Typological Word Order Correlations with Logistic Brownian Motion

Kai Hartung[1], Gerhard Jäger[2], Sören Gröttrup[1], and Munir Georges[1,3,4]

[1]Technische Hochschule Ingolstad, Germany
[2]Universität Tübingen, Germany
[3]Research Institute AImotion Bavaria, Ingolstadt, Germany
[4]Intel Labs, Munich, Germany
{kai.hartung, soeren.groettrup, munir.georges}@thi.de
gerhard.jaeger@uni-tuebingen.de

Abstract

In this study we address the question to what extent syntactic word-order traits of different languages have evolved under correlation and whether such dependencies can be found universally across all languages or restricted to specific language families. To do so, we use logistic Brownian Motion under a Bayesian framework to model the trait evolution for 768 languages from 34 language families. We test for trait correlations both in single families and universally over all families.

Separate models reveal no universal correlation patterns and Bayes Factor analysis of models over all covered families also strongly indicate lineage specific correlation patters instead of universal dependencies.

1 Introduction

Over the long time of their history, humans have come to develop a wide variety of languages. Overall, these languages share basic structural similarities. The nature and extent of these similarities have been subject of many theories. As language structure is mostly described in terms of syntax trees, these are often used as explanatory models for structural confines of possibly observable word orders in different languages. The explanations for such confines differ from Chomskian innate universal grammar (Chomsky, 1986), to more general aspects of computational ease of for the brain (Hawkins, 2009). On the other hand, increasing data about languages and computational means have prompted examinations of which structural universalities can be found empirically (Greenberg, 1963; Dryer, 1992; Dunn et al., 2011; Jäger, 2018).

More recent representatives of this start also to account for the historical evolution and relationships of the researched languages (Dunn et al., 2011; Jäger, 2018). By treating the evolution of languages analogous to biological evolution, they apply a method from bioinformatics (Pagel et al., 2004) to model historical transition rates between word-order traits. For Dunn et al (2011) the focus lay on four language families: Austronesian, Bantu, Indo-European and Uto-Aztecan. Across those they found widely differing trait correlations, proposing that correlations arise only specific to lineages and not from cognitive factors universally determining language evolution, but instead due to more local cultural evolution.

Jäger (2018) applied the models on a set of 34 families including models which cover all families at once comparing universal with lineage-specific correlations. This comparison resulted in a group of word-order traits being correlated universally. This suggests that phylogenetic models for individual language families cannot fully capture the universal correlations between different families.

In this work, we attempted to test this assumption with a different phylogenetic model. We used Brownian Motion to describe the evolutionary process of the trait change over time. This type of application is common in a biological phylogenetic context. It is described by, e.g., Harmon (2018). The Brownian Motion part of the model below is also based on this description. We then applied logistic linkage to model the observed categorical features of word order with a binomial distribution.

The paper is structured as follows. Section 2 describes the Brownian Motion models used. Section 3 will give a short overview of the data. The experimental setup is described in Section 4. Finally, we present the results in Section 5, followed by the conclusions in Section 6.

2 Brownian Motion Model

A Brownian Motion model is used in this study. It describes the process of evolutionary change of specified traits through time as a multivariate normal distribution. We denote it here as

$$x \sim MultiNormal(a, V) \tag{1}$$

with means a and Variance-Covariance matrix V, describing the distribution of the observed traits x.

Let V be the variance-covariance matrix which is computed of two matrices C and R. C encodes the degree of relation t between species such that languages with a longer shared history in a phylogenetic tree have a stronger covariance. We then define R as the correlation matrix. It is composed of the traits' evolutionary rates σ_n^2 and their correlations σ_{12}:

$$R = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \qquad (2)$$

V is computed by the Kronecker product as shown below on the example of a family with two traits and two species:

$$V = R \otimes C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes \begin{pmatrix} t_1 & t_{12} \\ t_{12} & t_2 \end{pmatrix}$$
$$= \begin{pmatrix} \sigma_1^2 \cdot t_1 & \sigma_{12} \cdot t_1 & \sigma_1^2 \cdot t_{12} & \sigma_{12} \cdot t_{12} \\ \sigma_{12} \cdot t_1 & \sigma_2^2 \cdot t_1 & \sigma_{12} \cdot t_{12} & \sigma_2^2 \cdot t_{12} \\ \sigma_1^2 \cdot t_{12} & \sigma_{12} \cdot t_{12} & \sigma_1^2 \cdot t_2 & \sigma_{12} \cdot t_2 \\ \sigma_{12} \cdot t_{12} & \sigma_2^2 \cdot t_{12} & \sigma_{12} \cdot t_2 & \sigma_2^2 \cdot t_2 \end{pmatrix}$$

The means a and the correlation R are parameters to be estimated. The traits x and phylogenetic matrix C are observed data as described in the following section.

