0.84. The 95% cred-

$$\rho \sim \text{Uniform}(-1, 1) \quad (1)$$

$$\sigma_1, \sigma_2 \sim \text{LogNormal}(0, 1) \quad (2)$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (3)$$

$$z_i \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad (4)$$

$$c_1 \sim \mathcal{N}(0, 2) \quad (5)$$

$$\zeta_k \sim \text{LogNormal}(0, 2), \quad c_{k+1} = c_k + \zeta_k \quad (6)$$

$$P(\text{affix} = 1) = P(z_{\text{affix}} < c_1) \quad (7)$$

$$P(\text{affix} = i) = P(c_{i-1} < z_{\text{affix}} < c_i) \quad \text{for } i = 2, 3, 4 \quad (8)$$

$$P(\text{affix} = 5) = P(z_{\text{affix}} > c_4) \quad (9)$$

$$P(\text{preposition}) = \frac{1}{1 + e^{-z_{\text{adposition}}}} \quad (10)$$

Model 1: Bivariate normal model for latent variables in the affix–adposition association.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
cutpoints <sub>1</sub>	-0.36	0.00	0.15	-0.66	-0.08	11381	1.00
cutpoints <sub>2</sub>	0.54	0.00	0.16	0.25	0.86	11638	1.00
cutpoints <sub>3</sub>	2.03	0.00	0.27	1.55	2.62	3683	1.00
cutpoints <sub>4</sub>	3.50	0.01	0.42	2.78	4.39	3229	1.00
$\sigma_1$	2.03	0.01	0.38	1.34	2.81	2627	1.00
$\sigma_2$	2.11	0.01	0.41	1.35	2.89	4721	1.00
$\rho$	0.84	0.00	0.08	<b>0.68</b>	<b>0.97</b>	3625	1.00

**Table 3.** Posterior summary statistics for the bivariate model associating affix type and adposition type.

ible interval (shown in bold) does not include zero, indicating that there is credible positive correlation between affix type and adposition type.

**Population size and phoneme inventory size** As the scatterplot in Figure 2 suggests, both population sizes and sound inventory sizes are log-normally distributed; i.e., the logarithms of these quantities are approximately normally distributed. The plot also suggests that a bivariate linear model using log-transformed population size and log-transformed phoneme inventory size as associated variables is appropriate. Following standard practice (McElreath, 2016), both variables were standardized to have mean 0 and standard deviation 1. Since both variables are continuous, no latent variables and link functions are needed.

Formally, the model is specified in Model 2 ( $x_i$  consists of the values for standardized values for log-transformed population size and log-transformed phoneme inventory size for language  $i$ ).

This model was fitted using the *rstan* package in R, and the results are summarized in Table 4. The model converged successfully (as indicated Rhat and n\_eff). The crucial outcome

$$\rho \sim \text{Uniform}(-1, 1) \quad (11)$$

$$\sigma_1, \sigma_2 \sim \text{LogNormal}(0, 1) \quad (12)$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (13)$$

$$\mu \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}) \quad (14)$$

$$x_i \sim \mathcal{N}(\mu, \Sigma) \quad (15)$$

$$(16)$$

Model 2: Bivariate normal model for segment inventory size/population size association.

is the fact that the 95% credible interval for the correlation coefficient  $\rho$  does not include zero, indicating that there is a significant positive correlation between population size and phoneme inventory size.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
$\mu_1$	0.00	0.00	0.026	-0.050	0.050	47271	1.00
$\mu_2$	0.00	0.00	0.026	-0.050	0.051	44708	1.00
$\sigma_1$	1.00	0.00	0.018	0.966	1.037	48340	1.00
$\sigma_2$	1.00	0.00	0.018	0.965	1.037	48035	1.00
$\rho$	0.35	0.00	0.023	<b>0.304</b>	<b>0.393</b>	47673	1.00

**Table 4.** Posterior summaries for the Bayesian linear bivariate model modelling log-transformed phoneme inventory size and log-transformed population size.

### 3.3. Taking phylogenetic non-independence into account I: Hierarchical models

The models presented above assume that the languages in the sample are independent. However, this is a very strong assumption, and it is very likely that it is not true. In this section, I will show how to take this non-independence into account.

A fairly straight-forward way to do this is to use a hierarchical model, where the languages are grouped into families. (This approach has been pioneered by Atkinson, 2011; see also the discussion in Jaeger, Graff, Croft, and Pontillo, 2011.)

Each language belongs to a language family. For this study, I assume the classification from Glottolog (Hammarström et al., 2020). Isolate languages are treated as their own family.

Starting with the running example of the affix-adposition association, the basic setup is as above. For both features, we assume a continuous latent variable, which is linked to the observed feature value via a link function.

In the hierarchical model, these latent variables are composed of two components: a family-level component and a language-level component. For both components, I assume a bivariate normal distribution, with correlation coefficients  $\rho_f$  and  $\rho_l$  respectively. The family-level component is assumed to be the same for all languages in a family, while the language-level component is assumed to be independent for each language.

