A Statistical Model for the Joint Inference of Vertical Stability and Horizontal Diffusibility of Typological Features

Yugo Murawaki[1#], Kenji Yamauchi[2*]

[1]Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

[2]LINE Corporation, 4-1-6 Shinjuku, Shinjuku-ku, Tokyo, 160-0022, Japan

[#]Corresponding author: murawaki@i.kyoto-u.ac.jp

Abstract

A major pursuit within the study of language evolution is to advance understanding of the historical behavior of typological features. Previous studies have identified at least three factors that determine the typological similarity of a pair of languages: (1) vertical stability, (2) horizontal diffusibility, and (3) universality. Of these factors, the first two are of particular interest. Although observed data are affected by all three factors to a greater or lesser degree, previous studies have not jointly modeled them in a straightforward manner. Here, we propose a solution that is derived from the field of cultural anthropology. We present a simple and extensible Bayesian autologistic model to jointly infer the three factors from observed data. Although a large number of missing values in the dataset pose serious difficulties for statistical modeling, the proposed model can robustly estimate these parameters as well as missing values. Applying missing value imputation to indirectly evaluate the estimated parameters, we quantitatively demonstrated that they were meaningful. In conclusion, we briefly compare our findings with those of previous studies and discuss future directions.

1 Introduction

Linguistic typology is concerned with the systematic comparison of language structures. Major linguistic types, or *features*, by which the world's languages are classified, include basic word order (Order of Subject, Object and Verb) and the presence or absence of tone. For example, basic word order concerns the ordering of subject, object, and verb: English is SVO (Subject-Verb-Object) while Japanese is SOV (Subject-Object-Verb). Features of linguistic typology constitute a promising resource that can potentially be used to uncover the evolutionary

[*]The work of the second author was done when he was a student in the Graduate School of Informatics, Kyoto University.

history of languages. It has been argued that in exceptional cases, typological features can reflect a time span of 10,000 years or more (Nichols, 1994).

One of the greatest challenges entailed in typology-based phylogenetic inference is that in contrast to lexical evidence, which are the default choice of historical-comparative linguistics, typological features are characterized by uncertainty. Assessing whether words grouped together are cognates that exhibit regular sound correspondences requires binary judgments for which the proficiency of logic-based human inference has already been demonstrated. Because of the arbitrary nature of linguistic signs, it is extremely unlikely that unrelated words coincidentally follow sound laws. Thus, once a sufficient number of cognates are identified within a group of languages, we can reasonably conclude that they share a common ancestor.[1] By contrast, the sharing of the same basic word order, for example, an SOV is only a weak indicator of shared ancestry, because this pattern is repeated multiple times in multiple places. Whereas human beings are not adept at handling uncertainties, we believe that computer-intensive statistical methods (Tsunoda et al., 1995; Dunn et al., 2005; Longobardi and Guardiano, 2009; Murawaki, 2015) can overcome this problem, because they provide effective ways of combining weak signals.

In this article, we demonstrate a step forward toward developing typology-based phylogenetic inference. Specifically, we engage with the problem of the obscuring of vertical (phylogenetic) signals by horizontal (spatial or areal) transmission. Typological features, like loanwords, can be borrowed from one language to another. In fact, a non-tree-like mode of evolution has emerged as one of the central topics in linguistic typology (Trubetzkoy, 1928; Campbell, 2006).

However, the tree model, which is the standard statistical model used for phylogenetic inference, only reflects vertical transmission. The task of incorporating both vertical and horizontal transmission within a statistical model is notoriously challenging because of the excessive flexibility of horizontal transmissions. Although the tree model also requires a vast search space, this issue can be handled using computer-intensive models (Felsenstein, 1981; Gray and Atkinson, 2003). Whereas the degree of uncertainty can be intuitively posited to increase as we trace the states of languages back to the past, the tree model, when used for this process, simultaneously reduces the degree of freedom by repeatedly merging nodes into a parent. Horizontal transmission elicits a degree of freedom that cannot be currently modeled without imposing some strong assumptions (Daumé III, 2009; Nelson-Sathi et al., 2010).

Consequently, here we pursue a line of research in linguistic typology that draws on information on the current distribution of typological features without explicitly requiring the reconstruction of previous states (Nichols,

[1] Once genetic relationship of languages is established, other types of evidence such as the ordering of regular sound changes and historical changes in inflectional paradigms would be needed to subgroup them.

1992, 1995; Parkvall, 2008; Wichmann and Holman, 2009). Studies that have applied this approach have attempted to quantify how *stable* typological features are. In their original forms, typological features constitute a mosaic exhibiting varying degrees of stability. If we succeed in discovering stable features, these could be regarded as typological equivalents of *basic vocabulary*, or as a list of basic concepts. Words within the list are assumed to be resistant to borrowing (Swadesh, 1971), and some words are even considered to be extremely stable (Pagel et al., 2013). Such stable traits can potentially enable deep phylogenetic relations existing among the world's languages to be uncovered.

The question then is how is the notion of stability connected to vertical and horizontal transmission? Answering this question is not as simple as we may think. Previous studies offer a surprisingly wide variety of perspectives on this question (Dediu and Cysouw, 2013). However, it is evident that at least three factors must be considered. These factors are: (1) vertical stability, (2) horizontal diffusibility, and (3) universality.[2] In other words, there are three possible explanations for the sharing of the same feature value by a pair of languages:

1. Two languages may be phylogenetically related, with the feature demonstrating a strong phylogenetic association (i.e., vertical stability).

2. The two languages may be spatially proximate, with the feature demonstrating a strong spatial association (i.e., horizontally diffusibility).

3. The third explanation lies in a bias term. In other words, the more often the feature value appears globally, the higher the probability of assuming the value is.[3]

Of course, these three explanations are not mutually exclusive. The current distribution of typological features is more or less affected by all three factors. Thus, the first step toward developing typology-based phylogenetic inference is to model all these factors jointly.

1.1 A Solution from Cultural Anthropology

We suggest that a parallel can be drawn between cultural anthropology and linguistics with respect to vertical and horizontal transmission. Within cultural anthropology, these two modes of vertical and horizontal transmission are known as phylogenesis and ethnogenesis, respectively (Collard and Shennan, 2000). Phylogenesis involves the transmission of cultural traits primarily from an ancestral population to its descendants, whereas ethnogenesis entails strong influences resulting from transmissions between populations.

[2]Like Dediu and Cysouw (2013) and others, we treat each typological feature separately. Modeling dependencies of a typological feature with other typological features (Murawaki, 2015) or extralinguistic features such as social structure (Lupyan and Dale, 2010) and climate (Everett et al., 2015) are out of scope of the present study.

[3]While one of the main goals of linguistic typology is to uncover universals, universality in our framework is nothing more than a factor that can be explained neither phylogenetically nor areally. We will revisit this point in Section 3.3.4.

A model that is potentially capable of disentangling these factors has already been proposed by Towner et al. (2012). This model is a variant of the autologistic model (Besag, 1974), which has been widely used to model the spatial distribution of a feature. The model assumes that the value of a random variable depends probabilistically on the values of its neighbors. Towner et al. (2012)'s extension incorporates two neighbor graphs: one for vertical transmission and the other for horizontal transmission. These graphs are associated with scalar parameters indicating the relative strength of the transmission modes.

