Laura Becker* and Matías Guzmán Naranjo

Replication and methodological robustness in quantitative typology

https://doi.org/10.1515/lingty-2023-0076
Received September 19, 2023; accepted October 11, 2024; published online February 25, 2025

Abstract: Replication and replicability are fundamental tools to ensure that research results can be verified by an independent third party, reproducing the original study and ideally finding similar results. Yet, replication has not played a very important role in language typology so far, with most of the discussion around replication concerned with different types of language samples and sampling methods. This study addresses the issue of replication in typology in a different way. We use the original datasets of four previous typological studies (Berg 2020; Dryer 2018; Seržant 2021; Shcherbakova et al. 2023) to show how statistical modeling can be used to test methodological robustness in typology. We do so employing advanced statistical bias controls, namely phylogenetic regression for genetic effects and a Gaussian Process for contact effects. While we could replicate some of the original results, parts of our findings differed from the original ones, revealing important methodological insights. Our comparisons show that more advanced statistical techniques that can model the phylogenetic and contact relations between languages pick up more complex patterns in the data than traditional sampling methods, and they capture more of the real relations between languages and their effects on linguistic structure.

Keywords: replication; methodological robustness; quantitative typology; statistical bias controls; phylogenetic regression; Gaussian process

1 Introduction

Replication and replicability are fundamental tools for ensuring that research results can be verified by an independent third party, reproducing the original study and ideally finding similar results. If so, then, more certainty can be attributed to the

***Corresponding author: Laura Becker [laura bɛkɐ]**, University of Freiburg, Freiburg, Germany, E-mail: laura.becker@linguistik.uni-freiburg.de. https://orcid.org/0000-0002-1835-9404
Matías Guzmán Naranjo [matias gusman naranxo], University of Freiburg, Freiburg, Germany. https://orcid.org/0000-0003-1136-6836

results due to cumulative evidence. Thus, replication serves the purpose of consolidating the findings, as they are arguably more robust when being reproduced.

Yet, replication has not played a very important role in language typology so far, with most of the discussion around replication concerned with different types of language samples and sampling methods (e.g. Dryer 1989; Haspelmath and Siegmund 2006; Maddieson 2006; Widmann and Bakker 2006). This study addresses the issue of replication in typology in a different way. We use the original datasets of four previous studies and use different methods from the original studies to test to what extent the results are method-dependent or robust across different methods used for analysis. In this study, we use statistical modeling for our analyses, showing how such techniques can be used to test the replicability of typological studies. The fact that there is no single best (statistical) approach to analyzing typological data is, we find, still under-appreciated in typology, and results are not independent of the methods used for data analysis. The objective of this paper is to raise awareness that the statistical tools chosen for analysis matter, that they require transparency and scrutiny as does the data and the annotation process, and that applying new methods to old data is a useful and necessary process to consolidate typological findings. We selected the following four test cases: Dryer (2018) on the order of elements in the noun phrase, Seržant (2021) on contact effects in Slavic morphosyntax, Shcherbakova et al. (2023) on the complexity trade-off hypothesis between nominal and verbal grammatical marking, and Berg (2020)[1] on the association between gender marking on nouns and different types of pronouns.[2] There is no specific reason for choosing these papers other than the fact that the authors made their datasets available. For full disclosure, we did not know whether our results would consolidate or call into question the original findings beforehand.

Since the purpose of this paper is to gauge the effect of the particular method used for analysis on the results, we use the original data without additional modifications for all four case studies. Therefore, we will not be concerned with questions regarding the particular choices made by the authors in the data collection and annotation for the original studies. Our purpose is not to contest the linguistic work of the papers in question, but simply to check the original results against a different statistical technique. More specifically, we will follow Guzmán Naranjo and Becker (2022) and Verkerk and Di Garbo (2022) in using phylogenetic regression to control for genetic effects and a Gaussian Process to control for contact and areal

1 For reasons of space, our replication of Berg (2020) is discussed in Appendix C. Appendices A-D can be found as a Supplementary File in the online version of this paper, and also as "replication_appendix.pdf" in the Supplementary Materials at https://osf.io/9b2zk/.

2 In fact, Dryer (2018) replicates Greenberg's Universals 20, and Berg's study is a replication (conceptual and in terms of sampling) of Greenberg's Universal 43.

effects (cf. Section 3). As we will show in Sections 4, 5, and 6 for three test cases, some findings are robust and can be corroborated with our methods, while others cannot be confirmed. This underlines how important it is to be aware of statistical methods having an impact on the results as well; they need to be chosen with as much care as the linguistic choices concerning the dataset and annotation, and they need to be reported with transparency to allow for evaluation and replication. We discuss this in more detail in Section 7.

2 Replication in typology

2.1 Defining the relevant notions

Different notions have been used around the issues of replication and replicability. A proper overview would go beyond the purposes of this paper. We will therefore only introduce the notions as they are used in remainder of this study.[3] Before turning to replication and replicability, we need to clarify what we mean by robust findings. We will define robustness following Goodman et al. (2016) as shown in (1). Applied to typology, robust findings then need to hold across (i) different language samples, (ii) alternative linguistic analyses and annotations as well across (iii) different statistical methods.

(1) **Robustness**
Robustness refers to the stability of experimental conclusions to variations in either baseline assumptions or experimental procedures. (Goodman et al. 2016: 4)

The second essential notion for this paper is that of replication. Replication can be understood in many different, or rather more or less strict ways. We define replication in a broader sense, loosely adapting the definition of Schmidt (2009: 91), including the idea of uncertainty (cf. Gelman 2018; Vasishth and Gelman 2021):

(2) **Replication**
Replication is a methodological tool based on a repetition procedure that is involved in assessing or reducing the amount of uncertainty regarding previous research results. In doing so, it can be used to establish a piece of knowledge of our world.

3 For more details on different types and uses of replication and replicability, cf. Gawne and Berez-Kroeker (e.g. 2018); Goodman et al. (2016); Machery (2020) and references therein.

Note that our definition of replication does not rely on the outcome of the replication study. Whether or not it confirms earlier results is irrelevant for its classification as a replication study in this sense.[4] Repetition can establish knowledge because it can establish stability, i.e. robustness in case the original results can be confirmed (cf. Schmidt 2009). In case repetition does not confirm earlier results, it leads to a justified increase in uncertainty regarding those earlier results and reveals the need for further research to arrive at more conclusive results. An exact replication of a previous study means that the data and annotation as well as the analysis are identical to the original ones. While hardly carried out in practice besides as part of reviewing, exact replications are highly important theoretically and correspond to the minimal requirements of replicability of an empirical study. We define replicability as follows:

(3) **Replicability**[5]
Replicability corresponds to the potential of exact replication. It guarantees that another independent scientist can use the same data and follow the same procedure as in the original study, obtaining the same results.

Replicability thus makes research results independently verifiable and ensures credibility. It has long been recognized as a research standard across different research disciplines (e.g. Donoho 2010; Gelman 2018; Goodman et al. 2016; Schmidt 2009) and has become a more prominent issue in linguistics as well (e.g. Aguilar-Sánchez 2014; Berez-Kroeker et al. 2018; Bisang 2011; Gawne and Berez-Kroeker 2018; Grieve 2021; Harris et al. 2006; Himmelmann 1998; Kobrock and Roettger 2023; Maxwell 2012). We will return to the issue of replicability in the discussion in Section 7.1.

2.2 Replication in typology: status quo

In typology, replication has mostly been carried out in that a research question of a previous study has been re-addressed with a different sample and/or different linguistic definitions and annotation choices.[6] A number of typological studies fall into

4 We use the term 'confirm' to mean that the results of the replication study are in agreement with those of the original study. Of course, this does not imply that the results are necessarily true or correct; they can in principle both be erroneous.

5 Replicability is also referred to as reproducibility in the literature; we regard the two terms as interchangeable and use "replicability" for consistency with the term "replication".

6 Broadly speaking, the method of triangulation also falls within approaches to replication in typology. Triangulation refers to the combination of different empirical approaches to study the same phenomenon in order to test how robust results are across methods and to, ideally, find converging evidence. In typology, triangulation often combines cross-linguistic generalization with corpus studies and/or artificial language learning experiments (e.g. Levshina 2022; Martin et al. 2019; Saldana et al. 2021; Tal et al. 2022).

this category. One example of topics or questions that have been revisited in a number of papers throughout the years is word order universals (e.g. Donohue 2011; Dryer 1992, 2011; Siewierska and Bakker 1996; Song 2012; Steele 1978; Tomlin 1986). These studies do not necessarily make the replication element explicit and they do not test for methodological robustness, which is why they are less relevant for the purposes of the present study.