The formal specification of the model given in Model 3.

$$\rho_f, \rho_l \sim \text{Uniform}(-1, 1) \quad (17)$$

$$\sigma_{f,1}, \sigma_{f,2}, \sigma_{l,1}, \sigma_{l,2} \sim \text{LogNormal}(0, 1) \quad (18)$$

$$\Sigma_f = \begin{pmatrix} \sigma_{f,1}^2 & \rho_f \sigma_{f,1} \sigma_{f,2} \\ \rho_f \sigma_{f,1} \sigma_{f,2} & \sigma_{f,2}^2 \end{pmatrix} \quad (19)$$

$$\Sigma_l = \begin{pmatrix} \sigma_{l,1}^2 & \rho_l \sigma_{l,1} \sigma_{l,2} \\ \rho_l \sigma_{l,1} \sigma_{l,2} & \sigma_{l,2}^2 \end{pmatrix} \quad (20)$$

$$z_{f,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_f) \quad (21)$$

$$z_{l,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_l) \quad (22)$$

$$z_i := z_{f,i} + z_{l,i} \quad (23)$$

$$c_1 \sim \mathcal{N}(0, 2) \quad (24)$$

$$\zeta_k \sim \text{LogNormal}(0, 2), \quad c_{k+1} = c_k + \zeta_k \quad (25)$$

$$P(\text{affix} = 1) = P(z_{\text{affix}} < c_1) \quad (26)$$

$$P(\text{affix} = i) = P(c_{i-1} < z_{\text{affix}} < c_i) \quad \text{for } i = 2, 3, 4 \quad (27)$$

$$P(\text{affix} = 5) = P(z_{\text{affix}} > c_4) \quad (28)$$

$$P(\text{preposition}) = \frac{1}{1 + e^{-z_{\text{adposition}}}} \quad (29)$$

Model 3: Hierarchical bivariate model with family-level random effects for affix–adposition association.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
cutpoints <sub>1</sub>	-1.186	0.014	0.469	-2.275	-0.399	1147	1.00
cutpoints <sub>2</sub>	0.187	0.007	0.409	-0.549	1.091	2983	1.00
cutpoints <sub>3</sub>	2.390	0.033	0.775	1.394	4.448	560	1.01
cutpoints <sub>4</sub>	4.382	0.058	1.241	2.943	7.811	464	1.01
$\sigma_{f,1}$	3.162	0.038	0.886	2.029	5.589	541	1.00
$\sigma_{f,2}$	6.927	0.058	2.782	3.626	13.986	2283	1.00
$\sigma_{l,1}$	1.723	0.042	0.864	0.635	4.030	421	1.01
$\sigma_{l,2}$	2.086	0.028	1.166	0.748	5.060	1679	1.00
$\rho_l$	0.811	0.003	0.144	<b>0.477</b>	<b>0.993</b>	2191	1.00
$\rho_f$	0.636	0.002	0.110	<b>0.390</b>	<b>0.816</b>	4047	1.00

**Table 5.** Posterior summaries for the ordinal–binary bivariate model with family-level and language-level covariance. Cutpoints are used to map a latent continuous variable to ordinal affix positions; correlations are estimated separately at the family ( $f$ ) and language ( $l$ ) level.

This model was fitted using *rstan*. The results are summarized in Table 5. The crucial outcome is the fact that the 95% credible interval for the correlation coefficients  $\rho_f$  and  $\rho_l$  are clearly positive and do not include zero. This indicates that there is a significant positive correlation between affix type and adposition type, even when taking family-level structure into account.

The hierarchical model provides a massively better fit of the data, with a log-Bayes factor of approximately 280 in favor of the hierarchical model.

Likewise, I fitted a hierarchical model to the population size and phoneme inventory size data. The model specification is given in Model 4.

$$\rho_f, \rho_l \sim \text{Uniform}(-1, 1) \quad (30)$$

$$\sigma_{f,1}, \sigma_{f,2}, \sigma_{l,1}, \sigma_{l,2} \sim \text{LogNormal}(0, 1) \quad (31)$$

$$\Sigma_f = \begin{pmatrix} \sigma_{f,1}^2 & \rho_f \sigma_{f,1} \sigma_{f,2} \\ \rho_f \sigma_{f,1} \sigma_{f,2} & \sigma_{f,2}^2 \end{pmatrix} \quad (32)$$

$$\Sigma_l = \begin{pmatrix} \sigma_{l,1}^2 & \rho_l \sigma_{l,1} \sigma_{l,2} \\ \rho_l \sigma_{l,1} \sigma_{l,2} & \sigma_{l,2}^2 \end{pmatrix} \quad (33)$$

$$\mu \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}) \quad (34)$$

$$z_{f,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_f) \quad (35)$$

$$z_{l,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_l) \quad (36)$$

$$x_i := \mu + z_{f,i} + z_{l,i} \quad (37)$$

Model 4: Bivariate model with family-level random intercepts for segment inventory/population size association.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
$\mu_1$	-0.578	0.003	0.065	-0.706	-0.450	579	1.02
$\mu_2$	-0.352	0.003	0.070	-0.484	-0.209	550	1.02
$\sigma_{l,1}$	0.696	0.000	0.013	0.671	0.722	7346	1.00
$\sigma_{l,2}$	0.682	0.000	0.013	0.656	0.708	6597	1.00
$\sigma_{f,1}$	0.753	0.001	0.053	0.654	0.864	2727	1.00
$\sigma_{f,2}$	0.880	0.001	0.061	0.765	1.009	2942	1.00
$\rho_l$	-0.010	0.000	0.028	<b>-0.064</b>	<b>0.045</b>	6484	1.00
$\rho_f$	0.543	0.002	0.076	<b>0.383</b>	<b>0.681</b>	2041	1.00

**Table 6.** Posterior summaries for the extended bivariate model with both residual- and family-level covariance components.

The fitted model is summarized in Table 6.