Towner et al. (2012) applied the autologistic model to examine cultural traits (features) of North American Indian societies such as the presence or absence of agriculture, the tendency toward exogamy, and types of social structure. They constructed a spatial neighbor graph using longitudinal and latitudinal data on the societies, as well as a phylogenetic neighbor graph, created by collapsing known language families. Applying this model, they empirically demonstrated that both transmission modes were non-negligible for the majority of traits. The same model was subsequently applied to folktales of Indo-European societies by da Silva and Tehrani (2016). Applied at an intermediary step toward discovering folktales with deep historical roots, the model was used to filter out those tales with strong horizontal signals.

Although the key idea of jointly modeling vertical and horizontal transmissions was built into the model formulated by Towner et al. (2012), it was not readily applicable to typological data. For this reason, we modified the model and added a new inference algorithm, as described in this article. Our model primarily differs from the above described model in two ways. First, whereas the original model only deals with binary features, our extended model can handle categorical features. Although the original database of cultural traits was coded categorically, Towner et al. (2012) selected its small subset by excluding unsuitable features and merging feature values. Second, because Towner et al. (2012) only focused on high-coverage features, they simply excluded languages with missing values from neighbor graphs. However, the linguistic typological database is too sparse to ignore missing values. The situation is unlikely to change in the foreseeable future because of the thousands of languages in the world, there is ample documentation for only a handful. Consequently, we take a Bayesian approach because it is known for its robustness in modeling uncertainties (Murphy, 2012:Chapter 2). We used a Markov chain Monte Carlo (MCMC) method to jointly infer scalar parameters and missing values The source code is available from GitHub at: `https://github.com/murawaki/bayes-autologistic`.

1.2 Brief Review of Previous Studies

Though a thorough review of previous studies is beyond the scope of this article, to the best of our knowledge, there are no previous studies in linguistic typology that have jointly modeled all three factors in a straightforward manner. More details on these earlier studies can be found in a review by Wichmann (2015), as well as

in a review and empirical comparison conducted by Dediu and Cysouw (2013). Here we briefly describe how three proposed approaches, described in the literature, have ignored one or more factors or have suffered from contamination.

Dediu (2010) used the tree model to infer stability (hereafter, stability *D*). The implication of his study was that stability *D* does not explicitly reflect horizontal diffusibility. He assumed that a feature is passed on from a parent to a child within a family tree and obeys a continuous-time Markov chain model in which the transition rate matrix is constrained to have only one parameter. This parameter is estimated for each of several given trees and is then merged into one index, that is, stability *D*. In general, a transition rate matrix controls not just the speed of change (i.e., instability), but it also determines the stationary distribution toward which the model converges. We can consider the stationary distribution to reflect universality because it is where vertical influence vanishes. However, the stationary distribution in Dediu's model is trivial, because only one parameter controls the transition rate matrix. A binary-coded feature, for example, is doomed to take value 0 with probability 0.5 and 1 with probability 0.5. In short, stability *D* also ignores universality.

Parkvall (2008) presented a series of calculations for deriving a stability index (stability *P*). First, he computed the Herfindahl-Hirschman index for a given group G, defined either phylogenetically or areally (spatially):

$$D_G = 1 - \sum_{j=1} (p_j)^2,$$

where p_j is the relative frequency of the feature value j in G. Because this was a diversity index, he calculated a homogeneity index, H_G, using the reciprocal of D_G. He then averaged H_G across all phylogenetic groups to obtain H_{FAM}. H_{ARE} denoted the areal equivalent of H_{FAM}. As a final step, he divided H_{FAM} by H_{ARE} to obtain the stability *P*.

Stability *P* takes into account both vertical stability (H_{FAM}) and horizontal diffusibility (H_{ARE}). However, D_G suffers from noise because G does not reflect purely phylogenetic (or spatial) signals. A model that performs one-step disentanglement rather than a combination of noisy estimation and correction would thus be ideal. The question of universality remains, which is obscured by the series of calculations. However, H_{FAM} should probably be seen as an amalgam of vertical stability and universality (the same is true of H_{ARE}).

The stability index developed by Wichmann and Holman (2009) (stability *W*) is similar to that of Parkvall, but only considers phylogenetic groups. Whereas Wichmann and Holman introduced three metrics, here we focus on metric C, which has been found to perform the best within simulation experiments. For each phylogenetic group, G, they calculated V_G, the proportion of language pairs in G that shared the same feature value. Next, they averaged V_G across all phylogenetic groups weighted by the inverse of the square of the number of

languages in G:

$$R = \sum_{G} \frac{V_G}{\sqrt{|G|}}.$$

Similarly, they calculated U as the proportion of pairs of unrelated languages sharing the same value. Stability *W* was obtained by correcting for U as follows:

$$\frac{R - U}{1 - U}.$$

Though stability *W* accounts for vertical stability and universality, it does not explicitly consider horizontal diffusibility. Wichmann and Holman (2009) justified this exclusion by pointing out that diffusibility is largely area-specific. Nevertheless, it is desirable to control for horizontal diffusibility when computing vertical stability.

To summarize, the methods reviewed above either simply ignore one or more factors, or they entail the performance of noisy estimation, followed by a correction step. In what follows, we will show that all three factors can jointly be modeled in a straightforward manner. We also argue that it is better to keep apart vertical stability and horizontal diffusibility, rather than conflating them into a single stability index.

2 Materials and Methods

2.1 Dataset

Although the proposed model does not depend on a particular dataset, in this article we focus on one database, namely, the online edition of the World Atlas of Language Structures (WALS) (Haspelmath et al., 2005). As of 2017, this atlas covered 2,679 languages and 192 typological features. However, the associated language-feature matrix was very sparse, containing less than 15% of the items.

An item of the language-feature matrix was defined as a categorical value. For example, Feature 81A, "Order of Subject, Object and Verb" had seven possible values: `SOV`, `SVO`, `VSO`, `VOS`, `OVS`, `OSV` and `No dominant order`, and each language incorporated one of these seven values. We used these categorical feature values without making any modifications. Although mergers of some fine-grained feature values appear to be desirable (Daumé III and Campbell, 2007; Greenhill et al., 2010; Dediu, 2010; Dunn et al., 2011), we leave this work for a future investigation. Some of the previous studies discarded languages with few observed features (Murawaki, 2015; Takamura et al., 2016). To avoid introducing arbitrariness by ourselves, we included all but 72 languages in the database. The excluded languages were sign languages and pidgin and creole languages, all of which have been categorized in WALS as belonging to a dummy phylogenetic group labeled