Finally, logistic linking is introduced to model the categorical word-order traits x as a binomial distribution with probabilities p. Their logits, $logistic(p)$ are modelled as Brownian Motion:

$$x \sim Binomial(p) \qquad (3)$$
$$logistic(p) \sim MultiNormal(a, V) \qquad (4)$$

3 Word-order Traits & language families

The data used in this paper are of two kinds. The first are language features that are to be modeled. The second are phylogenetic data used as basis for the evolutionary model. Combined, these provide data for 768 languages from 34 language families as described in more detail by Wichman et al. (Wichman et al., 2016). All data were kindly provided by Gerhard Jäger(2018). We provide the data as part of this work[1].

We considered the same eight word-order traits (table 1) as in Dunn et al.(2011) and Jäger (2018).
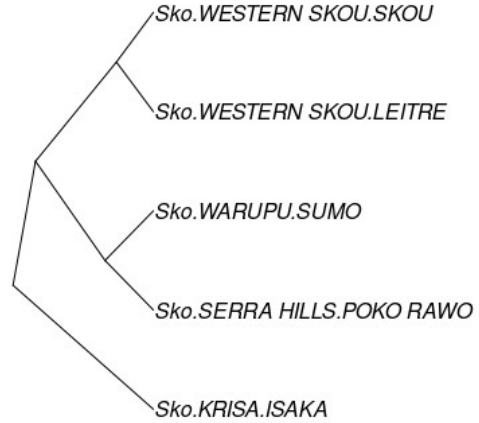
Figure 1: Example phylogenetic tree of the Sko language family.

These word order traits are based on the World Atlas of Languages (Dryer and Haspelmath, 2013), but their respective values are summarized into three possible values per trait. Each of these three values represents order configurations for the concerned syntactic elements. The first two classes represent the most dominant values. The third class summarizes all alternative configurations, appearing more rarely overall. For this work, we summarized the data as binary to be used in the binomial model. The distinction is only between a language having a value of the first majority class or not.

The phylogenetic data represent the languages' historical relations in the form of binary trees as for example in Figure 1. The trees' edge lengths represent the time since the last common ancestor split up. They have been estimated by Jäger (2018) and range from 10 to 1000 samples per family.

4 Experiments

In our experiments, we compared variations of this model with different base assumptions:

(A), we fitted two models to each of the 34 language families separately and pairwise to each of the eight traits. One version assumes strict independence of the traits by defining the trait correlations σ_{12} in R as constants with value 0. In the other version, the correlation between the characteristics is taken into account by keeping the values in R as parameters to be estimated. In this way, we could test whether we could find any correlation patterns by examining the families separately.

(B), we fitted three models for each trait pair, describing all language families together. Here

trait	0	1
AN	!adjective-noun	adjective-noun
PN	!postpositions	postpositions
ND	!demonstrative-noun	demonstrative-noun
NG	!genitive-noun	genitive-noun
NNum	!numeral-noun	numeral-noun
VO	!object-verb	object-verb
NRc	!relative clause-noun clause	relative clause-noun
VS	!subject-verb	subject-verb

Table 1: Word-order features based on WALS (Dryer and Haspelmath, 2013). The features are summarised as binary such that every feature is given either '1' for the listed value being present, or '0' for it being absent.

the first version assumes the same universal correlation R over all families. The second assumes distinct lineage-specific correlation R_F for each family. The last assumes strict universal independence for all families. The different versions were each compared with three different metrics. The first of these are Bayes Factors (Kass and Raftery, 1995), which allow a comparison with previous works (Dunn et al., 2011; Jäger, 2018). Bayes Factors above 10 indicate strong, values above 100 decisive evidence (Kass and Raftery, 1995). Further, we adopted 5 as minimum threshold for meaningful comparisons from Dunn et al. (2011). The Bayes Factors were obtained using bridgesampling (Gronau et al., 2020).