There are several noteworthy points to be made. Unlike in the non-hierarchical model, the estimated posterior means ( $\mu_1$  and  $\mu_2$ ) are much smaller than 0. Converting back from the normalized log-scales to the original scales, they are approximately at 3,200 population size and 30 segments. This reflects the fact that languages from large families have lower weight and isolates a higher weight in the hierarchical model.

Even more noteworthy, we find that the language-level correlation coefficient  $\rho_l$  is essentially zero, with a 95% credible interval of  $(-0.064, 0.045)$ , while we find a pronounced positive correlation at the family level, with a posterior mean of 0.543 and a 95% credible interval of  $(0.383, 0.681)$ . This indicates that the correlation between population size and phoneme inventory size is driven by the family-level structure of the data.

A discussion of the implications of this finding are deferred to the next subsection.

### 3.4. Taking genetic non-independence into account II: Phylogenetic models

Family-level random intercepts capture only a part of the genetic non-independence of languages. They have two main shortcomings. First, they ignore the internal structure of language families. For instance, the dependency between Spanish and Portuguese is modeled as equally strong as the dependency between Spanish and Greek, since both pairs are in the same family. This can be mitigated to a certain degree by adding sub-family as random effects, but this still does not capture the intricate structure of language families.

Second, family-level random intercepts assume the degree of dependency within families is the same for all families. This ignores the difference between shallow families such as Mongolic (estimated age of divergence 700-800 years; Janhunen, 2024) and deep families such as Afro-Asiatic (estimated age of divergence earlier than 10,000 years; Ehret, 1979).

To address these limitations, researchers increasingly turn to **phylogenetic comparative methods** (PCMs; see Harmon 2019 for a book-length overview), which explicitly model the evolutionary relationships among languages based on a phylogenetic tree. These methods treat languages (or species, or whatever empirical phenomena are being modeled) not as independent datapoints, but as part of an evolutionary process shaped by descent from common ancestors. The phylogenetic tree encodes these relationships, including both the topology (who is related to whom) and branch lengths (how much change has happened since divergence).

Constructing family trees of languages is one of the core tasks of classical historical linguistics. However, the trees produced by the comparative method are only of limited use for PCMs, since (a) they do not include branch lengths, and (b) they are often underspecified since they rely on the very strict criterion that a clade can only be assumed if a shared innovation can be demonstrated.

Within the past twenty-five years, the new field of *computational historical linguistics* has emerged which uses methods adapted from computational biology to produce binary-branching phylogenetic trees of languages. Starting point for these methods are parallel word lists, such as the Swadesh list or similar collections of basic vocabulary items. From these lexical data, discrete (usually binary) features are extracted that classify languages in various ways. These features are ideally historically inert, i.e., their values are mostly inherited faithfully from an ancestor language to its descendants. If a mutation occurs, it is passed on to the descendants. If a sufficient number of such features is available, the tree can be automatically reconstructed.

The most widely used type of such features are derived from **cognate classes**, which are sets of words in different languages that are derived from a common ancestor. For instance, the English word *mother* and the Hindi word माता (/ma:ta:/) are cognates, as they both derive from the Proto-Indo-European word *\*méh<sub>2</sub>tēr*. All languages sharing a word for the concept of *mother* that is derived from the same Proto-Indo-European word have the value 1 for this feature, while all languages that do not share this cognate have the value 0.

This approach, while widely used, has two limitations. First, it is based on *manual cognacy annotation*, which is a time-consuming and error-prone process. Second, cognate classes are, by definition, confined to a single language family. According to the classical comparative method, a language family is a maximal group of languages for which a common ancestor can be demonstrated, and a cognate class is a set of words for which a common ancestor can be reconstructed. Therefore it is not possible to have cognate classes that span multiple language families. This is a serious limitation, as it means that PCMs cannot be used for comparative studies spanning multiple language families.

There are various proposals in the literature to overcome these limitations. E.g, features can be extracted from the phonetic transcriptions of basic vocabulary word lists directly via machine learning, thereby sidestepping both above-mentioned problems.

In this study, I will use the method proposed in Jäger (2018). This method uses automatic cognate clustering along the lines of Jäger, List, and Sofroniev (2017), and combines those with simple features pertaining to the presence or absence of sound classes in the reflexes of basic concepts. Jäger (2018) uses the data from (Wichmann, Holman, & Brown, 2018) to produce a phylogenetic tree of the world's languages. It is demonstrated via various quality measures that the resulting tree is in good accordance with the tree produced by the classical comparative method.

The workflow from Jäger (2018) was replicated with the data from (Wichmann, Holman, & Brown, 2020) in (Jäger, 2025). The results are available at <https://osf.io/a97sz/>.

For the study concerning affix type and adposition type, I started with the maximum-likelihood world tree, which was rooted with the method described in (Tria, Landan, & Dagan, 2017) and pruned it to the 589 languages for which WALS data were available. This tree was converted to ultrametric form using the *R*-function *chronos* from the *ape* package (Paradis, Claude, & Strimmer, 2004).

The tree, as well as the sub-tree for the Indo-European language family, is shown in Figure 4 for illustration. It illustrates both strengths and weaknesses of the method. On the one hand, it is clearly visible that the tree captures the main genealogical patterns adequately, largely identifying language families and their internal sub-groupings. On the other hand, many details are at odds with received scholarship. For instance, both the Atlantic-Congo languages and the Afro-Asiatic languages are interspersed with other languages in the world tree, and the Nuclear Trans New Guinea family is not recognized at all. In the Indo-European sub-tree, e.g., the placement of English and of Greek are arguably historically incorrect.