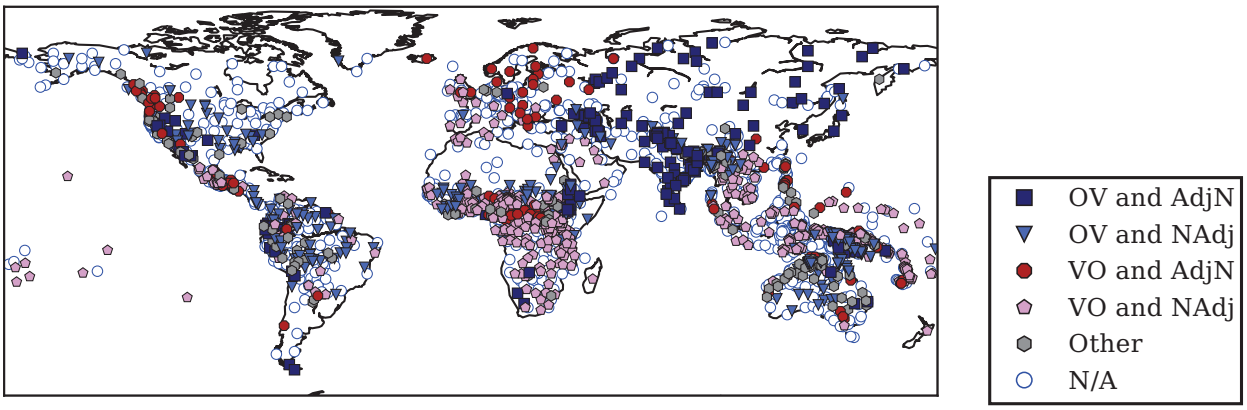


Figure 1: The geographical distribution of Feature 97A ("Relationship between the Order of Object and Verb and the Order of Adjective and Noun"). Each point denotes a language. There are a non-negligible number of missing values (denoted as `N/A`).

"other."

Out of the 192 features, we selected 82 features that covered at least 10% of the remaining languages and applied to the entire world. For example, we dropped Feature 144H ("NegSVO Order"), because it was only defined for SVO languages.

WALS was also the source for constructing phylogenetic and spatial neighbor graphs. This atlas provides two levels of grouping: family and genus. We used the genera level to construct a phylogenetic neighbor graph in which every pair of languages within a genus was linked. The analysis showed that the average number of phylogenetic neighbors was 30.8. Other resources such as Glottolog (Hammarström et al., 2016) and Ethnologue (Lewis et al., 2014) provided more detailed hierarchical classifications. However, there seemed to be no way to select groups of languages of comparable time depth.

Languages were found to be associated with single-point geographical coordinates (longitude and latitude), as shown in Figure 1. We constructed a spatial neighbor graph by linking all language pairs that were located within a distance of R km. Following da Silva and Tehrani (2016), we set the value of R at 1,000. The average number of spatial neighbors was 89.1.[4] The phylogenetic and spatial neighbor graphs had substantial overlap. On average, 71.4% of phylogenetic neighbors were simultaneously spatial neighbors while 20.1% of spatial neighbors were simultaneously phylogenetic neighbors.

2.2 Counting Functions

We begin our description of the model by introducing three counting functions corresponding to vertical, horizontal, and universal factors. A counting function receives the feature values of the world's languages and

[4]da Silva and Tehrani (2016) chose the distance threshold such that the average number of spatial neighbors was roughly the same as that of phylogenetic neighbors. In our case, the former is nearly three times as large as the latter, however.

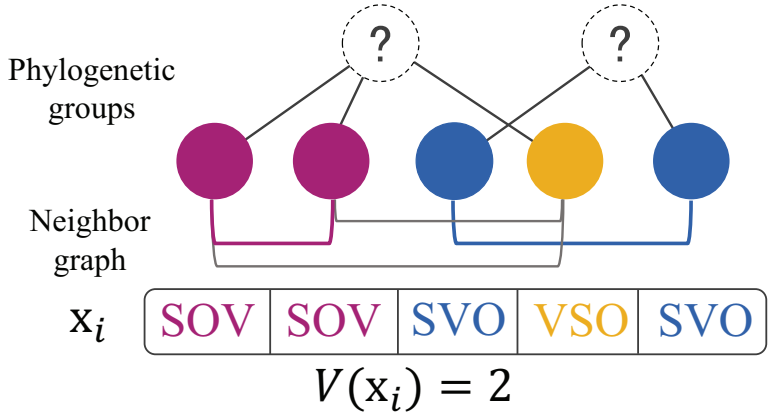


Figure 2: The phylogenetic neighbor graph and $V(\mathrm{x}_i)$. Each node denotes a language. Languages are grouped by common ancestors whose feature values are unknown. The neighbor graph consists of edges, each of which connects a pair of languages within the same group. In this example, two edges indicate the sharing of feature values.

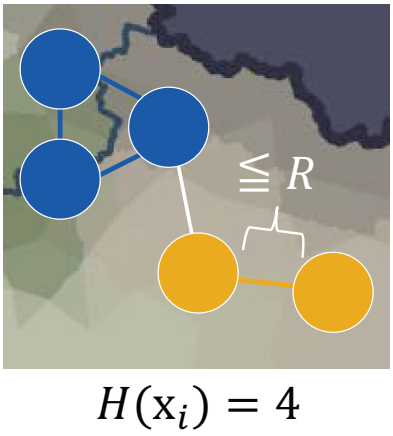


Figure 3: The spatial neighbor graph and $H(\mathrm{x}_i)$. Of the five edges in the neighbor graph, four indicate feature value-sharing.

returns a non-negative scalar. Formally, let $\mathrm{x}_i = (x_{1,i}, \cdots, x_{L,i}) \in \mathcal{X}_i$ be a sequence of feature values, where $x_{l,i}$ is the value of the l-th language for feature type i. $x_{l,i}$ takes one of F_i categorical values, and F_i differs according to feature types. For now, we can assume that there are no missing values.

The vertical factor is represented by $V(\mathrm{x}_i)$. It returns the number of pairs sharing the same value in the phylogenetic neighbor graph. Figure 2 provides an example. As the basis of this function, we intuitively posit that if the feature in question is vertically stable, then a phylogenetically defined group of languages will tend to share the same value. A more direct counting method would entail checking how many changes have occurred relating to parent-to-child transitions within the trees. However, this alternative method would be difficult to implement, because the feature values of ancestors remain unknown. Consequently, we chose a counting function that only refers to contemporary languages.

$H(\mathrm{x}_i)$ is the spatial equivalent of $V(\mathrm{x}_i)$, as illustrated in Figure 3. If the feature in question is horizontally

diffusible, then spatially close languages would be expected to frequently share the same feature value.

Last, $U_j(\mathrm{x}_i)$ is the number of languages that take the value of j. If j is universally common, then the value of $U_j(\mathrm{x}_i)$ must be large.

2.3 Autologistic Model

We now introduce the following parameters: vertical stability v_i, horizontal diffusibility h_i, and universality $\mathrm{u}_i = (u_{i,1}, \cdots, u_{i,F_i})$ for each feature i. Applying the autologistic model, the probability of x_i, conditioned on v_i, h_i and u_i, can be expressed as:

$$P(\mathrm{x}_i \mid v_i, h_i, \mathrm{u}_i) = \frac{\exp\left(v_i V(\mathrm{x}_i) + h_i H(\mathrm{x}_i) + \sum_j u_{i,j} U_j(\mathrm{x}_i)\right)}{\sum_{\mathrm{x}'_i \in \mathcal{X}_i} \exp\left(v_i V(\mathrm{x}'_i) + h_i H(\mathrm{x}'_i) + \sum_j u_{i,j} U_j(\mathrm{x}'_i)\right)}. \tag{1}$$

The denominator is a normalization term, ensuring that the sum of the distribution equals one. This term is computationally intractable, because there are $(F_i)^L$ possible assignments of x_i. However, as we will show, the parameters can be estimated without precisely calculating the normalization term.