In addition, we used the information criteria WAIC (Watanabe and Opper, 2010) & LOOIC (Vehtari et al., 2016). These were added to test for consistency across different metrics. They are both based on pointwise log-likelihoods and can become unreliable, when dealing with strongly dependent data points (McElreath, 2020). Both express model comparisons primarily in differences where the value of the difference is not directly amenable to assess how strongly one model is favoured over another. To address this, we utilized the `rethinking` package by McElreath (2020). This offers a convenient function to compare models via WAIC including the assignment of weights ranging from 0 to 1 to give an accessible overview over the relative quality of the models.

The models were implemented in the Stan Modeling Language (2020) which uses the NUTS sampler (Hoffman et al., 2014) for parameter estimation. The code is available here[2] . Based on the suggestion of (Gronau et al., 2020) the models are

run for 21,000 iterations after warm-up, to achieve reliable Bayes factors. Distributed on 14 chains, this becomes 1.500 iterations per chain, in addition to 1,500 warm-up iterations each. The sampling was done with the default control parameters for Stan's sampler. With these parameters the lineage-specific models of the ND-NNum trait pair and the independent models of the NG-VO & AN-VO trait pairs did return a too small effective sample size for the bridgesampling to be effective. So these models were rerun with `max_treedepth` raised from 10 to 15.
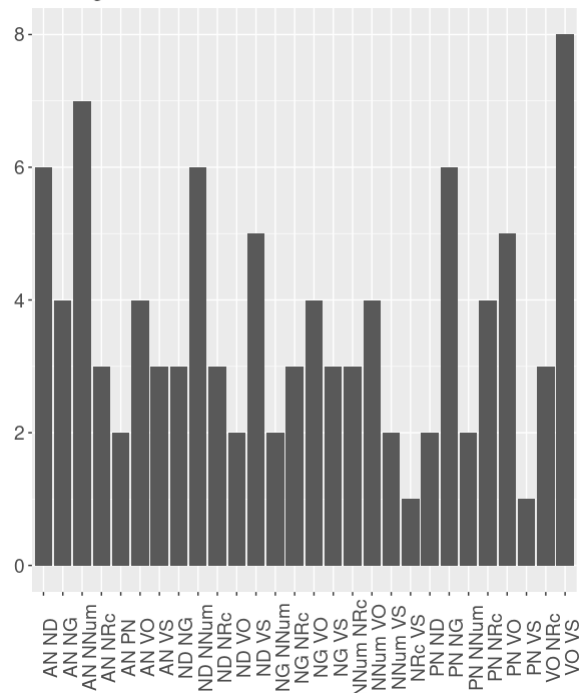

Figure 2: Numbers of families for which each trait pair was estimated to be correlated. Counts all cases for which Bayes Factors in favour of the dependent model have a value of at least five.

24

5 Results

For the single-family models, the comparisons did not yield Bayes factors above 5 for 15 of the 34 families. Therefore, no conclusions about preferred models are possible in those cases. Those trait correlations that did come up, only appeared in a minority of the covered families, not providing any pattern for universal correlations. Figure 2 shows for each trait pair in how many families it was estimated to be correlated.

The information criteria at face value find up to 22 families for which the same trait-pair is correlated. We filtered cases close to equivalence with WAIC weights to test how many cases have a robust preference. We chose the cutoff at 0.6. Filtering such cases results in less cross-family similarities, with no more than 3 families sharing the same trait pair as correlated. Cross-metric agreement: The percentage of model comparisons where Bayes Factors & WAIC weights favoured the same model or both don't favour any model at all is 85.5% with a LOOIC & WAIC agreement of 99.8%.

The model comparison between universal and lineage specific models strongly favour lineage specificity, with Bayes Factors favouring lineage-specific correlation for each trait pair (Figure 3).

In contrast to single family models, Bayes Factors and information criteria strongly disagree. Both WAIC & LOOIC strongly favour universal correlation models for each trait pair. WAIC weights are fully assigned to the universal models. Comparing lineage specificity with universal independence shows similar results. Bayes Factors strongly favour lineage-specificity but information criteria favour independence.