Also, it should be kept in mind that the branch lengths in the tree are not calibrated with information from the historical and archaeological record. They represent the amount of lexical change between divergence events. This is correlated with, but not identical to historical time.

In the context of the present study, the purpose of the phylogenetic tree is not to be a faithful representation of the history of language diversification. Rather, it is part of a statistical model, and it serves as a representation of the expected statistical dependencies between languages. In other words, the tree represents the similarity patterns of core vocabulary items, and it is used as a proxy for the expected similarity patterns of the features under investigation.

There are actually two questions that can be addressed when using phylogenetic control for an analysis of the association between two typological variables:

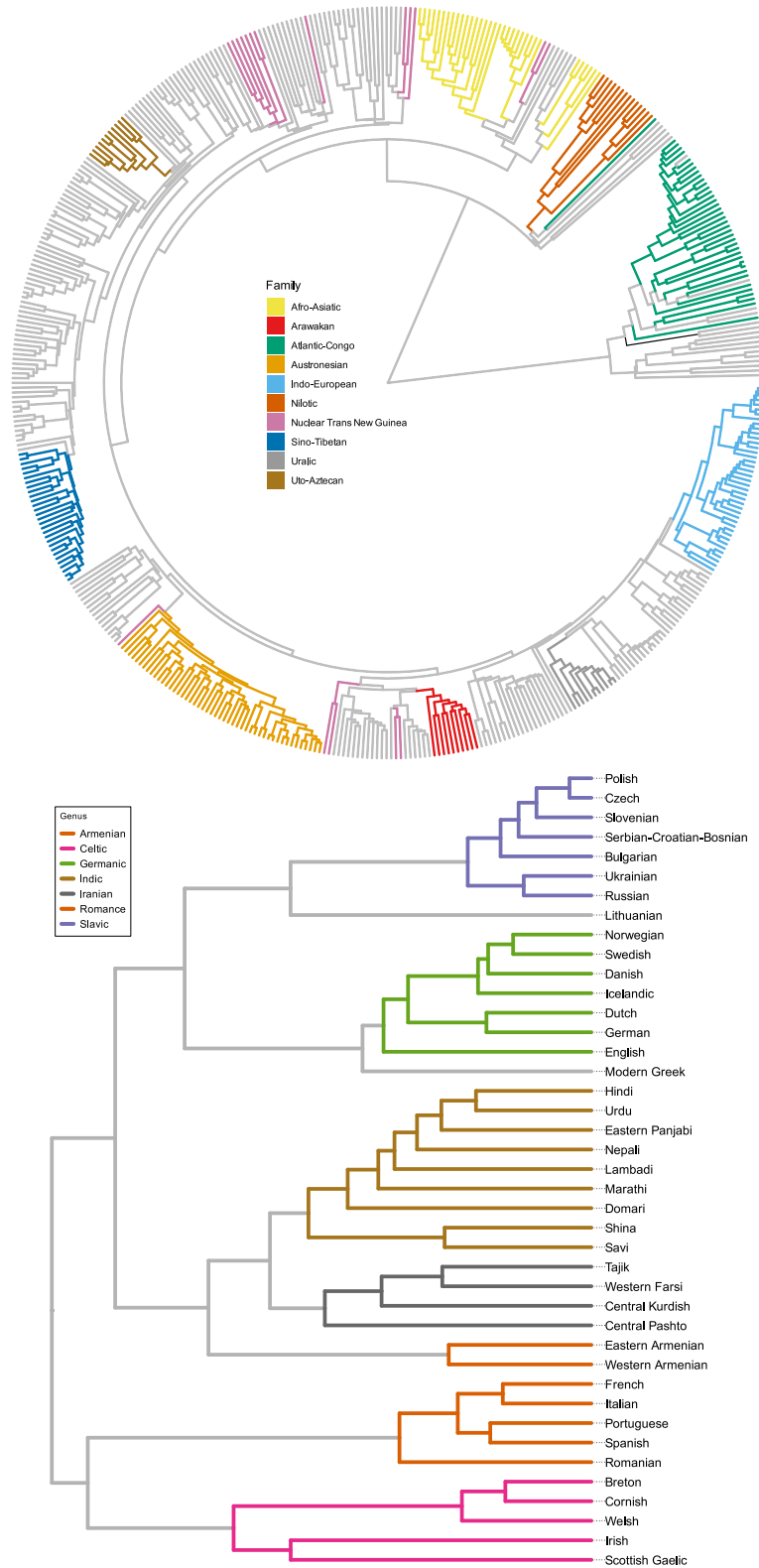
- When assuming that the response variable evolves on the tree, does knowledge of the predictor variable help to predict the response variable?
- When assuming that both variables evolve on the tree, is there a correlation between the diachronic changes of the two variables?

The first question is addressed by *phylogenetic regression* models, which assume that the response variable evolves on the tree, and that the predictor variable is a fixed effect. The second question is addressed by *phylogenetic correlation* models, which assume that both variables evolve on the tree.

As argued above, for observational data such as the ones we are dealing with here, a phylogenetic regression model is not appropriate. Rather, we want to model the distributions of both variables, and the dependency between them. This can be achieved via a bivariate phylogenetic correlation model.

In this study, I will generally assume that continuous variables evolve along the branches of a tree following the *Ornstein-Uhlenbeck process* (Hansen, 1997; Uhlenbeck & Ornstein, 1930).