The autologistic model can be interpreted in terms of the competition associated with possible assignments of x_i for the probability mass 1. If a given value, x_i, has a relatively large $V(\mathrm{x}_i)$, then setting a large value for v_i enables it to appropriate fractions of the mass from its weaker rivals. However, if too large a value is set for v_i, then it will be overwhelmed by its stronger rivals.

To acquire further insights into the model, let us consider the probability of language l taking the value k, conditioned on the rest of the languages, $\mathrm{x}_{-l,i}$:

$$P(x_{l,i} = k \mid \mathrm{x}_{-l,i}, v_i, h_i, \mathrm{u}_i) \propto \exp\left(v_i V_{l,i,k} + h_i H_{l,i,k} + u_{i,k}\right), \tag{2}$$

where $V_{l,i,k}$ is the number of language l's phylogenetic neighbors that assume the value k, and $H_{l,i,k}$ is its spatial counterpart. The derivation of Eq. (2) is explained in Section S.1 of this article. The probability of language l assuming value k is expressed by the weighted linear combination of the three factors in the log-space. If the vertical stability parameter v_i is positive, then the probability will increase with a rise in the number of phylogenetic neighbors that assume value k. However, this probability depends not only on the phylogenetic neighbors of language l, but it also depends on its spatial neighbors and on universality. How strongly these factors affect the stochastic selection is controlled by v_i, h_i, and u_i.

We employed a Bayesian approach because it is known for its robustness in modeling uncertainties. We set prior distributions to v_i, h_i and $u_{i,j}$ as follows: $v_i \sim \mathrm{Gamma}(\kappa, \theta)$, $h_i \sim \mathrm{Gamma}(\kappa, \theta)$, and $u_{i,j} \sim \mathcal{N}(0, \sigma^2)$,

where κ, θ and σ were hyperparameters. Note that by setting Gamma priors, we constrained v_i and h_i to always be positive. Negative v_i and h_i would be difficult to justify from a linguistic point of view, because they indicate that languages attempt to differentiate themselves from other languages to which they are related. We set shape $\kappa = 1.0$, scale $\theta = 1.0$, and standard deviation $\sigma = 10.0$. These priors were not non-informative because we wanted to penalize extreme values. However, they were sufficiently gentle in the regions where these parameters typically resided.

The fact that the phylogenetic and spatial neighbor graphs had substantial overlap was likely to cause difficulty disentangling the two factors. As a possible solution to the problem, we considered a model variant with an additional parameter m_i:

$$P(\mathrm{x}_i \mid v_i, h_i, m_i, \mathrm{u}_i) = \frac{\exp\left(v_i \tilde{V}(\mathrm{x}_i) + h_i \tilde{H}(\mathrm{x}_i) + m_i M(\mathrm{x}_i) + \sum_j u_{i,j} U_j(\mathrm{x}_i)\right)}{\sum_{\mathrm{x}'_i \in \mathcal{X}_i} \exp\left(v_i \tilde{V}(\mathrm{x}'_i) + h_i \tilde{H}(\mathrm{x}'_i) + m_i M(\mathrm{x}'_i) + \sum_j u_{i,j} U_j(\mathrm{x}'_i)\right)}, \tag{3}$$

where $M(\mathrm{x}_i)$ counts languages that are simultaneously phylogenetic and spatial neighbors, $\tilde{V}(\mathrm{x}_i) = V(\mathrm{x}_i) - M(\mathrm{x}_i)$, and $\tilde{H}(\mathrm{x}_i) = H(\mathrm{x}_i) - M(\mathrm{x}_i)$. By tying v_i and h_i to disjoint sets of languages, we expected the autologistic model to detect vertical and horizontal signals more easily.

2.4 Inference

We can now proceed to the parameter estimation. Let x_i be decomposed into an observed portion $\mathrm{x}_i^{\mathrm{obs}}$ and the remaining missing portion be $\mathrm{x}_i^{\mathrm{mis}}$. $\mathrm{x}_i^{\mathrm{mis}}$, in addition to v_i, h_i and u_i, needed to be inferred.

We used a standard MCMC sampling algorithm (Bishop, 2006:Chapter 11). Specifically, we employed Gibbs sampling to draw posterior samples of v_i, h_i, u_i and $\mathrm{x}_i^{\mathrm{mis}}$ from

$$P(v_i, h_i, \mathrm{u}_i, \mathrm{x}_i^{\mathrm{mis}} \mid \mathrm{x}_i^{\mathrm{obs}}) \propto P(v_i, h_i, \mathrm{u}_i) P(\mathrm{x}_i \mid v_i, h_i, \mathrm{u}_i) = P(v_i) P(h_i) P(\mathrm{u}_i) P(\mathrm{x}_i \mid v_i, h_i, \mathrm{u}_i), \tag{4}$$

where hyperparameters are omitted for brevity. The algorithm is sketched in Section S.2.

Note that we needed to sample v_i (and h_i and $u_{i,j}$) from $P(v_i \mid h_i, \mathrm{u}_i, \mathrm{x}_i) \propto P(v_i) P(\mathrm{x}_i \mid v_i, h_i, \mathrm{u}_i)$. This belongs to a class of problems known as sampling from doubly-intractable distributions (Møller et al., 2006; Murray et al., 2006). While it remains a challenging problem in statistics, it is not difficult to approximately sample the variables if we give up theoretical rigorousness (Liang, 2010). The details of the algorithm we used can be found in Section S.3.[5]

[5]Towner et al. (2012) performed model selection to assess the necessity of vertical and horizontal factors. In Bayesian statistics, the standard choice for model selection is to use Bayes factors. However, estimating Bayes factors for the autologistic model is challenging because it entails yet another layer of intractability (and thus this task is called a triply intractable problem) (Friel, 2013), not to mention

3 Results and Discussion

3.1 Missing Value Imputation

While most previous studies have involved a subjective analysis of stability indices, the quality of estimated parameters can be quantitatively measured. This quantitative measuring ability is an advantageous feature of a probabilistic model with predictive ability. Specifically, estimated parameters can be indirectly evaluated by means of missing value imputation. If the autologistic model predicts missing values better than reasonable baselines, we can say that the estimated parameters are justified. Although no ground truth exists for the missing portion of the database, missing value imputation can be evaluated by hiding some observed values and verifying the effectiveness of their recovery.

For each feature type, we conducted tenfold cross validation. We randomly partitioned observed values into 10 blocks that were almost equal in size. For each of the 10 runs, we treated an individual block the same as other missing values and performed posterior sampling. After 200 burn-in iterations, we ran 2,495 iterations and collected samples with the interval of 5 iterations. Among the collected samples, we chose the most frequent feature value for each language as the final output.

The autologistic model had two variants: the basic model (Autologistic-Basic) defined in Eq. (1) and the extension (Autologistic-Disjoint) defined in Eq.(3). They were compared with three baselines.