Sparked by a side note discussion in Corbett (2005), replication in typology became an explicit topic of debate in a 2006 issue of *Linguistic Typology*. The 2006 discussion mainly centered around the question of how exactly replication and reproduction can and should be understood in typology, i.e. at which levels of research is replication useful and desirable. In this vein, Haspelmath and Siegmund (2006: 74) make a more concrete proposal as to how replication can apply to typological work. Updating their classification of four levels allows us to distinguish five levels of replicability in typology as shown in (4).

(4) Levels of replicability in typology
   a. replicability of the **primary data collection**
   b. replicability of the **grammatical description**
   c. replicability of the **linguistic analysis & annotation**
   d. replicability of the typological generalization based on **different samples**
   e. replicability of the analysis based on **different (statistical) methods**

Levels (a) and (b) relate to the primary data collection and language documentation itself. Since our focus is on quantitative typological studies that usually do not involve primary data collection, we will not discuss replicability of levels (a) and (b) further.[7] Level (c) involves the coding of the linguistic phenomena at hand; this includes the theoretical definitions and choices as well as the categorization of the phenomena under investigation. We are only aware of one study that explicitly tests for replicability across different ways of categorizing the data, namely Nichols et al. (2006). The authors show that their findings on the distribution of morphological complexity remain similar when using inflectional, derivational as well as lexical inflectional metrics.

Level (d) tests the generalizability of the results, using the same linguistic categorizations and methods, to new data. An early example of a replication study that tackles this issue is Dryer (1989), where he shows that the results in Nichols (1986) concerning head marking orders were biased by the sample used. Using a

7 Nevertheless, we acknowledge that data robustness on those two levels is crucial for any typological study and we refer the reader to discussions of data transparency, replicability and robustness in the language documentation literature (e.g. Gawne and Berez-Kroeker 2018; Himmelmann 1998).

more balanced sample which took contact and areal distributions into account, Dryer (1989) produced completely different results. That replication in typology most importantly consists of verifying previous findings with new language samples is also reflected by the contributions to the *Linguistic Typology* debate on replication in 2006. The issue includes four empirical studies; three out of those studies focus on varying the language sample in order to subject previous findings to replication (Haspelmath and Siegmund 2006; Maddieson 2006; Widmann and Bakker 2006). Maddieson (2006) not only uses different convenience samples but tests previous findings with areal as well as random sub-samples of the original dataset.

Level (e) comes in at the highest level, checking to what extent the findings are robust when the same sample with the same linguistic annotations is analyzed with different methods. As will be discussed in Section 2.3, not many quantitative studies in typology evaluate the impact that the methods used can have on the results obtained. The purpose of the present study thus is to draw attention to this, i.e. that the methods used for analysis influence the results, by replicating previous studies to test how method-dependent or robust the results really are.

2.3 Replication for methodological robustness in typology

There is no single, objectively adequate solution to model a typological phenomenon, but building a model (statistical or not) necessarily involves a number of different choices that have to be motivated and that can influence the results. So far, not much work has focused on replicating typological studies using the same data but applying a new statistical method.

There are only a handful of notable exceptions to this gap in the literature, and the original study tends to make (strong) conclusions that do not fit in with the general theoretical expectations in the field, e.g. Atkinson (2011) and Chen (2013). Atkinson (2011) reports a world-wide decline in phonemic diversity from Africa, arguing that those findings support a global serial founder effect with Africa as the point of origin. Chen (2013) finds an association between the obligatory use of grammatical future tense and savings behavior of individuals. In both cases, the replication studies (Jaeger et al. 2011; Roberts et al. 2015; Van Tuyl and Pereltsvaig 2012) reveal that the effects found in the original studies disappear with more rigorous statistical bias controls for family and areal effects, calling into question the original conclusions.

Two other studies that have been replicated for methodological robustness establish an association between an environmental factor and a linguistic property. Everett (2017) reports a relation between ambient humidity and the vowel-consonant ratio, concluding that languages in drier climates use fewer vowels. Similarly, Maddieson (2018) finds that languages spoken in areas with higher temperatures tend to have higher sonority scores. Hartmann (2022) carries out a replication of both studies, using the original data and more sophisticated bias controls. As in the

replication studies mentioned above, Hartmann (2022) finds that using more careful statistical controls for potential family and areal biases greatly reduces the effects found in both original studies. He therefore concludes that the original findings could not be replicated.

Another example of replication in quantitative typology is Schmidtke-Bode and Levshina (2018), who replicate Bickel et al. (2015) on differential case marking and so-called "scale effects". Simply put, scale effects refer to cross-linguistic tendencies of objects being unmarked or case marked depending on some of their inherent and contextual properties such as animacy, referential status and discourse-pragmatic prominence. Applying the Family Bias method to a cross-linguistic sample, Bickel et al. (2015) do not find strong universal evidence for such scale effects with differential case marking. In their replication study, Schmidtke-Bode and Levshina (2018) use the same data but analyze them with mixed effect regression. In contrast to the original study, Schmidtke-Bode and Levshina (2018) do find robust cross-linguistic support for universal scale effects.

We are only aware of one replication study for methodological robustness which could confirm the original findings. Everett et al. (2016), replicating Everett et al. (2015), find phonemic tones to be more likely to develop in warmer climates than in colder or desiccated ones. Everett et al. (2016) react to methodological criticism from Hammarström (2016), adjust their statistical model and replicate their original results.

While some of those studies have received much criticism from the linguistic community, they have to be given credit from a data transparency point of view, for making all data and code publicly available. This should of course be the standard for typological studies, but many studies do not publish the full dataset and code. Without this, evaluating and replicating their theoretical decisions and methodology would not have been possible, and we would have missed a constructive theoretical and methodological discussion in quantitative typology.

3 The current approach

3.1 Evaluating methodological robustness

As mentioned in Section 2, an important but largely ignored function of replication is the evaluation of the methodological robustness of the statistical methods. Roberts (2018) notes that "if the same core components cause the same result across a range of alternative models, then the results are robustly due to those core components." We adapt this idea in the present study by evaluating how robust effects in the data are when using a different statistical approach for analysis. Crucially, we use the same dataset, i.e. sample and annotation, as in the original study. This leads to a controlled environment where we can test how much the results depend on the analysis alone, having eliminated variation across samples and annotation decisions. If the results

of the previous studies can be replicated when using more advanced statistical techniques, we can be somewhat more confident about the effects found in the original studies. If our replications lead to different results, we should interpret the original results as less certain.

Importantly, this does not require the original study to make use of statistical tests. A typological study based on a language sample and annotation of some linguistic feature can (in part) be quantitative in that it minimally counts the occurrence of different values of that feature to assess their distributions. By now, there are a multitude of different statistical methods that have been proposed for typological work, from simple chi-square tests (see Dryer 1992 for an early example), to mixed effect models (e.g. Jaeger et al. 2011), and more recently the use of phylogenetic regression (Verkerk and Di Garbo 2022) and Gaussian Processes for areal controls (Guzmán Naranjo and Becker 2022), as well as other types of phylogenetic models (Jäger and Wahle 2021). Despite the abundance of different available techniques, there is very little work comparing how robust results are across these techniques when applied to the same datasets. This is, however, crucial if we want to assess how confident we can be about previous findings. We think that this applies especially to the field of typology, where bias control, e.g. for phylogenetic and contact effects, has traditionally been done manually in the sampling process itself. The analysis of the data then often no longer includes any statistical methods to control for sampling biases. Studies which make use of statistical modeling do not necessarily control for biases in the sampling process, but use convenience samples instead and build bias control into the statistical modeling. Against this background, it is important to evaluate whether typological results are robust across those two fundamentally different families of approaches.