The Ornstein-Uhlenbeck process (OU process henceforth) is a stochastic process that de-



**Figure 4.** **Top:** Phylogenetic tree of the world's languages. The tree was pruned to the 589 languages for which WALS data were available. **Bottom:** Subtree for the Indo-European languages.

scribes the evolution of a continuous trait over time. The OU process can be described as random walk on a leash. The trait evolves according to a random walk, but it is also subject to a restoring force that pulls it back towards a stable equilibrium value. This means that the trait tends to fluctuate around a mean value, rather than drifting away from the origin indefinitely.

It is characterized by three parameters: the long-term average  $\mu$ , the *drift* parameter  $\lambda$ , which describes the tendency of the trait to return to its mean value, and the *diffusion* parameter  $\sigma$ , which describes the amount of random variation in the trait. When it has value  $x_0$  at time  $t = 0$ , the probability density function of the trait at time  $t$  is given by:

$$x_t \sim \mathcal{N}\left(\mu + (x_0 - \mu)e^{-\lambda t}, \frac{\sigma^2}{2\lambda}(1 - e^{-2\lambda t})\right) \quad (38)$$

The long-term equilibrium distribution when  $t \rightarrow \infty$  is given by:

$$x_t \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\lambda}\right) \quad (39)$$

The model assumes that the value of the trait (latent variables for the affix-adposition study and log-transformed population size and phoneme inventory size for the population size study) is drawn at random from the equilibrium distribution (39). The value of the trait at a daughter node follows the distribution in (38) with the value at the parent node as  $x_0$  and the length of the branch as  $t$ . The parameters  $\mu$ ,  $\lambda$ , and  $\sigma$  are estimated from the data.

For a given phylogenetic tree, the predicted covariance of a trait value evolving according to OU between two languages  $i$  and  $j$  is given by:

$$\text{Cov}(x_i, x_j) = \frac{\sigma^2}{2\lambda} e^{-\lambda t_{ij}}, \quad (40)$$

where  $t_{ij}$  is the length of the path between languages  $i$  and  $j$  in the tree. The resulting variance-covariance matrix denoted by  $\Sigma_{OU}$ .

Since the phylogenetic tree used here is possibly unreliable above the family level, I combine this phylogenetic component with family-level intercepts. The model specification is given in Model 5.

The model was fitted using the *rstan* package in R. The results are summarized in Table 7. As in the hierarchical model, we find clear evidence for a positive correlation between affix type and adposition type, both at the family level (mean  $\rho_f = 0.554$ , 95% credible interval (0.215, 0.811)) and at the language level (mean  $\rho_l = 0.544$ , 95% credible interval (0.198, 0.896)).

Again, this model constitutes a massive improvement over the previous models, with a log-Bayes factor of approximately 925 in favor of the phylogenetic model.

It is important to appreciate that the language-level correlation holds between the diachronic changes of the two variables. Informally put, if a language evolves towards a more prefixing type for inflectional morphology, it tends to become more likely to be prepositional, and vice versa.

It is an advantage of an explicit phylogenetic model that ancestral state reconstruction is automatically conducted as a side effect. These reconstructions can be extracted from the fitted model. Figure 5 shows the reconstructed ancestral states for both latent variables for the Indo-European sub-tree for illustration.

$$\rho_f, \rho_l \sim \text{Uniform}(-1, 1) \quad (41)$$

$$\sigma_{f,1}, \sigma_{f,2}, \sigma_{l,1}, \sigma_{l,2}, \lambda_1, \lambda_2 \sim \text{LogNormal}(0, 1) \quad (42)$$

$$\Sigma_f = \begin{pmatrix} \sigma_{f,1}^2 & \rho_f \sigma_{f,1} \sigma_{f,2} \\ \rho_f \sigma_{f,1} \sigma_{f,2} & \sigma_{f,2}^2 \end{pmatrix} \quad (43)$$

$$\Sigma_l = \begin{pmatrix} \sigma_{l,1}^2 & \rho_l \sigma_{l,1} \sigma_{l,2} \\ \rho_l \sigma_{l,1} \sigma_{l,2} & \sigma_{l,2}^2 \end{pmatrix} \quad (44)$$

$$z_{f,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_f) \quad (45)$$

$$z_l \sim \mathcal{N}(\mu, \Sigma_{OU} \otimes \Sigma_l) \quad (46)$$

$$z_i := z_{f,i} + z_{l,i} \quad (47)$$

$$c_1 \sim \mathcal{N}(0, 2) \quad (48)$$

$$\zeta_k \sim \text{LogNormal}(0, 2), \quad c_{k+1} = c_k + \zeta_k + 10^{-2} \quad (49)$$