Universal Values are derived from the empirical distribution of $\mathrm{x}_i^{\mathrm{obs}}$.

Global majority The most frequently occurring value among $\mathrm{x}_i^{\mathrm{obs}}$ is selected and this value is always the output.

Neighborhood The neighbors of each language are compiled in $\mathrm{x}_i^{\mathrm{obs}}$ and a value is derived from the empirical distribution. If both graphs have observed neighbors, one is chosen randomly. If one of the two graphs has no observed neighbor, then the other is chosen. If neither is available, then the universal baseline provides a fallback option.

Table 1 shows the overall results of the analysis. The universal baseline performed the worst, followed, in ascending order, by the global majority, and the neighborhood. The autologistic model outperformed the three baselines. The two model variants achieved comparable performance. The estimated parameters were evidently reasonable.

Let us now take a closer look at the results. Table S.1 depicts the accuracy of missing value imputations for each feature. Figure 4 and Table 2 compare the performance of the autologistic model with those of the

the marginalization of missing values. We leave it for future work.

Table 1: Results of missing value imputation.

Model	Accuracy (%)	
	Macro-average	Micro-average
Universal	42.21	42.05
Global majority	54.85	54.20
Neighborhood	57.87	58.95
Autologistic-Basic	60.32	62.03
Autologistic-Disjoint	**60.92**	**62.84**

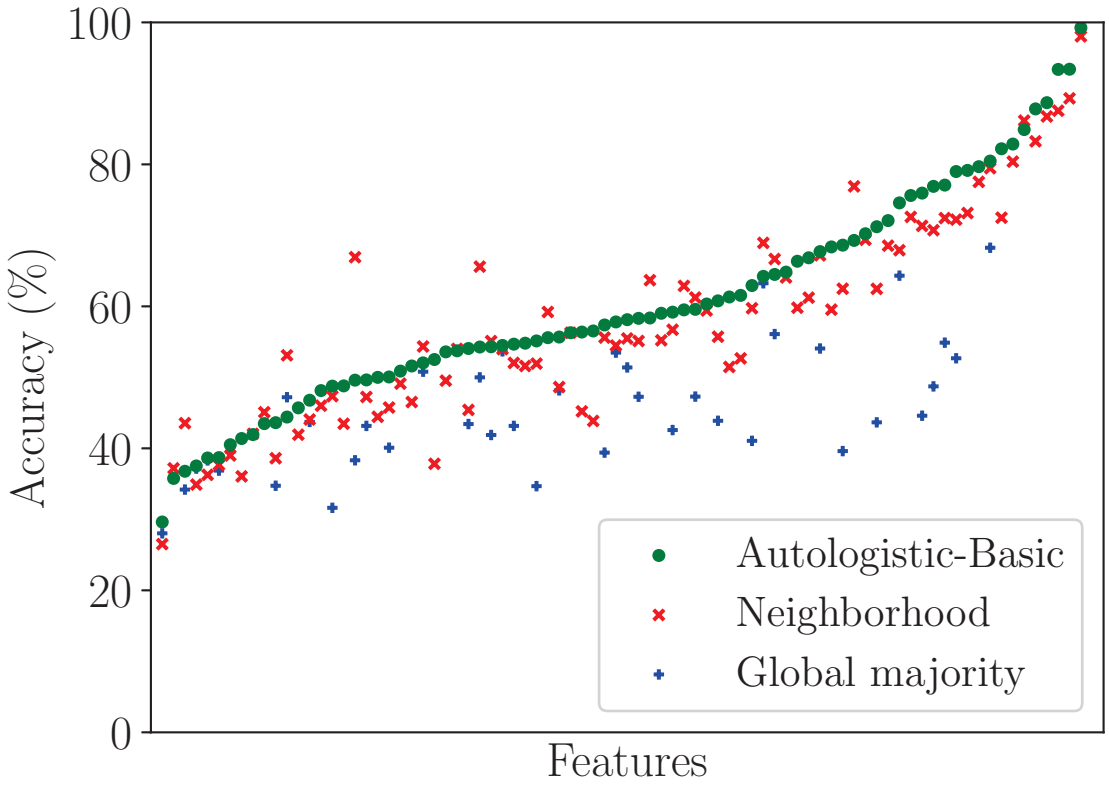


Figure 4: Accuracy per feature.

baselines. Although the autologistic model outperformed the neighborhood baseline in relation to the majority of features, it demonstrated performance loss for some of the features. For these features, the autologistic model fell in between the neighborhood and the global majority baselines, suggesting that the effect of universality was overestimated.

3.2 Estimated Parameters

We now turn to an examination of the estimated parameters. The model settings were the same as those applied in the preceding experiment, with the exception that all of the observed values were used.

Figure 5 depicts estimated parameters for the autologistic model. Note that comparing the absolute values of a v_i and an h_i makes no sense because they are tied with different neighbor graphs. Table 3 lists the top

Table 2: Features ranked by performance gain of Autologistic-Basic compared with the neighborhood baseline. The top five and bottom five ranked features are shown.

	Feature	Accuracy (%)	Gain (%)
16A	Weight Factors in Weight-Sensitive Stress Systems	52.52	+14.69
15A	Weight-Sensitive Stress	56.54	+12.68
100A	Alignment of Verbal Person Marking	56.38	+11.17
116A	Polar Questions	61.32	+9.85
111A	Nonperiphrastic Causative Constructions	82.20	+9.71
53A	Ordinal Numerals	36.77	−6.77
119A	Nominal and Locational Predication	69.29	−7.61
17A	Rhythm Types	44.41	−8.70
9A	The Velar Nasal	54.27	−11.32
118A	Predicative Adjectives	49.61	−17.32

five and bottom five features ranked according to v_i and h_i (Tables S.2 and S.3 for the complete lists). While Figure 5 represents each feature as a single point, we can use posterior samples to measure the uncertainty about the estimated parameters, as visualized in Figure S.3.

As we stated in Section 2.1, the phylogenetic and spatial neighbor graphs had substantial overlap. In fact, the arithmetic mean of v_i was moderately correlated with that of h_i for Autologistic-Basic (Spearman's $\rho = 0.454$). While the disjointness extension reduced the correlation to 0.170, we refrain from concluding that two neighbor graphs exhibited exactly what we expected them to exhibit, namely vertical and horizontal signals.

We investigated the effects of coverage (the ratio of languages whose feature values are present) on the estimated parameters. The results are shown in Figure S.1. Coverage moderately positively correlated with v_i and h_i (Spearman's $\rho = 0.416$ and 0.425, respectively). This can be explained by the fact that word order-related features, which happened to be of high coverage thanks to Matthew Dryer (Haspelmath et al., 2005), tended to vertically stable and horizontally diffusible. Not surprisingly, coverage showed moderate negative correlations to 95% highest posterior density (HPD) credible intervals ($\rho = -0.414$ for v_i and -0.473 for h_i), meaning that the autologistic model was more uncertain about features with larger numbers of missing values. We also checked correlation statistics regarding the number of unique values for each feature. All we found was very weak to weak correlations (Figure S.2).