3.2 Statistical bias control

This section gives a brief overview of the statistical methods that we use to control for phylogenetic and contact bias. For a more detailed description of these techniques, we point the reader to Verkerk and Di Garbo (2022), Guzmán Naranjo and Becker (2022) and Guzmán Naranjo and Mertner (2023), as well as the Supplementary Materials for the concrete computational implementations. The models of this study were coded using Stan (Carpenter et al. 2017) and in some cases also the brms package (Bürkner 2017) in R 4.3.0 (R Core Team 2023).[8] We will discuss more details regarding the models used for each of the three case studies in Sections 4.2, 5.3 and 6.3 respectively.

8 The Supplementary Materials can be found at https://osf.io/9b2zk/?view_only=c3c021ddaa9749c9a88af86109d331b0. In all cases, we fitted our models ensuring that there were no divergent transitions, our effective sample size was large enough, and that Rhat values were close to 1. In all models, we used (weakly) informative priors. We found that the estimates of interest were very robust to

3.2.1 Phylogenetic regression

To control for phylogenetic effects we make use of a method called phylogenetic regression.[9] The idea of phylogenetic regression is that we want to control for the whole structure of the phylogenetic tree, i.e., languages which are closer to each other in the tree are expected to be more similar due to shared inheritance. To model this idea, we add intercepts for each language but we force the estimates of the intercepts to be correlated according to the structure of the tree. If two languages are close to each other in the tree, their estimates will be very close to each other, and two languages on completely different branches of the tree can be as different as they need to. This way of modeling family relations is more flexible than adding intercepts per family or genus (see Jaeger et al. 2011 for an example of this approach), as it does not represent relatedness between languages in a categorical way. Instead, it captures relatedness in a gradual way in that the intercepts of languages that are more closely related are forced to be more correlated than the intercepts of languages which are less closely related.

Although still being a relatively new technique in typology, adding a phylogenetic term has been shown to be an effective control in several studies (Bentz et al. 2015; Guzmán Naranjo and Becker 2022; Verkerk and Di Garbo 2022). It could be shown to be able to deal with bias resulting from multiple related languages in a sample. Here, we follow Guzmán Naranjo and Becker (2022) and use the Glottolog tree (Hammarström et al. 2022).[10] To estimate the relative time depth of each node, we assume that genetic relatedness is proportional to the number of shared nodes in a tree. While there are alternative methods for estimating relative time depth like calculating lexical distances with lexical datasets as proposed by Wichmann and Rama (2021), our experience is that these methods do not improve model performance but increase modeling complexity. Note that we do not actually need the real time depth of the trees; we only require a useful approximation of the relations between languages within families.

prior selection, and we illustrate this with two prior sensitivity analyses for the Shcherbakova et al. (2023) and Seržant (2021) studies (found in the corresponding scripts in the Supplementary Materials) https://osf.io/9b2zk/.

9 An exhaustive mathematical description of phylogenetic effects can be found in de Villemereuil and Nakagawa (2014).

10 However, in this study, we do not make use of micro-families as in Guzmán Naranjo and Becker (2022). Also, we use the phylogenetic trees from the original study in our replication of Shcherbakova et al. (2023) instead of the Glottolog tree.

3.2.2 Gaussian Process

Traditional bias control in the sampling process has focused much more on phylogenetic dependencies, and areal control has usually consisted of limiting the number of languages in the sample per macroarea (or other comparable areas).[11] In this study, we use a Gaussian Process (GP) to control for contact bias.[12] A GP uses a distance matrix between the observations in the dataset to estimate the spatial covariance of the observations. In a GP, two observations which are located closely together can have a strong influence on each other, with the strength of the influence between observations decaying non-linearly with increasing distance. Crucially, this decay follows a Gaussian curve, meaning that it has a non-linear structure. Therefore, the strength of influence quickly drops to zero for observations which are further apart. In this paper we use Euclidean distance between languages, using the coordinate data (latitude and longitude) of the language's location from Glottolog (Hammarström et al. 2022). This is more of a practical choice for now, constrained by the spatial information available for a large number of languages. In principle, a GP can be used with other distance metrics as well that capture the spatial properties of languages in a more realistic way.[13] For other examples of GPs used to control for spatial effects in typological studies see Guzmán Naranjo and Becker (2022) and Guzmán Naranjo and Mertner (2023).

Thus, the advantage of using a GP to model areal or contact effects is that languages in contact can be included and that this information can be used by the model to estimate how much of the variation contact accounts for. Moreover, it accommodates contact effects as non-linear, reflecting that distance between languages has different effects depending on the linguistic density of the area.

4 Case study: Dryer (2018) on the order of elements in the noun phrase

4.1 Overview and results of the original study

Dryer (2018) surveys different word orders of elements in the nominal domain with a large typological sample. The elements examined are the demonstrative (DEM),

11 There are a few other, more principled, approaches to control for contact bias. See the overview in Guzmán Naranjo and Becker (2022: 22–26) for more details.

12 For a discussion of the mathematics behind GPs, see Rasmussen (2004); Williams and Rasmussen (2006).

13 As of now, the current approach is the most realistic statistical approach to areal and contact control on a global scale.

numeral (NUM), adjective (A) and noun (N). Two examples to illustrate different word orders in the nominal domain are given in (5) and (6). In (5), the nominal expression has an initial noun with all additional elements following it (N-A-DEM-NUM). Example (6) shows the opposite, with all modifying elements preceding the noun (DEM–A–NUM–N).

(5) Akha (Sino-Tibetan, akha1245) (Dryer 2018: 800)

tshɔ́hà	*jɔmɔ̀*	*xhǿ*	*njì*	*ɣà*
person.N	good.A	`those.DEM	two.NUM	CLF

'those two good persons'

(6) Dhivehi (Indo-European, dhiv1236) (Dryer 2018: 800)

mi	*ranngaḷu*	*tin*	*fot*
this.DEM	good.A	three.NUM	book.N

'these three good books'

Dryer (2018) provides a dataset with 1,096 languages, but out of these only 593 are coded for word order.[14] There are 24 logically possible orders between demonstrative, numeral, adjective and noun. Out of those, 18 are attested in Dryer's data.[15] The distribution of attested word orders is shown in Figure 1. Certain areal patterns start to become apparent from Figure 1 already, such as the order DEM–NUM–A–N being predominant in Eurasia and the order DEM–NUM–N–A mostly being found in the Americas. Dryer's paper is particularly insightful for the purposes of this study because it proposes a new sampling technique to control for genetic and areal bias in order to estimate a so-called "adjusted frequency". In other words, he aims at estimating the expected frequency of each possible word order after controlling for genetic and areal correlations. The present replication study will focus on this adjusted frequency count.

4.2 Models of the replication study

It is common in typology to use regression to examine whether some (set of) variable(s) is a good predictor of another variable (e.g. Guzmán Naranjo and Becker 2022; Jaeger et al. 2011; Sinnemäki 2020). However, regression models can also be used to estimate the expected proportions of the values of a single linguistic feature, which provides insights into how common a given value is across languages. In this case, we can use a categorical regression model to estimate the proportions of the 18 different word orders in the world's languages based on Dryer's sample. The resulting

14 After applying his sampling method, Dryer (2018) analyzes the data of 576 languages.

15 Table 1 in Appendix A summarizes the original results, showing the number of languages, genera, and the adjusted frequencies of all the 24 orders.

Figure 1: Distribution of word orders based on Dryer's (2018) sample.

estimates thus correspond to the expected global proportions after having controlled for areal and genetic correlations.

Thus, for the first case study, we fitted a categorical model with a phylogenetic term and a Gaussian Process (using the languages' latitude and longitude information as predictors).[16] In order to compare the methodological robustness of the results in an even more detailed way, we fitted the following four models:

1. `m_base`: no controls, intercept only
2. `m_gp`: contact effects
3. `m_phylo`: phylogenetic effects
4. `m_gp+phylo`: contact and phylogenetic effects

4.3 Results of the replication study

The results of the first replication study are given in Figure 2. In addition to the proportions of word orders estimated by the four models, Figure 2 shows the observed proportion of each word order (green) and the adjusted frequency as calculated by Dryer (2018) as a proportion (black). The estimates of `m_base` can be seen in beige, the ones of `m_phylo` in light blue, the estimates of `m_gp` in red, and the ones of `m_gp+phylo` in dark blue. All model estimates additionally include 50 % (bold) and 95 % (light) uncertainty intervals. This means that, given the data and the model, we can be 50 % or 95 % certain that the proportion of a given word order will fall in that interval.[17] There are four important observations that we can take from Figure 2.[18] First, the estimates of `m_base` are essentially the same as the observed values, although including some uncertainty. This works as a sanity check that the model is performing as expected.