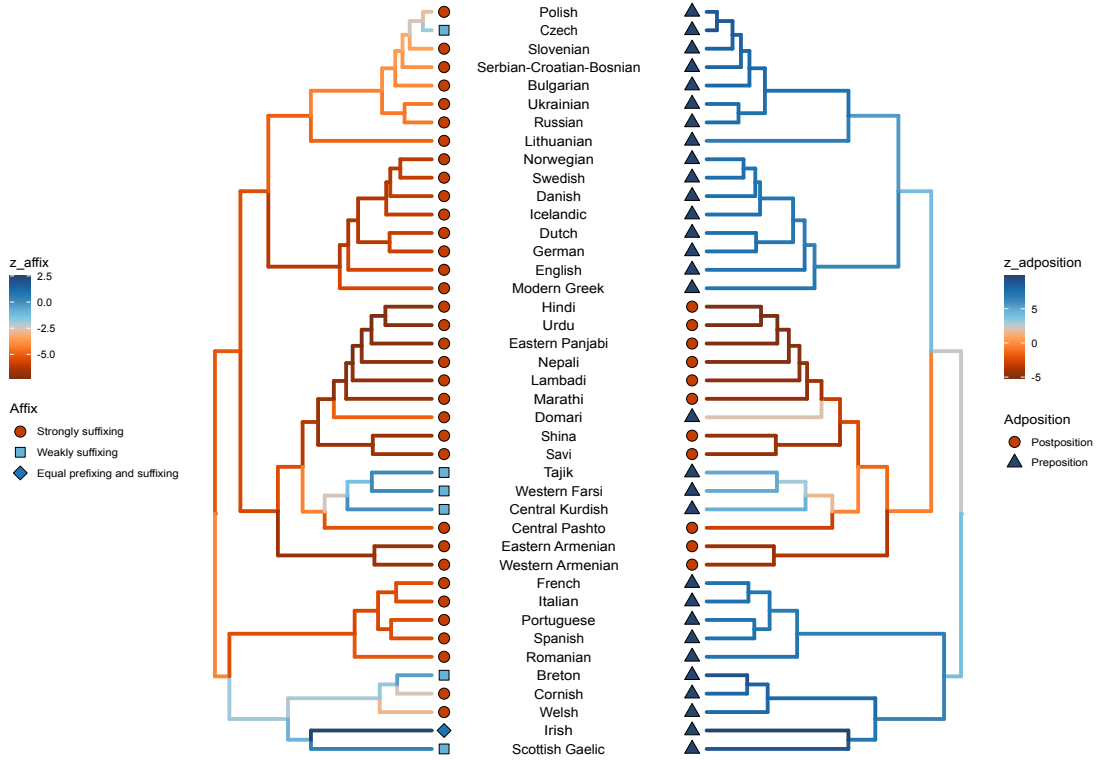
$$P(\text{affix} = 1) = P(z_{\text{affix}} < c_1) \quad (50)$$

$$P(\text{affix} = i) = P(c_{i-1} < z_{\text{affix}} < c_i) \quad \text{for } i = 2, 3, 4 \quad (51)$$

$$P(\text{affix} = 5) = P(z_{\text{affix}} > c_4) \quad (52)$$

$$P(\text{preposition}) = \frac{1}{1 + e^{-z_{\text{adposition}}}} \quad (53)$$

Model 5: Phylogenetic correlation model for affix–adposition association.



**Figure 5.** Ancestral state reconstructions for the Indo-European sub-tree. The left panel shows the reconstructed values for the affix latent variable, while the right panel shows the reconstructed values for the population size and adposition latent variable. Markers at the leaves indicate observed features.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
cutpoints <sub>1</sub>	-0.511	0.009	1.615	-3.656	2.653	33649	1.00
cutpoints <sub>2</sub>	2.014	0.020	1.727	-1.294	5.500	7129	1.00
cutpoints <sub>3</sub>	6.057	0.055	2.283	2.044	11.116	1692	1.00
cutpoints <sub>4</sub>	9.698	0.086	2.982	4.880	16.654	1207	1.00
$\mu_1$	0.505	0.009	1.616	-2.677	3.667	36011	1.00
$\mu_2$	-0.755	0.009	1.926	-4.494	3.065	46803	1.00
$\sigma_1$	2.618	0.030	0.886	1.413	4.865	890	1.01
$\sigma_2$	3.950	0.039	1.898	1.718	8.879	2393	1.00
$\lambda_1$	0.115	0.000	0.039	0.048	0.201	7201	1.00
$\lambda_2$	0.051	0.000	0.021	0.018	0.101	25272	1.00
$\rho_l$	0.544	0.003	0.188	<b>0.198</b>	<b>0.896</b>	3375	1.00
$\rho_f$	0.554	0.002	0.153	<b>0.215</b>	<b>0.811</b>	6404	1.00

**Table 7.** Posterior summaries for the bivariate Ornstein–Uhlenbeck model with family-level correlation  $\rho_f$ , residual correlation  $\rho_l$ , and fitted drift rates  $\lambda_1, \lambda_2$ . Cutpoints map the latent variable for the ordinal response to observed categories.

The tendency for correlated evolution can be observed, for instance, within the Indo-Iranian and the Celtic subfamilies.

The fact that the inter-family correlation is in the same range as the within-family correlation is consistent with the uniformitarian hypothesis whereas the same mechanisms driving evolution within families are also responsible for the distribution across families. Therefore we can conclude with high confidence that a preference for prepositions is universally associated with a preference for suffixes, and vice versa.

Let us now turn to the other case study. The model setup is similar, except that the co-evolving continuous variables are directly observed here. So the model specification is:

$$\rho_f, \rho_l \sim \text{Uniform}(-1, 1) \quad (54)$$

$$\sigma_{f,1}, \sigma_{f,2}, \sigma_{l,1}, \sigma_{l,2}, \lambda_1, \lambda_2 \sim \text{LogNormal}(0, 1) \quad (55)$$

$$\Sigma_f = \begin{pmatrix} \sigma_{f,1}^2 & \rho_f \sigma_{f,1} \sigma_{f,2} \\ \rho_f \sigma_{f,1} \sigma_{f,2} & \sigma_{f,2}^2 \end{pmatrix} \quad (56)$$

$$\Sigma_l = \begin{pmatrix} \sigma_{l,1}^2 & \rho_l \sigma_{l,1} \sigma_{l,2} \\ \rho_l \sigma_{l,1} \sigma_{l,2} & \sigma_{l,2}^2 \end{pmatrix} \quad (57)$$

$$\mu \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}) \quad (58)$$

$$z_{f,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_f) \quad (59)$$

$$z_l \sim \mathcal{N}(\mu, \Sigma_{OU} \otimes \Sigma_l) \quad (60)$$

$$x_i := \mu + z_{f,i} + z_{l,i} \quad (61)$$

Model 6: Phylogenetic correlation model for segment inventory/population size association.