As a final step, we compared estimated parameters with statements within the literature. There is no clear-cut methodological solution for undertaking a comparison of previously proposed stability indices with a combination of two parameters, namely, vertical stability v_i and horizontal diffusibility h_i. We believe that the two distinct concepts of vertical stability and horizontal diffusibility should be kept apart, rather than being conflated into a single stability index. As a tentative solution, we devised our own index by subtracting h_i from v_i.

Table 3: Features ranked with v_i and h_i of Autologistic-Basic. The top five and bottom five features are shown for v_i and h_i. See Tables S.2 and S.3 for complete lists and Tables S.4 and S.5 for Autologistic-Disjoint.

(a) Features ranked with the vertical stability parameter v_i.

	Feature	v_i
73A	The Optative	0.0281
33A	Coding of Nominal Plurality	0.0280
6A	Uvular Consonants	0.0278
97A	Relationship between the Order of Object and Verb and the Order of Adjective and Noun	0.0251
86A	Order of Genitive and Noun	0.0243
101A	Expression of Pronominal Subjects	0.0035
105A	Ditransitive Constructions: The Verb 'Give'	0.0034
130A	Finger and Hand	0.0034
72A	Imperative-Hortative Systems	0.0030
18A	Absence of Common Consonants	0.0024

(b) Features ranked with the horizontal diffusibility parameter h_i.

	Feature	h_i
144A	Position of Negative Word With Respect to Subject, Object, and Verb	0.0222
97A	Relationship between the Order of Object and Verb and the Order of Adjective and Noun	0.0182
93A	Position of Interrogative Phrases in Content Questions	0.0166
81A	Order of Subject, Object and Verb	0.0163
82A	Order of Subject and Verb	0.0163
34A	Occurrence of Nominal Plurality	0.0024
8A	Lateral Consonants	0.0023
94A	Order of Adverbial Subordinator and Clause	0.0022
18A	Absence of Common Consonants	0.0019
111A	Nonperiphrastic Causative Constructions	0.0013

Table 4 summarizes the findings of the comparison. Whereas we used a list created by Wichmann and Holman (2009), only its subset is shown here for the following reasons. First, some of the statements referred to feature *values* rather than feature *types*. Second, some of the statements were below the coverage threshold we set in Section 2.1. Thus, although our results were not strongly supported, they were largely consistent with previous findings. Somewhat surprisingly, tone was judged only moderately horizontally diffusible by the autologistic model even though it is known as a typical areal feature. As Wichmann and Holman (2009) observed, horizontal diffusion appears to exhibit area specificity. A possible solution to the heterogeneity is to introduce a hierarchical Bayesian model to h_i, in a way similar to relaxed molecular clock methods of evolutionary biology (Drummond and Bouckaert, 2015:Chapter 4.3): instead of letting the single parameter h_i control the whole world, we first draw a global parameter and then use it to draw area-specific parameters that are close to the global parameter on average but can be distant from it.

Table 4: A comparison of the estimated parameters with statements in the literature compiled by Wichmann and Holman (2009). For practical purposes, we labeled $v_i \geq 0.01$ as vertically stable, $0.01 > v_i \geq 0.005$ as moderate, and $v < 0.005$ as unstable. Similarly, we labeled $h_i \geq 0.01$ as horizontally diffusible, $0.01 > h_i \geq 0.005$ as moderate, and $h < 0.005$ as undiffusible. For the tentative stability index $v_i - h_i$, we set two thresholds, namely, 0.003 and -0.003, to create a three-way division.

	Statements in the literature	Corres. Feature	Estimated parameters v_i	h_i	$v_i - h_i$	Agreement
i.	Dem-N, Num-N Adj-N less stable than Gen-N and Rel-N (Hawkins, 1983:93)	88A	0.0136	0.0134	0.0002	Yes
		89A	0.0147	0.0150	−0.0003	
		87A	0.0183	0.0106	0.0077	
		86A	0.0243	0.0140	0.0102	
ii.	Place of adposition appears to be stable (Nichols, 1995:352); Adpositions are stable (Croft, 1996:206–7)	90A	0.0168	0.0043	0.0125	Yes
		85A	0.0228	0.0144	0.0084	
iii.	Definite articles are stable (Croft, 1996:206–7)	37A	0.0126	0.0102	0.0024	Moderate
iv.	Tones are stable but also areal (Nichols 1995:343, 2003:307)	13A	0.0091	0.0025	0.0066	Partial
v.	Cases are stable (Nichols, 2003:286, 307)	51A	0.0164	0.0110	0.0054	Yes
vi.	Numerical classifiers do not have high probabilities for inheritance (Nichols, 2003:299)	55A	0.0051	0.0071	−0.0020	Yes
vii.	A-removing inflection (or very regular derivation) on verbs (passive, etc.) is stable (Nichols, 1995:343)	107A	0.0070	0.0140	−0.0070	No?

3.3 Possible Extensions

The autologistic model is conceptually simple and easy to extend. In this section, we discuss several possible extensions that can be made to this model.

3.3.1 Phylogenetically and Spatially Coherent Samples

To account for uncertainty about data, a Bayesian model sometimes incorporates posterior samples generated by another Bayesian model (Pagel et al., 2004). Given that our model yielded reasonably good results on missing value imputation, we consider that the samples of $\mathrm{x}_i^{\mathrm{mis}}$ are potentially useful for such a model.

It is noteworthy that another approach exists for imputing missing values (Murawaki, 2015; Takamura et al., 2016). This approach is based on the idea that some features depend on others. Greenberg (1963) presents several rules that hold true for the world's languages, one of which stipulates the following: In languages with prepositions, the genitive almost always follows the governing noun, whereas in languages with postpositions, it almost always precedes. The scope of practical applications is not limited to pairs of features but instead uses a set of features to predict missing features. During a preprocessing step, Murawaki (2015) used a variant of multiple correspondence analysis (Josse et al., 2012) to impute missing values. Takamura et al. (2016) chose a

logistic model to investigate the predictive power of features. In their experiments, the discriminative classifier was given all but one feature of a given language and the value of the remaining feature was predicted. One language was repeatedly selected for evaluation and the classifier was trained using all of the other languages in the typological database. Compared with the dependency-based approach, the autologistic model can be considered a proximity-based approach.

Because of different experimental settings, we cannot directly compare our results with those of dependency-based methods. However, there appears to be a greater degree of improvement for the latter compared with the results obtained from the baselines. It is noteworthy that proximity is implicitly exploited by the dependency-based approach, because languages with similar feature value combinations often happen to be phylogenetically or spatially close. In fact, Takamura et al. (2016) conducted a type of ablation experiment in which they modified the training data by removing languages that shared an ancestral population in common with the target language or by excluding languages that exhibited spatial proximity to the target. They demonstrated that accuracy declined in both settings. This finding implies that vertical and horizontal clues are useful for imputing missing values, although they do not control for the tendency of lower volumes of training data to decrease test accuracy. Nevertheless, the proximity-based approach is largely complementary to the dependency-based one.

An advantage of the proximity-based approach over the dependency-based one is that the former produces phylogenetically and spatially coherent samples. Because a dependency-based model imputes missing features of each language independently of those of other languages, imputed languages are likely to disrupt vertical and horizontal signals. A combined model, which we would like to work on in the future, would be expected to improve accuracy of missing value imputation while retaining phylogenetic and spatial coherence.