Second, for the most part, `m_phylo` does not seem to produce estimates that differ substantially from those of `m_base` and the observed values. This suggests that there is actually not much of a phylogenetic bias in the sample to begin with. This observation is further supported by the contrast between the estimates of `m_phylo+gp` and `m_gp`, which are, for all word orders, very close to each other and mostly overlapping.

16 See the Supplementary Materials for the implementation.

17 We focus on the 50 % uncertainty intervals because there is too much uncertainty in the estimates at larger intervals. This does not mean that the model is performing poorly; rather, it means that we cannot reach strong conclusions about the likely value of the expected proportions.

18 Figure 2 shows so-called equal-tailed intervals. That means that the probability of being above the interval is the same as being below it. This is not the only possible way of constructing intervals, with a common alternative being the High Density Posterior Interval (HDPI), which is the narrowest interval that contains a given percentage of the posterior. We have included the equivalent HDPI of Figure 2 in the Supplementary Materials under `dryer/code/plots/expected-value-hdpi.pdf`.

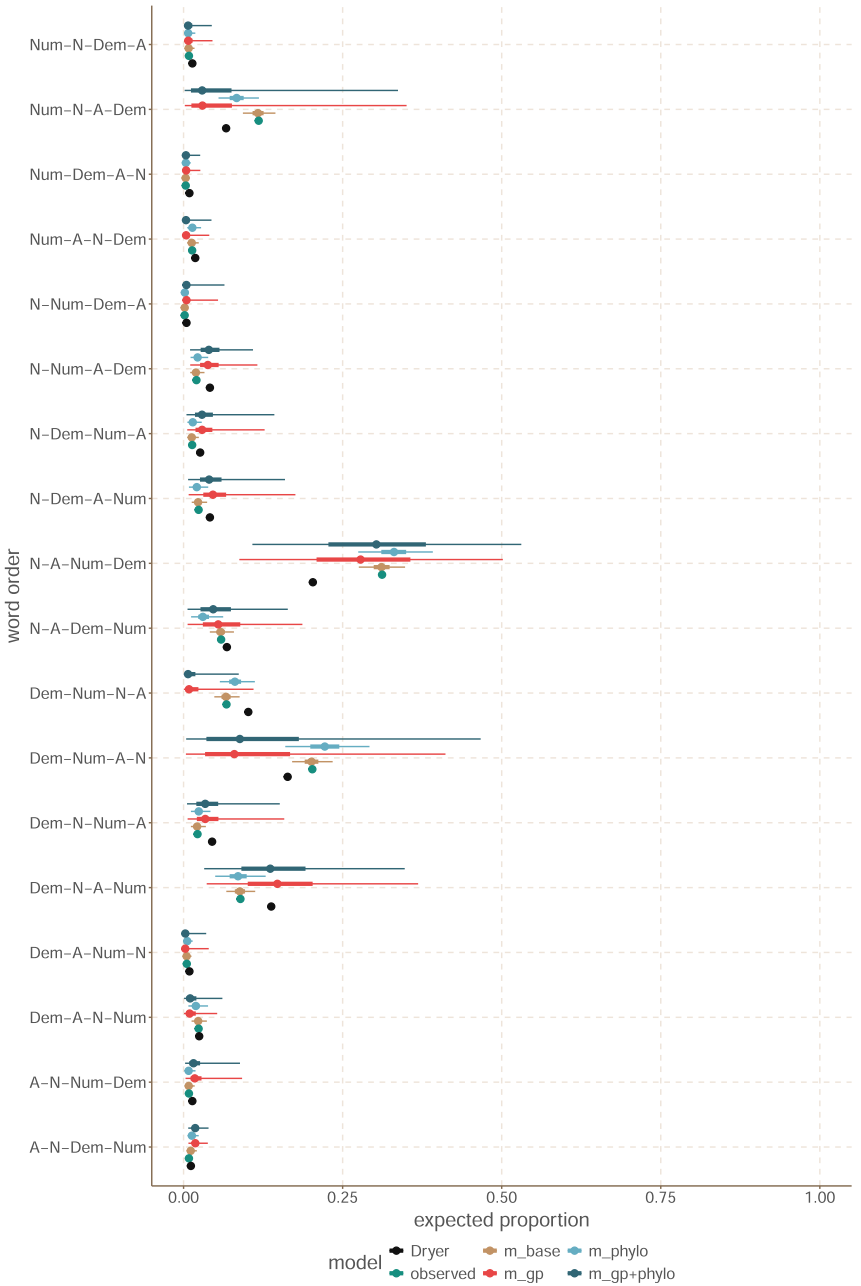

Figure 2: Original and replication results.

Third, if we compare the estimates of the full model m_phylo+gp with the observed proportions, we see that some observed proportions seem to be slightly biased, particularly for the orders DEM–NUM–A–N, NUM–N–A–DEM and DEM–N–A–NUM. In the first two cases, the observed proportion is substantially higher than what is estimated by m_phylo+gp with controls for phylogenetic and contact biases. The adjusted frequencies as calculated by Dryer (2018) are slightly more conservative than the observed proportions but are also fairly high compared to the mean estimates of m_phylo+gp. In those two cases, the observed proportions likely overestimate the actual proportions, as the estimates produced by m_phylo+gp are smaller. For DEM–N–A–NUM, the model and Dryer agree that the observed proportion likely underestimates the actual proportions. Other orders such as N–NUM–A–DEM, N–DEM–NUM–A and N–DEM–A–NUM also seem to be biased in the observed counts, although rather by a small if not negligible amount.

Lastly, the results in Figure 2 show that uncertainty intervals are generally larger with the two models that include a GP, m_gp and m_phylo+gp, than with the other two models m_base and m_phylo. This is especially the case for those orders that are more frequent in the sample and that have higher observed proportions. It is important to note that larger uncertainty intervals have nothing to do with model performance, i.e. how much of the variance a model can explain. Larger uncertainty intervals do not mean that the models are "worse" than the ones with smaller uncertainty intervals.[19] Instead, what the larger uncertainty intervals of m_gp and m_phylo+gp reflect is that the GP (which models contact effects) can account for some of the variation observed in the data. As a result, the model attaches a higher level of uncertainty to the expected proportion, which is supposed to represent the real proportion of a word order in the world's languages. Thus, given the data, the model cannot be very certain what the real proportions of those word orders are, as much of their occurrences can be accounted for by contact and not by an independent, general preference. Conversely, the models that do not include the GP are overly confident about the expected proportion of a given word order. Without any contact information, they ignore that the distribution of orders could be due to a different factor than the observed proportions (and phylogenetic effects in the case of m_phylo) and thus allow for more confidence regarding the expected proportion of an order. It is crucial to understand that this confidence is not a good thing, because this model does not represent the reality very well. The more complex model m_phylo+gp shows that contact effects play an important role in accounting for the

19 In fact, m_phylo+gp is the best model in terms of performance, meaning it can account for most of the variation in the observed orders. We tested model performance by approximate leave-one-out cross-validation. See the Supplementary Materials for the result of the model comparisons, and see Section 5 for a more detailed description of the technique used.

variation of word orders, and that the current sample is simply not sufficient to make more certain predictions about the real proportions, once we control for the variation due to contact and areal effects.

We now turn to comparing the method of adjusted frequencies from Dryer (2018) to the results of our full model m_phylo+pg. It is impressive that Dryer's results often coincide with the estimates of the model or fall within its 50 % uncertainty interval for most orders. For the orders of NUM–N–A–DEM and DEM–NUM–A–N, Dryer's results are somewhat further away from our point estimates. Still, his method of adjusted frequency corrects in the same direction as the m_phylo+gp estimate from the observed proportions.