The model was fitted using the *rstan* package in R. The results are summarized in Table 8. As in the hierarchical model, we find clear evidence for a positive correlation between population size and phoneme inventory size at the family level (mean 0.57, credible interval (0.39, 0.72)) but essentially zero correlation at the language level (mean  $-0.01$ , credible

interval  $(-0.07, 0.04)$ ).

Statistical model comparison clearly favors the phylogenetic model, with a log-Bayes factor of approximately 117 in comparison to the hierarchical model.

Parameter	Mean	SE_Mean	SD	2.5%	97.5%	n_eff	Rhat
$\sigma_x$	0.807	0.001	0.037	0.739	0.885	3012	1.00
$\sigma_y$	0.740	0.001	0.034	0.678	0.812	3463	1.00
$\lambda_x$	0.648	0.001	0.068	0.527	0.794	2974	1.00
$\lambda_y$	0.549	0.001	0.059	0.445	0.676	3724	1.00
$\mu_x$	-0.583	0.001	0.065	-0.712	-0.458	2652	1.00
$\mu_y$	-0.334	0.002	0.076	-0.483	-0.185	2417	1.00
$\rho_l$	-0.013	0.000	0.028	<b>-0.067</b>	<b>0.041</b>	18278	1.00
$\rho_f$	0.566	0.001	0.084	<b>0.389</b>	<b>0.718</b>	9843	1.00

**Table 8.** Posterior summaries for parameters in the bivariate phylogenetic OU model with family-level correlation. Residual correlation  $\rho$  includes 0 in the 95% credible interval.

This result is in stark contrast to the findings of the affix-adposition association. We find no evidence for a correlated diachronic evolution of population size and segment inventory size. But how do we account for the family-level correlation?

When two variables are correlated, two causal scenarios have to be considered. The correlation may be due to a direct causal relationship between the two variables, or it may be due to a third variable exerting causal influence on both observed variables. The absence of small-scale coevolution essentially excludes the possibility of a direct causal link. Therefore, the family-level correlation must be due to a third variable. This common cause is arguably connected to geography, paired with contingent historical events.

These findings reinforce the conclusions reached by Moran et al. (2012) via a hierarchical regression model.

### 3.5. Summary

The results of the two case studies illustrate the importance of controlling for genealogical dependencies in typological data. The association between affix position and adposition type is a robust diachronic tendency, observable both within and across language families. In contrast, the apparent correlation between phoneme inventory size and population size disappears at the level of individual languages when phylogenetic relationships are taken into account, suggesting that the observed association is a by-product of shared ancestry or geography rather than a result of co-evolution.

It is also instructive to look at the results of statistical model comparison for both studies. The methods PSIS-LOO (*Pareto-smoothed importance sampling leave-one-out cross-validation*, as implemented in the R-package *loo*, Vehtari et al. 2020) and log-Bayes Factor (using the R-package *bridgesampling*, Gronau, Singmann, and Wagenmakers 2020) were used to compare the models. The results are summarized in Table 9. (The former method compares the *expected log pointwise predicted density*; a higher value indicates a better fit to the data.)

According to both methods, the fit of the data massively improves when adding family-level random intercepts. Adding phylogenetic control improves the fit even further. This indicates that the added complexity of using the phylogenetic comparative method is clearly justified when conducting typological studies.

Model Type	Affix–Adposition		Segment Inventory–Population Size	
	Bayes factor (log)	$\Delta$ elpd	Bayes factor (log)	$\Delta$ elpd
<i>Vanilla Model</i>	–1305	–407	–1022	–1194
<i>Hierarchical Model</i>	–916	–146	–140	–116
<i>Phylogenetic Model</i>	0	0	0	0

**Table 9.** Model comparison for both case studies. Columns 2–3 correspond to the affix–adposition case study, and columns 4–5 to the segment inventory–population size case study. Higher elpd and log-Bayes values indicate better model fit.

#### 4. Conclusion

This chapter has reviewed recent advances in computational typology, with a particular focus on the role of quantitative and phylogenetic methods in the investigation of language universals and structural correlations. Classic typological hypotheses, such as Greenberg’s Universal 27, as well as more recent proposals concerning the interaction of linguistic and non-linguistic variables, have been re-examined using statistical models that account for genealogical and areal dependencies.

A central methodological contribution is the application of bivariate models for mixed data types – including ordinal and binary variables – embedded in hierarchical and phylogenetic frameworks. These models allow for the joint modeling of two traits, taking into account both family-level structure and phylogenetic inertia via the Ornstein–Uhlenbeck process. This approach facilitates the distinction between correlations driven by shared descent and those that reflect co-evolution under common functional pressures.

The empirical findings illustrate the utility of this approach. The association between affix position and adposition type proves to be a robust diachronic tendency, observable both within and across language families. In contrast, the apparent correlation between phoneme inventory size and population size disappears at the level of individual languages when phylogenetic relationships are taken into account, suggesting that the observed association is a by-product of shared ancestry or geography rather than a result of co-evolution.

Overall, these results highlight the importance of controlling for non-independence in typological data. Computational models that integrate genealogical information offer a principled way to test claims about universals, revealing whether they reflect true evolutionary tendencies or are artifacts of sampling and historical contingency. As typological datasets continue to grow in coverage and detail, such models will play a central role in refining the understanding of cross-linguistic patterns and their underlying causes.