3.3.2 Parameters for Feature Values

We defined the parameters of vertical stability v_i and horizontal diffusibility h_i for each feature $type$. However, some feature values could be more vertically stable and/or horizontally diffusible than others. The procedure for estimating these parameters for each feature *value* is a simple one. It merely requires replacing v_i with $\mathrm{v}_i = (v_{i,1}, \cdots, v_{i,F_i})$, and h_i with $\mathrm{h}_i = (h_{i,1}, \cdots, h_{i,F_i})$ and modifying the model accordingly.

3.3.3 Geography and Cultural Contacts

A spatial neighbor graph was constructed by connecting all languages located within a distance range of 1,000 km. This appears to be a rather arbitrary method, because, for example, the Himalayan range is much more difficult to cross than the Eurasian Steppe. A sophisticated landscape-based (Bouckaert et al., 2012) and/or human-centric (Marck, 1986) model is thus needed to take account of these ecological factors.

A much simpler extension to the model comprises a weighted variant of the spatial neighbor graph in which a spatially distant pair of languages carries a smaller weight. For the weighted variant, the number of pairs $H(\mathrm{x}_i)$ is replaced with the sum of the edge weights. Preliminary experiments indicated that this simple weighting scheme did not lead to significant differences in the results, only slightly improving performance in terms of missing value imputation (from 62.03% to 62.16% in microaveraged accuracy). We also tested the spatial neighbor graph with different distance thresholds. In terms of missing value imputation, the 500 km threshold yielded the best performance, but performance gaps were not large (Figure S.4). Thus, tuning the distance threshold is of minor importance.

Another factor that needs to be considered is cultural contacts, because the duration of contact of a pair of languages is not necessarily proportional to spatial proximity. If some proxy data for the degree of cultural contacts could be obtained, these data could be incorporated into $H(\mathrm{x}_i)$.

3.3.4 Explaining the Universality Factor

In the present study, universality is another name for a bias term. It was incorporated into the autologistic model only to correct for the *default* distribution explained neither by vertical stability nor horizontal diffusibility. However, uncovering universals is one of the main goals of linguistic typology.

In fact, the autologistic model can be viewed from the opposite side. To uncover something truly universal, we need to correct for the fact that languages are not independent (Daumé III and Campbell, 2007), and the autologistic model does its job. Thus, explaining the universality factor by means of extralinguistic features (Lupyan and Dale, 2010; Everett et al., 2015), for example, would be interesting future work. If data for such features are available, they can be straightforwardly incorporated into the model.

In conclusion, we have presented a statistical model that jointly models the distribution of typological features in terms of vertical stability, horizontal diffusibility, and universality. The application of missing value imputation demonstrated that the estimated parameters were meaningful. We believe that the present study will contribute to the broader interdisciplinary research community on language evolution. Though we have emphasized statistical modeling in this article, a future in-depth analysis of estimated parameters would be fruitful. Although computer intensive, the autologistic model is conceptually simple and extensible. Extensions to the model are a promising future direction.

Acknowledgments

The key ideas and early experimental results were presented at the 26th International Conference on Computational Linguistics (Yamauchi and Murawaki, 2016). This article, which presents updated results, is a substantially extended version of the earlier conference paper. This work was partly supported by JSPS KAKENHI Grant Number 26730122.

References

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012. doi: 10.1126/science.1219669.

Lyle Campbell. Areal linguistics. In *Encyclopedia of Language and Linguistics, Second Edition*, pages 454–460. Elsevier, 2006.

Mark Collard and Stephen J. Shennan. Ethnogenesis versus phylogenesis in prehistoric culture change: a case-study using European Neolithic pottery and biological phylogenetic techniques. In Colin Renfrew and Katherine V. Boyle, editors, *Archaeogenetics: DNA and the Population Prehistory of Europe*. McDonald Institute for Archaeological Research, 2000.

William Croft. *Typology and Universals*. Cambridge University Press, 1st edition, 1996.

Sara Graça da Silva and Jamshid J Tehrani. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science*, 3(1), 2016. doi: 10.1098/rsos.150645.

Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601, 2009.

Hal Daumé III and Lyle Campbell. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, 2007.

Dan Dediu. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1704):474–479, 2010. doi: 10.1098/rspb.2010.1595.

Dan Dediu and Michael Cysouw. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS ONE*, 8(1):1–20, 2013. doi: 10.1371/journal.pone.0055009.

Alexei J. Drummond and Remco R. Bouckaert. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press, 2015.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, 2005. doi: 10.1126/science.1114615.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011. doi: 10.1038/nature09923.

Caleb Everett, Damián E. Blasi, and Seán G. Roberts. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, 112(5):1322–1327, 2015. doi: 10.1073/pnas.1417413112.

Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981. doi: 10.1007/BF01734359.

Nial Friel. Evidence and Bayes factor estimation for Gibbs random fields. *Journal of Computational and Graphical Statistics*, 22(3):518–532, 2013. doi: 10.1080/10618600.2013.778780.

Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003. doi: 10.1038/nature02029.

Joseph H. Greenberg, editor. *Universals of language*. MIT Press, 1963.

Simon J. Greenhill, Quentin D. Atkinson, Andrew Meade, and Russel D. Gray. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693):2443–2450, 2010. doi: 10.1098/rspb.2010.0051.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, 2016. URL `http://glottolog.org`.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, 2005.

John A. Hawkins. *Word Order Universals*. Academic Press, 1983.

Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29(1):91–116, 2012. doi: 10.1007/s00357-012-9097-0.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World, 17th Edition*. SIL International, 2014. Online version: http://www.ethnologue.com.

Faming Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010. doi: 10.1080/00949650902882162.

Giuseppe Longobardi and Cristina Guardiano. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706, 2009. doi: 10.1016/j.lingua.2008.09.012.

Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PLOS ONE*, 5(1): e8559, 2010. doi: 10.1371/journal.pone.0008559.

Jeffrey C. Marck. Micronesian dialects and the overnight voyage. *The Journal of the Polynesian Society*, 95 (2):253–258, 1986.

Jesper Møller, Anthony N. Pettitt, R. Reeves, and Kasper K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006. doi: 10.1093/biomet/93.2.451.

Yugo Murawaki. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, 2015. doi: 10.3115/v1/N15-1036.

Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 2010. doi: 10.1098/rspb.2010.1917.

Johanna Nichols. *Linguistic Diversity in Space and Time*. University of Chicago Press, 1992.

Johanna Nichols. The spread of language around the Pacific rim. *Evolutionary Anthropology: Issues, News, and Reviews*, 3(6):206–215, 1994. doi: 10.1002/evan.1360030607.

Johanna Nichols. Diachronically stable structural features. In Henning Andersen, editor, *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics, Los Angeles 16–20 August 1993*. John Benjamins Publishing Company, 1995.

Johanna Nichols. Diversity and stability in languages. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 283–310. Oxford University Press, 2003.

Mark Pagel, Andrew Meade, and Daniel Barker. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5):673–684, 2004. doi: 10.1080/10635150490522232.