Although adjusted in the same direction, Dryer's adjusted proportion of 0.2 for N–A–NUM–DEM is substantially lower than our mean estimate of 0.3, and it lies outside of the 50 % uncertainty interval. Importantly, this order is the most frequent one observed, and it is very common in three areas that also have a high linguistic density in the sample: West Africa around the Gulf of Guinea, Mainland South East Asia and Melanesia. This can be seen in Figure 1, where the order of N–A–NUM–DEM is coded by "J". The areal distribution of this order is likely the reason for the adjusted proportion of Dryer being much lower than the estimate of m_phylo+gp. To control for phylogenetic and areal biases, Dryer applied his sampling method to each word order separately (see footnote 8 in Dryer 2018 for an explanation). Dryer's (2018) method thus had to exclude most of the datapoints in those three areas for the N–A–NUM–DEM order due to their geographical closeness. This then led to a much lower adjusted proportion of this word order than the estimate of the model that takes into account the other word orders attested in this area. Dryer (2018) is aware of this "ceiling effect" of his method and briefly puts it into context in footnote 8 (Dryer 2018: 803). While this methodological choice may be justified in the context of his particular study, it needs to be highlighted for potential future studies that may apply Dryer's method to a context in which investigating the actual proportions is part of the research question.

Returning to the results in Figure 2, the order DEM–NUM–N–A is the only case in which Dryer corrects in the opposite direction than m_phylo+gp. In this case, m_phylo+gp concludes that the observed value over-represents this word order, while Dryer's method assumes that it under-represents it. For a better understanding of this discrepancy, we can look at Figure 3, which shows the distribution of DEM-NUM-N-A as opposed to all other orders.

Figure 3 strongly suggests that the order DEM-NUM-N-A is localized around certain areas, namely Western Europe, Eastern Turkey, Amazonia, Central Mexico, and potentially North America more broadly. Dryer's method does not seem to pick up on this areality, and therefore corrects the proportion by increasing it. The models with a GP (m_gp and m_phylo+gp), however, assign a large portion of the variance to this areal pattern and thus estimate the remaining expected proportion to be lower than the observed one. In fact, the expected proportion based on those two models for the DEM–NUM–N–A order is close to 0. We can interpret this as the DEM–NUM–N–A order being

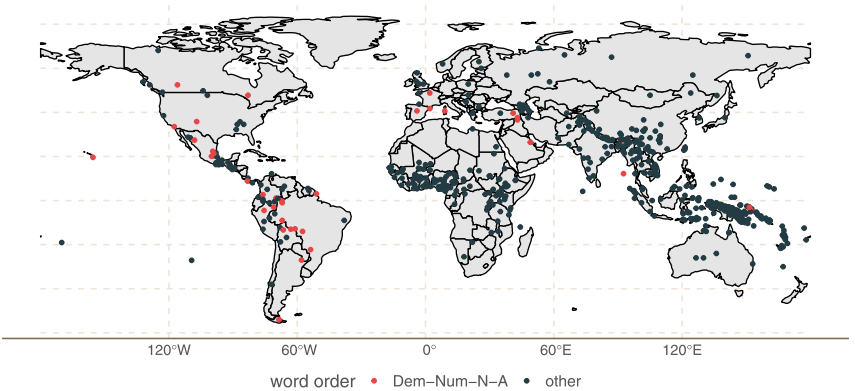

Figure 3: Distribution of DEM–NUM–N–A versus other word orders.

extremely rare cross-linguistically if it were not for the spread by contact in a few selected regions.

4.4 Taking stock

We have seen that Dryer's (2018) approach to calculating adjusted frequencies results in very similar estimates to the full model `m_phylo+gp` in our replication study. There does seem to be a small amount of disagreement between the two methods especially when areal patterns are involved. Based on this example, the statistical model seems to be more able to deal with such cases than Dryer's sampling method, but it is difficult to say with certainty which method is better for this particular case. Overall, it is good that we find much agreement across different techniques. This means that we can be more confident about the expected proportions of the different word orders in the nominal domain.

5 Case study: Seržant (2021) on contact effects in Slavic morphosyntax

5.1 Overview of the original study

Seržant (2021) examines the factors that contribute to the innovation and retention of grammatical properties over the course of time. Specifically, he examines the role of contact and areality in Slavic on (i) the retention of Proto-Indo-European person-number indexes and (ii) the innovation of the partitive markers. We will discuss the first part only.

Seržant (2021) focuses on six verbal person-number indexes (1SG, 2SG, 3SG, 1PL, 2PL, 3PL) in Indo-European, Tibeto-Burman, Turkic, Uralic, Dravidian and Semitic.[20] In total, his sample includes 150 languages from those six families. To examine the role of contact and areality on the development of person-number indexes, Seržant (2021) studies the distribution of what he introduces as the "verbal paradigm decay factor". The decay factor is a metric that measures to what extent the contrasts present in the proto-language are preserved in its modern descendants. The decay factor is a number bounded between 0 and 1, with a decay of 0 meaning that the original paradigm is preserved in its entirety, whereas a decay factor of 1 corresponds to the total loss of the original person-number indexes.

To accommodate the various transition stages in between these two extremes, Seržant proposes three indicators with their own measure to calculate the decay factor. He takes a paradigm to decay if (i) there is reduction in the number of segments of markers which have a morphological impact on the paradigm, (ii) the contrast between two cells is lost (i.e. when syncretism emerges), or if (iii) markers that are phonetically zero develop. Seržant operationalizes these three indicator metrics as follows. For (i), the decay is calculated as the number of segments in the modern paradigm divided by the number of segments in the paradigm of the proto-language. The decay for (ii) is given by the number of syncretisms minus the total number of potential syncretism. For (iii), the decay is measured as the total number of cells with zero markers. The total decay factor of a language is then calculated as the mean of the normalized metrics for (i), (ii) and (iii).

5.2 Original results

One of the main conclusions drawn by Seržant (2021) is that there is an East-West cline in terms of the decay in the verbal paradigms of several language families in Eurasia. Regarding Slavic languages, he concludes that the ones spoken in closer proximity to Uralic languages retain more of their original paradigms than languages spoken further to the west. This result is mainly based on visual inspection of the distribution of decay factors on the map (Figure 1 in Seržant 2021: 72). Figure 4 reproduces this map (including the adjusted location information, cf. Section 5.3), where we can see the decay factor for all 150 languages in the sample. A high decay factor is shown in orange and red, a low decay factor in blue and violet.

20 Table 2 in Appendix B illustrates the data person-number indexes from Indo-European languages together with their decay factors. Seržant (2021) generally represents each language by one set of person-number indexes. This set corresponds to the markers used with the present tense, except for Semitic, for which the imperfective indexes are used (Seržant 2021: 68).

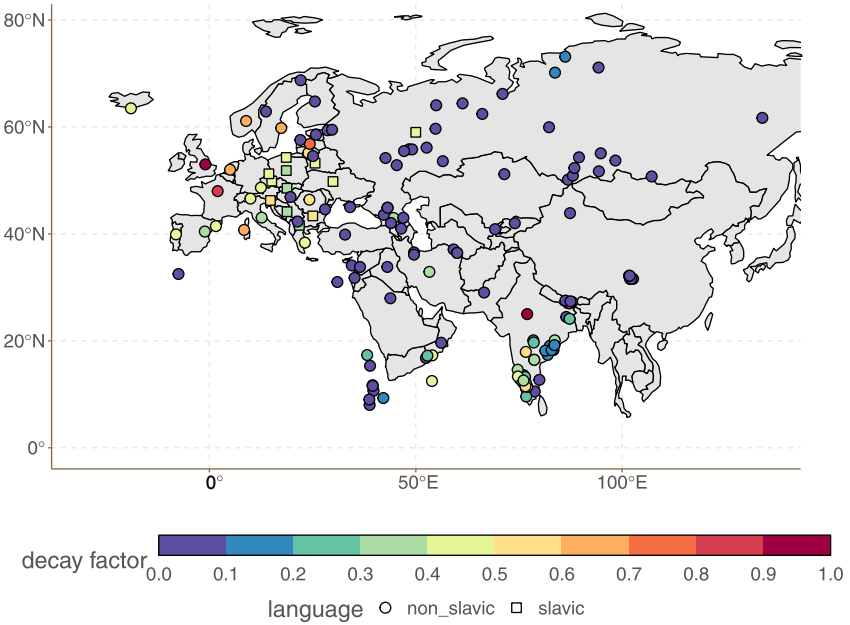

Figure 4: Paradigm decay factor across Eurasian languages based on Seržant's (2021) sample.