As briefly alluded to above, phylogenetic control does not fully eliminate the problem of non-independence. For instance, it does not take language contact into account. While active research in this direction is underway, the integration of phylogenetic and geostatistical methods is still in its infancy and provides ample opportunities for future research.

## Acknowledgements

This research was supported by the DFG Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement 834050).

## Data and code availability

The data and code for the analyses presented in this chapter are available at [https://codeberg.org/profgerhard/computational\\_typology\\_routledge](https://codeberg.org/profgerhard/computational_typology_routledge).

## References

- Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332(6027), 346–349.
- Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., ... Lowe, J. B. (2018). *The AUTOTYP database, release 0.1*. <https://github.com/autotyp/autotyp-data>.
- Bynon, T. (2004). Approaches to morphological typology. In G. Booij, C. Lehmann, & J. Mugdan (Eds.), *Morphology: An international handbook on inflection and word-formation* (Vol. 2, pp. 1221–1231). Berlin: De Gruyter Mouton.
- Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., ... others (2020). Northeuralex: A wide-coverage lexical database of Northern Eurasia. *Language resources and evaluation*, 54, 273–301.
- Dixon, R. M. W. (1994). *Ergativity*. Cambridge, UK: Cambridge University Press.
- Dryer, M. S. (2013a). Order of adposition and noun phrase (v2020.4) [Data set]. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.13950591>
- Dryer, M. S. (2013b). Prefixing vs. suffixing in inflectional morphology (v2020.4) [Data set]. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.13950591>
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.4)* [Data set]. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.13950591>
- Ehret, C. (1979). On the antiquity of agriculture in Ethiopia. *The Journal of African History*, 20(2), 161–177.
- Elffers, E., et al. (2008). Georg von der Gabelentz and the rise of general linguistics. In L. van Driel & T. Janssen (Eds.), *Ontheven aan de tijd. linguïstisch-historische studies voor jan noordegraaf* (pp. 191–200). Amsterdam & Münster: Stichting Neerlandistiek VU & Nodus.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–448.
- Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, 112(5), 1322–1327.
- Greenberg, J. (1963). Some universals of grammar with special reference to the order of meaningful elements. In *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29.
- Guzmán Naranjo, M., & Becker, L. (2021). Statistical bias control in typology. *Linguistic Typology*, 26, 605 - 670.

- Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2020). *Glottolog 4.2.1*. Jena: Max Planck Institute for the Science of Human History.
- Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, *51*(5), 1341–1351.
- Harmon, L. (2019). *Phylogenetic comparative methods*. Independent Traverse City, MI, USA.
- Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*(2), 388–400.
- Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*(2), 281–319.
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, *5*.
- Jäger, G., List, J.-M., & Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*. ACL.
- Janhunen, J. (2024). The Mongolic language family. In E. Vajda (Ed.), *The languages and linguistics of Northern Asia: Language families* (Vol. 10, pp. 75–122). Walter de Gruyter GmbH & Co KG.
- Jäger, G. (2025). *Global-scale phylogenetic inference from asjp19*. Open Science Framework. Retrieved from <https://doi.org/10.17605/OSF.IO/A97SZ> (Version 1)
- Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (Sixteenth ed.). SIL International. (Online version: <http://www.ethnologue.com>)
- List, J.-M., Forkel, R., Greenhill, S. J., Rzymiski, C., Englisch, J., & Gray, R. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, *9*, 316.
- Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.
- McElreath, R. (2016). *Statistical rethinking. a Bayesian course with examples in R and Stan*. Boca Raton: CRC Press.
- Michaelis, S. M., Maurer, P., Haspelmath, M., & Huber, M. (Eds.). (2013). *Apics online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from <https://apics-online.info/>
- Moran, S., & McCloy, D. (2019). *PHOIBLE 2.0*. Available online at <http://phoible.org>, Accessed on 2019-06-15.
- Moran, S., McCloy, D., & Wright, R. (2012). Revisiting the population vs phoneme-inventory correlation. *Language*, *88*(4), 877–893.
- Naroll, R. (1961). Two solutions to Galton’s problem. *Philosophy of Science*, *28*(1), 15–39.
- Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290.
- R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/> (R package version 4.3.1)
- Roberts, S., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS ONE*, *8*(8), e70902.
- Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latache, J. J., ... Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. , *9*(16).
- Stan Development Team. (2022). Stan reference manual, v 2.31.0 [Computer software manual]. Retrieved from <https://mc-stan.org/>
- Stan Development Team. (2024). *RStan: the R interface to Stan*. Retrieved from <https://mc-stan.org/> (R package version 2.32.6)
- Tria, F. D. K., Landan, G., & Dagan, T. (2017). Phylogenetic rooting using minimal ancestor deviation. *Nature ecology & evolution*, *1*(7), 0193.
- Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, *36*(5), 823.
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. <https://mc-stan.org/loo/>. (R

- package version 2.4.1)
- von der Gabelentz, G. (1901). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse* (Reprint, with an introduction by E. Coseriu, 1984 ed.). Tübingen: Gunter Narr.
- Wichmann, S., Holman, E. W., & Brown, C. H. (2018). *The ASJP database (version 18)*. <http://asjp.clld.org/>.
- Wichmann, S., Holman, E. W., & Brown, C. H. (2020). *The ASJP database (version 19)*. <http://asjp.clld.org/>.
- Wichmann, S., Holman, E. W., & Brown, C. H. (2022). *The ASJP database (version 20)*. <http://asjp.clld.org/>.