Universal trait pair correlations, based on the dependent-independent model comparisons center around the Adposition-Noun order (Figure 4).

6 Conclusions

We applied logistic Brownian Motion under a Bayesian framework to model the trait evolution for 768 languages from 34 language families. The models for single language families and those across all families indicate no universal word-order trait correlations across language families when compared using Bayes Factors.

This result matches with the observations made by Dunn et al.(2011), but contradicts the results over all families presented by Jäger (2018). Thus,
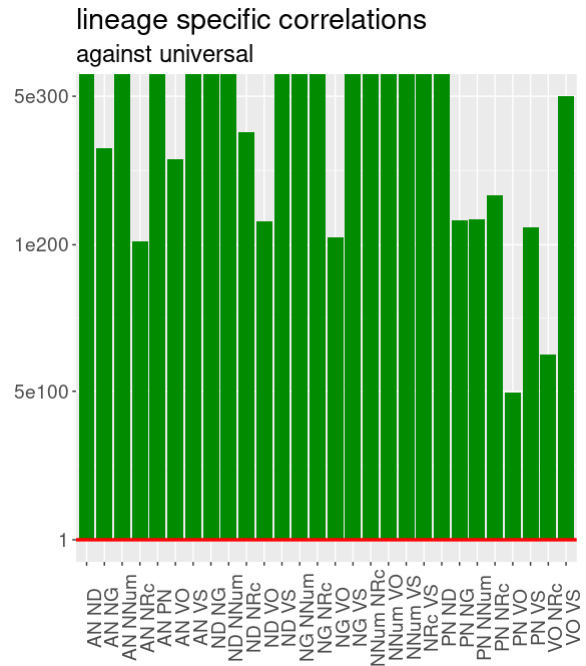

Figure 3: Model comparisons of lineage-specific vs universal correlations. The y-axis shows the strength of the Bayes Factors in favour of the models with lineage-specific correlations for each trait pair on the x-axis.
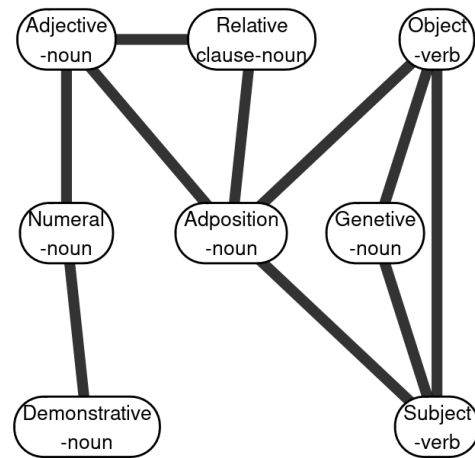

Figure 4: Universal trait correlations from directly comparing dependence and independence assumptions. For connected trait pairs, the dependence assumption is stronger in terms of Bayes Factors.

representing all families in one model did not add much information to single family models.

It is noteworthy, that the results from the information criteria contradict those from the Bayes Factors regarding the universal models. Although this could be attributed to some unreliability of the information criteria given the data and the nature of the model, it is still worth further investigation.

References

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Matthew S Dryer. 1992. The greenbergian word order correlations. *Language*, pages 81–138.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *In J. Greenberg, ed., Universals of Language. 73-113. Cambridge, MA*.

Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. 2020. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29.

Luke Harmon. 2018. *Phylogenetic comparative methods: learning from trees*. EcoEvoRxiv.

John A Hawkins. 2009. Language universals and the performance-grammar correspondence hypothesis. *Language universals*, pages 54–78.

Matthew D Hoffman, Andrew Gelman, et al. 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Gerhard Jäger. 2018. A bayesian test of the lineage-specificity of word order correlations. In *12th International Conference on Language Evolution (Evolang XII), Torun*.

Robert E Kass and Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

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

Stan Development Team. 2020. Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0. `https://mc-stan.org`.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2016. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Sumio Watanabe and Manfred Opper. 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Søren Wichman, Eric W. Holman, and Cecil H. Brown. 2016. The asjp database (version 17). `https://asjp.clld.org/`.