Seržant (2021: 72) mentions two hotbeds of decay, namely Northwestern Europe and India. He argues that the hotbed in India is not directly relevant for the Slavic languages, which is why he does not include this zone in his analysis. Instead, Seržant concentrates on the remaining patterns in Northern Eurasia, identifying an innovative zone in Northwestern Europe (high decay), a conservative zone in Northeastern Eurasia (low decay), and a transition zone (intermediate decay). He calls this the "East-West cline" and notes that "[…] it can be reasonably inferred that Slavic languages have retained the morphological functionality of their inflectional person-number indexing system from Proto-Indo-European into Early and Modern Slavic due to their geographic position on the East-West cline" (Seržant 2021: 74). Furthermore, he observes a similar East-West cline within Slavic. Besides this cline and the position of Slavic in the transition zone, Seržant (2021: 75–76) argues that language contact between Slavic and other languages in Northeastern Eurasia is an important component of explaining why Slavic languages (average decay of 0.15) have preserved so much of their paradigm structure in contrast to other modern Indo-European languages in West and Central Europe (average decay of 0.61).

5.3 Models of the replication study

Since the data in this study ranges from 0 to 1, a natural choice of model is a zero-one inflated beta regression model. Regular beta regression is used to model outcomes in the continuous, open interval (0, 1). A zero-one inflated beta regression model can deal with continuous data between 0 and 1, including the values 0 and 1. A zero-one inflated beta regression model consists of three components. The first component is a logistic regression model that decides whether an observation is modeled as continuous data in the interval (0, 1), or as binary data (0 or 1). The second component corresponds to the beta regression part, which models observations in the open interval (0, 1). The third component performs logistic regression and models the remaining observations which are either 0 or 1. As in the previous case study, we added a GP and a phylogenetic term to each of the three components of the model.

In this case we are interested in understanding the spatial effect, but also in exploring the hypothesis that there is a clear East-West cline. If we want to be certain that this cline is due to contact and not an artifact of inheritance, this cline needs to persist once phylogenetic effects are controlled for. In addition, we want to explore the effect of the genetic component on the observed decay factors. This is necessary to clarify how much of the observed patterns can actually be accounted for by inheritance alone. For this reason, we fitted six different models:[21]

1. `m_base`: no controls, intercept only
2. `m_cline`: linear effect of longitude, which represents the East-West cline as proposed by Seržant
3. `m_cline_spline`: non-linear effect of longitude (using a spline), which represents a non-linear East-West cline
4. `m_phylo`: phylogenetic effects
5. `m_gp`: contact effects
6. `m_gp+phylo`: contact and phylogenetic effects

5.4 Results of the replication study

We first focus on the spatial effects of models `m_cline`, `m_cline_spline`, `m_gp` and `m_gp+phylo`. To do so, Figure 5 shows the estimated areal effects of those four models. Since they all include a spatial component, the models make predictions across space which can be plotted as in Figure 5 and visually interpreted.

21 We also corrected some of the latitude and longitude information in the dataset, as some values were missing or conflated (e.g. all Kannada varieties were placed in the same location). The corrected dataset can be found in the Supplementary Material under serant/code/data/data-final.csv.

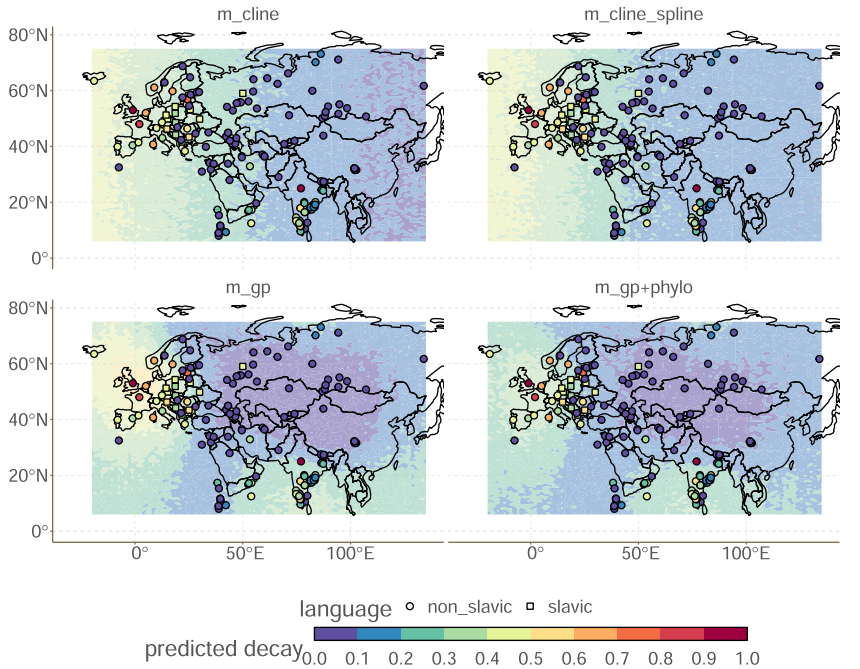

Figure 5: Predicted areal effects.

The left upper plot shows the spatial predictions from `m_cline`, which only uses the longitude information (i.e. horizontal position) of the languages to predict their decay factor. In other words, this model is built to test the assumption of a linear East-West cline of decay. The predicted areal effects by `m_cline` match Seržant's claim about a general East-West cline. With no additional information, `m_cline` predicts a cline in the decay factor decreasing from West to East, with Slavic languages having a decay factor somewhere in between Germanic and Romance on the higher end and Uralic on the lower end. The model `m_cline_spline` tests for the possibility of a general East-West cline which is not linear. As we can see in the right upper plot in Figure 5, we find an almost identical effect when assuming a non-linear East-West cline.

However, once we account for two-dimensional non-linear geographic effects with a GP as in the two lower plots in Figure 5, the picture changes substantially. In a way, the predictions of `m_gp` and `m_gp+phylo` match Seržant's observation of high decay hotbeds and his intuition that the East-West cline is not sufficient to account for the patterns found in Slavic. While Seržant (2021) derives these points from the raw data and theoretical considerations, the models `m_gp` and `m_gp+phylo` offer empirically more robust evidence. Both models no longer predict an East-West cline,

but rather two hotbeds of high decay in Western Europe and South India. Given that low decay corresponds to retention while high decay means a high degree of innovation, we can interpret the model predictions as showing Western Europe and South India to be two hotbeds, where innovation has started and spread from. Besides, Central Asia is predicted to be an area of low decay. Since this area is very large and reflects the absence of innovation, it can be taken as a default situation as opposed to the two hotbeds identified. Therefore, the spatial predictions by `m_gp` and `m_gp+phylo`, taking into account non-linear contact and areal effects, suggest an interpretation that is different from what Seržant concludes in the original study, even though his observations as such are compatible with our findings.

Thus, based on the data without any additional diachronic qualitative analysis of the contact situations in question, we can conclude that it is not so much that Slavic languages have a lower decay factor in their paradigm because they are in close contact with Turkic, Uralic and other languages in the conservative area of Northeastern Eurasia. Rather, Slavic languages are simply farther away from the hotbed of innovation in Western Europe (and the one in Southern India, for that matter), and have thus undergone less decay. In other words, the relevant property does not appear to be the contact with the languages that retained their paradigms but the lack of contact with languages that innovated their person-number indexes.

An important point not addressed so far is the comparison between `m_gp` and `m_gp+phylo`. As can be seen from their spatial predictions in Figure 5, the difference between the two models is minor. It does however show that a portion of the variance captured by the spatial component in `m_gp` is instead accounted for by the phylogenetic term in `m_gp+phylo`. The fact that the predicted spatial distribution of decay does not fundamentally change between `m_gp` and `m_gp+phylo` suggests that even when controlling for phylogenetic effects, spatial effects remain robust.

The other relevant question was whether these areal patterns are actually needed to account for the data, or whether genetic effects would be sufficient in this case. We can explore this question by comparing the predictive performance of different models, i.e. how much of the variation in the decay factors they can capture. The expectation is that if the spatial component is really necessary, then the model with a spatial component in addition to the phylogenetic term (`m_gp+phylo`) should have a better predictive performance than the model which only includes a phylogenetic term (`m_phylo`). If, however, `m_phylo` performs equally well or better than `m_gp+phylo`, then we have to conclude that there is no conclusive evidence for a spatial effect in the data.