Mark Pagel, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476, 2013. doi: 10.1073/pnas.1218726110.

Mikael Parkvall. Which parts of language are the most stable? *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(3):234–250, 2008.

Morris Swadesh. *The Origin and Diversification of Language*. Aldine Atherton, 1971.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 69–76, 2016.

Mary C. Towner, Mark N. Grote, Jay Venti, and Monique Borgerhoff Mulder. Cultural macroevolution on neighbor graphs: Vertical and horizontal transmission among western north American Indian societies. *Human Nature*, 23(3):283–305, 2012. doi: 10.1007/s12110-012-9142-z.

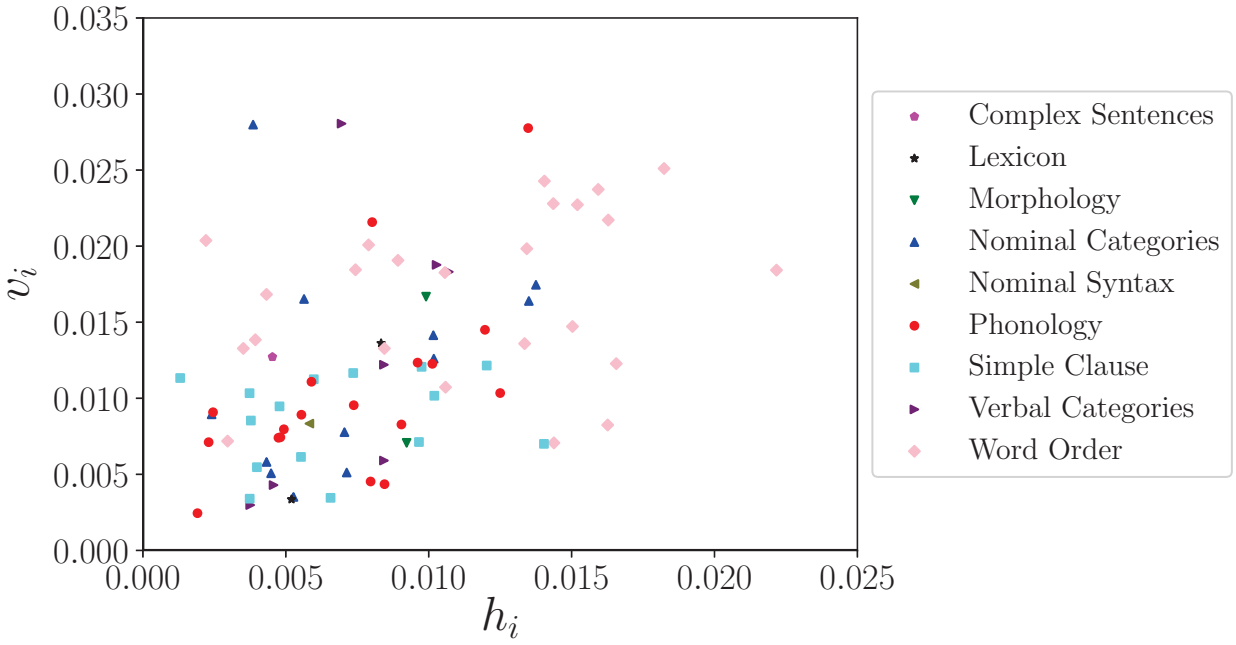
Nikolai Sergeevich Trubetzkoy. Proposition 16. In *Acts of the First International Congress of Linguists*, pages 17–18, 1928.

Tasaku Tsunoda, Sumie Ueda, and Yoshiaki Itoh. Adpositions in word-order typology. *Linguistics*, 33(4): 741–762, 1995. doi: 10.1515/ling.1995.33.4.741.

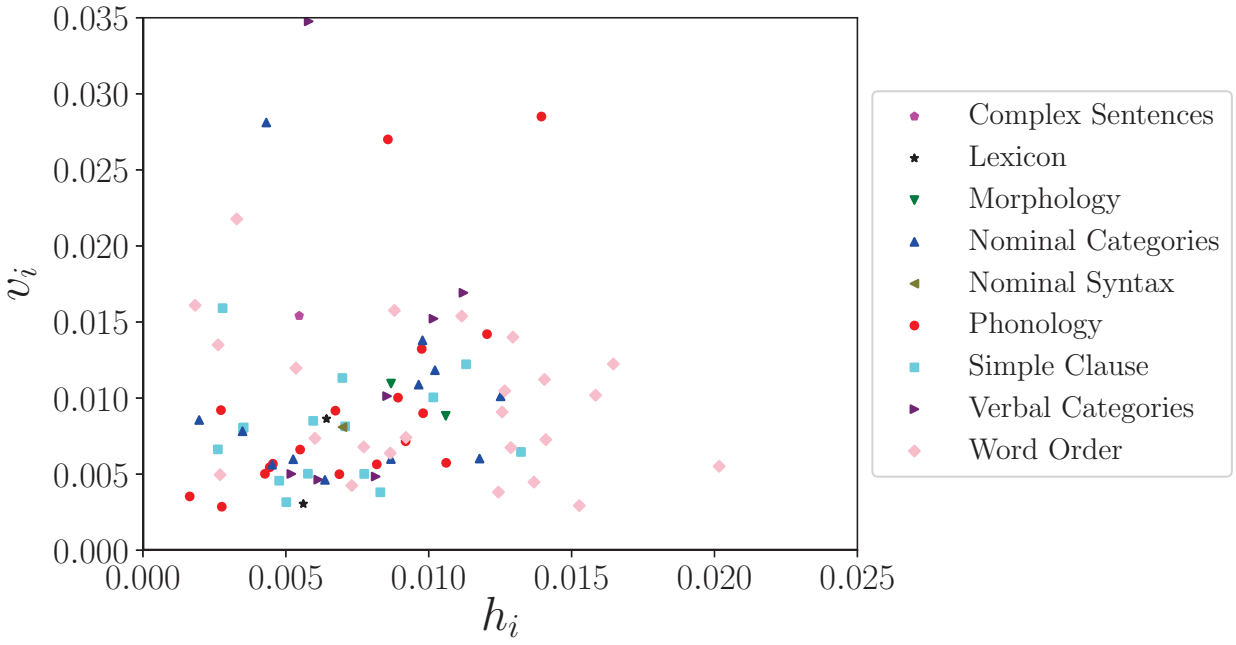
Søren Wichmann. Diachronic stability and typology. In Claire Bowern and Bethwyn Evans, editors, *The Routledge Handbook of Historical Linguistics*, pages 212–224. Routledge, 2015.

Søren Wichmann and Eric W. Holman. *Temporal Stability of Linguistic Typological Features*. Lincom Europa, 2009.

Kenji Yamauchi and Yugo Murawaki. Contrasting vertical and horizontal transmission of typological features. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 836–846, 2016.



(a) Autologistic-Basic.



(b) Autologistic-Disjoint.

Figure 5: Scatter plots of estimated parameters for the autologistic model. Each point denotes the arithmetic mean of the posterior samples. Larger v_i (h_i) indicates that feature i is more stable (diffusible). The shapes of the points represent broad categories of features (called *Area* in WALS).