We compare the predictive power of the different models using approximate leave-one-out cross-validation (LOO-CV). LOO-CV means that we re-fit the model based on all observations except for one observation at a time in order to make a prediction for that observation. This is repeated for all observations. We approximate this LOO-CV, using the method described in Vehtari et al. (2017). The metric for the comparison is ELPD, the Expected Log pointwise Predictive Density. While it is

Table 1: Approximate LOO-CV.

	ELPD difference	standard error
m_phylo	0.0	0.0
m_gp+phylo	−2.1	3.2
m_gp	−16.3	6.2
m_cline_spline	−57.3	6.5
m_cline	−58.1	6.4
m_base	−65.8	6.6

difficult to interpret in absolute terms, we can use the difference between the ELPD values of the models to compare their predictive performance. A higher ELPD difference value means that we expect the model to perform better, a lower ELPD difference value means that the model performs worse. This is shown in Table 1 for the six different models. Here, the models are arranged according to their performance, from best at the top to worst at the bottom. Table 1 shows the relative ELPD differences to the best performing model, whose value is set to 0. The negative sign indicates that the other models perform worse, and the absolute value quantifies how much worse the model is. The standard error in the last column tells us how certain we can be about the difference between models. It is common to require the ELPD difference to be at least twice as large as its standard error to draw any strong conclusions (Gabry et al. 2019; Vehtari et al. 2017).

From Table 1, we can conclude that there is no clear difference between `m_phylo` and `m_gp+phylo`. Although the model with only phylogenetic effects has a slightly higher ELPD, the standard error is larger than the difference, meaning that this difference can very well be due to chance alone. This does not exclude areal effects from having played a role, but it indicates that the areal patterns and phylogenetic relations in the data are highly correlated. In other words, both predictors contain very similar information about the distribution of decay.[22] This does not mean that contact and areal effects did not play a role in the distribution of decay in languages of Eurasia, but there is no clear evidence for contact and areal effects in the dataset.

We do observe two important differences, however. First, adding phylogenetic effects to `m_gp` marks a clear improvement (ELDP difference of 14.2). This suggests that areal effects alone cannot account for the variation in the data. Second, `m_cline`,

22 This issue cannot be resolved with the dataset as it is. A solution would be to build stronger priors for decay rates of person-number indexes across the world languages. This requires building a larger, global dataset, which would allow us to determine a global decay rate baseline. This baseline could be used as a prior in the models presented here, which could then make more informed assumptions about the likelihood of a decay rate simply being the result of inheritance or of contact.

which assumes a linear longitudinal effect, as well as the `m_cline_spline`, which assumes non-linear longitudinal effects, perform much worse than the other models. Their performance is similar to the one of `m_base`, which did not include any predictors. This means that adding longitude information as a linear or non-linear predictor does not really help to capture the variation in decay factors.

5.5 Taking stock

Even though we cannot fully disentangle the effects of family and contact in our models, the results show that there is little evidence for the conclusion drawn by Seržant (2021) that Slavic languages show comparatively little decay due to their contact with other languages in Northeastern Eurasia. If at all, our models suggest that Slavic languages are relatively far away from the two hotbeds of decay in Western Europe and South India. Our results point to the situation in Slavic resulting from a lack of contact with more innovative patterns, retaining more of the Proto-Indo-European person-number indexes by default. Neither do our results support evidence for the East-West cline. In that, our findings are in agreement with the latter part of Seržant's explanation, where he argues that the cline is not sufficient to capture the decay patterns. Our results go one step further, suggesting that there is no East-West cline, once two-dimensional non-linear areal patterns are fully considered. Moreover, our comparison of model performance suggests that the distribution of decay could also be a product of inheritance alone.

6 Case study: Shcherbakova et al. (2023) on the complexity trade-off hypothesis

6.1 Overview of the original study

Shcherbakova et al. (2023) test the complexity trade-off hypothesis examining grammatical coding in the nominal and verbal domain. Assuming a complexity trade-off, the overall degree of complexity should be comparable across languages, with an increase in complexity over time in one domain, e.g. nouns, being associated with a decrease in complexity in another domain, e.g. verbs. The authors use Grambank data from 244 languages to test for such a co-evolution between the development of coding complexity in the nominal and verbal domain. To do so, Shcherbakova et al. (2023) define 13 so-called feature groups of grammatical coding,

five for the nominal domain and eight for the verbal domain.[23] Each feature group is based on one or more Grambank features, all of which are binary and indicate whether a property is absent (0) or present (1). All feature groups receive a score between 0 and 1. If, e.g. two Grambank features are used to calculate the score of a given feature group, both features contribute proportionally to the score, i.e. 0.5, which is why scores can take values between 0 and 1 as well. The overall metrics for the degree of nominal and verbal coding correspond to the average scores of all nominal and verbal feature groups, respectively.

Shcherbakova et al. (2023) fit two models using BayesTraits (Pagel et al. 2004): one model infers the degree of co-evolution between nominal and verbal grammatical coding from the data, while the second model assumes independent evolution of nominal and verbal grammatical coding. The authors then compare which model can account better for the data in order to assess how likely co-evolution of grammatical coding in the two domains is. Model comparison is performed by calculating the Bayes factor (Burnham and Anderson 2002) between the two models; Shcherbakova et al. (2023) pre-determine to take a value >2 as weak evidence and a value >5 as strong evidence for co-evolution.

The authors fit two additional model series. One is performed on single feature groups (e.g. negation, case marking, tense marking) instead of an overall metric for the nominal versus the verbal domain in order to allow for a more fine-grained picture of co-evolution. The other one includes both types of models, but for selected language families, namely Austronesian, Sino-Tibetan and Indo-European. Shcherbakova et al. (2023) do so in order to assess to what extent co-evolution of nominal and verbal grammatical coding is lineage-specific.

6.2 Original results

The maps in Figure 6, reproducing Figure 4 in Shcherbakova et al. (2023: 161), show the distribution of the nominal and verbal metrics for grammatical coding. Scores closer to 0 indicate a lower degree of grammatical coding and scores closer to 1 represent a higher degree of grammatical coding. Figure 6 allows for two important observations: nominal scores are slightly lower than verbal scores across the board, and Mainland South East Asia appears as an area with particularly low scores for both metrics.

In the first part of the study, testing for the association between nominal and verbal grammatical coding in the whole sample, the authors do not find evidence in

23 The feature groups for the nominal domains are: case, number, gender/noun class, possession, definiteness. The feature groups for the verbal domain are: core argument indexing, transitivity, negation, tense, aspect, mood, non-core argument indexing, clause-related coding.

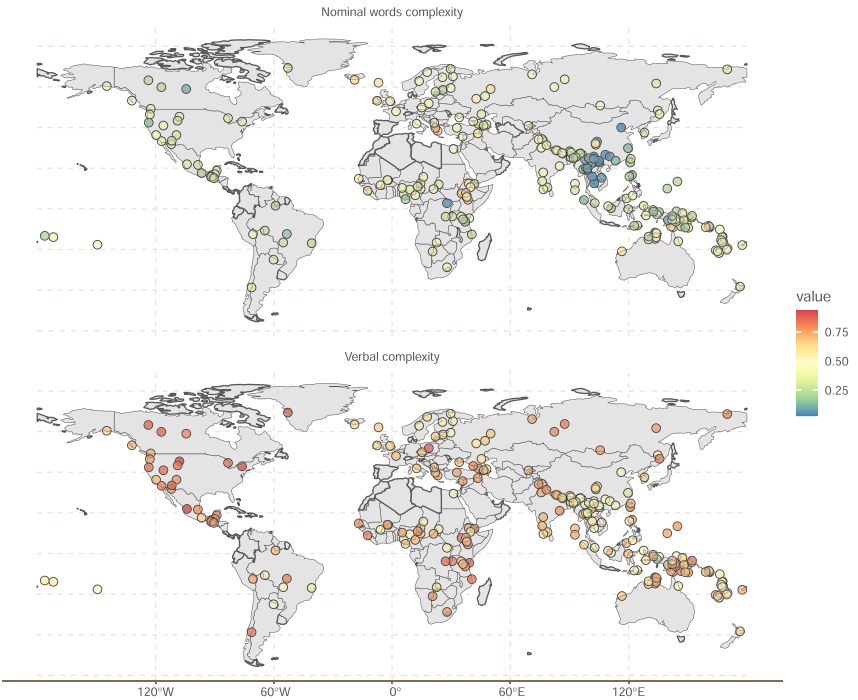

Figure 6: Distribution of complexity scores based on Shcherbakova et al.'s (2023) sample.

support of co-evolution (Bayes Factor = 1.81). They do, however, find a strong phylogenetic signal for both nominal and verbal metrics, which is why they test for co-evolution in Austronesian, Sino-Tibetan and Indo-European separately. The authors find no evidence for co-evolution in Austronesian, but weak support for the co-evolution of nominal and verbal grammatical coding in Sino-Tibetan (Bayes Factor = 3.84). The results for Sino-Tibetan show a positive correlation between nominal and verbal coding ($r = 0.5$), which means that an increase/decrease in nominal coding is accompanied by an increase/decrease in coding in the verbal domain, which is the opposite effect of what we would assume given the complexity trade-off hypothesis.

For Indo-European, Shcherbakova et al. (2023) use two distinct phylogenetic trees. They find strong evidence for co-evolution in Indo-European (Bayes Factor = 5.7) using the tree from Bouckaert et al. (2012). In this case, the authors find a negative correlation ($r = -0.58$) between nominal and verbal coding, which is what we would expect under the complexity trade-off hypothesis. However, when using a different phylogenetic tree from Chang et al. (2015), Shcherbakova et al. (2023) no longer find any support for co-evolution of nominal and verbal coding (Bayes

Table 2: Ten feature group pairs with strongest evidence for co-evolution.

Feature group 1	Feature group 2	Correlation r (95 % HPDI)	Bayes factor
Possession	Aspect	−0.27 (−0.28–0.26)	18.11
Articles	Negation	−0.20 (−0.21–0.19)	9.68
Possession	Negation	−0.14 (−0.15–0.13)	4.90
Gender	Tense	0.09 (0.08 0.10)	2.04
Case	Core arguments	0.11 (0.10 0.11)	2.77
Gender	Non-core arguments	0.11 (0.10 0.12)	2.90
Possession	Tense	0.11 (0.10 0.12)	3.18
Gender	Core arguments	0.13 (0.12 0.14)	4.35
Possession	Core arguments	0.14 (0.13 0.14)	4.60
Case	Mood	0.14 (0.13 0.15)	5.09

Factor = 0.9). The authors explain this difference by a number of languages being included in only one of the phylogenies (namely Panjabi, Gheg Albanian and the three Slavic languages Russian, Czech and Polish).

Given the inconclusive results for a potential co-evolution of grammatical features between the nominal and the verbal domain, in the second part of the study, Shcherbakova et al. (2023) test for correlation between all pair-wise combinations of single feature groups using the global phylogenetic tree. For ten (out of 40) feature group pairs, they find Bayes Factors >2, which the authors interpret as evidence for co-evolution. Table 2 lists those ten feature pairs together with their correlation coefficient r and its 95 % highest posterior density interval (HPDI), as well as their Bayes Factor values.[24] The feature pairs are arranged from the strongest negative (top) to the strongest positive correlation (bottom).

In Table 2, we see that the authors find three feature group pairs with a moderate negative correlation between −0.14 and −0.27. These pairs thus reflect a trade-off in grammatical coding: the increase of grammatical coding in one feature group is moderately correlated with the decrease of coding in the other feature group. For the remaining seven pairs, Shcherbakova et al. (2023) find a weak positive correlation between 0.09 and 0.14, suggesting that there is weak evidence for grammatical coding in both feature groups to develop or to be lost together.

6.3 Models of the replication study

The first part of the study consists of testing the co-evolution between overall grammatical coding in the nominal versus the verbal domain. To measure the

24 Figure 5 in Shcherbakova et al. (2023: 163) shows a visualization of these correlations; the values given in Table 2 are taken from Table D in the supplementary file of Shcherbakova et al. (2023).

correlation between the scores of the nominal and verbal metrics, we use a multivariate Beta regression model with correlated phylogenetic effects. A multivariate model predicts two or more outcomes simultaneously. Here we predict two outcomes at the same time, namely the score of the aggregated nominal metric and the score of the aggregated verbal metric.

Our model assumes that the phylogenetic effects are correlated across the two outcomes. This means that the intercepts for the nominal and the verbal metric are uncorrelated, and that the model builds correlation into the phylogenetic structure instead. Whether the correlation is captured in the outcomes themselves or in the phylogenetic structure used as a predictor has little effect on the correlation estimates. To show that this is indeed the case and that our approach is sound, we carried out a simulation study with synthetic data to compare this approach to one that estimates the correlation between two outcomes. This consists of simulating data from a distribution with known parameters (in this case the correlation between two random variables), and then trying to recover those parameters using a model. The results of both models are very similar (see the Supplementary Materials under `shcherbakova/code/simulation.r` for details).[25]

We fitted a series of two models, one with phylogenetic effects and one with phylogenetic effects and spatial effects with a Gaussian Process, to predict the association between the nominal and the verbal scores. The first model, which only uses the phylogenetic relations between languages as a predictor uses information equivalent to what Shcherbakova et al. (2023) use in their phylogenetic model.[26] The second model includes additional information about the spatial relations between the languages in the dataset, which can be used to assess the role of contact between languages in addition to their phylogenetic structure.

As in the original study, we fitted these two types of models for the entire dataset as well as for Austronesian, Sino-Tibetan and Indo-European. For Indo-European, we also used two versions of phylogenetic trees, taken from Bouckaert et al. (2012) (Indo-European$_1$) and Chang et al. (2015) (Indo-European$_2$). We thus fitted the following models for the first part of the study:

1. whole dataset:
   - `m_global`: correlated phylogenetic effects

25 An alternative approach to estimating the correlation between two outcomes - as used in our simulation - is multivariate normal (or Student) regression with a residual correlation structure. This type of model assumes that the outcomes themselves are correlated. However, a multivariate normal (or Student) regression model requires the outcome to be unbound continuous data. Because the scores of the nominal and verbal metrics used in this study are bound between 0 and 1, we cannot apply this type of model here.

26 As far as we are aware, BayesTraits cannot include spatial information into the model, which is why Shcherbakova et al. (2023) do not do so.

   - `m_global_gp`: correlated phylogenetic and contact effects
2. Austronesian
   - `m_austr`: correlated phylogenetic effects
   - `m_austr_gp`: correlated phylogenetic and contact effects
3. Sino-Tibetan
   - `m_sino`: correlated phylogenetic effects
   - `m_sino_gp`: correlated phylogenetic and contact effects
4. Indo-European
   - `m_indo1`, `m_indo2`: correlated phylogenetic effects
   - `m_indo1_gp`, `m_indo2_gp`: correlated phylogenetic and contact effects

For the second part of the study, i.e. testing the correlation between single feature groups, we used multivariate Beta regression as in the first part. In contrast to the pair-wise comparisons carried out by Shcherbakova et al. (2023), this type of approach allows us to fit a single model to predict the outcome of the 13 feature groups simultaneously. Examining the associations between the different feature groups in a single model has the advantage that we can detect interactions between more than two feature groups, which we could otherwise miss when testing for pair-wise correlations only. Additionally, using a single model produces more reliable estimates than fitting multiple pairwise models.[27] We fitted two models to assess potential contact effects in addition to phylogenetic effects:

5. – `m_13`: corr. phylo. effects for the 13 feature groups
   – `m_13_gp`: corr. phylo. and contact effects[28] for the 13 feature groups

6.4 Results of the replication study

The results of the first part, i.e. measuring the correlation between the nominal and verbal metrics in the global dataset, are shown in Figure 7. We see the posterior correlations for the models without (`m_global`) and with a spatial component (`m_global_gp`), respectively. Both models find a considerable positive correlation; there is only a minimal mass of the posterior correlation below zero, which corresponds to 0.002 (`m_global`) and 0.04 (`m_global_gp`) probability. In contrast to Shcherbakova et al. (2023), given the data and our models, we can be very confident that there is in fact a weak positive correlation between the amount of nominal and

27 In the Supplementary Materials we show the effect of performing pair-wise correlation models versus a single large model using synthetic data and known parameter values. We find that a single large model is better at finding the original parameter values than a series of pair-wise correlation models.

28 For computational reasons, we used an approximation with a spline in this model.

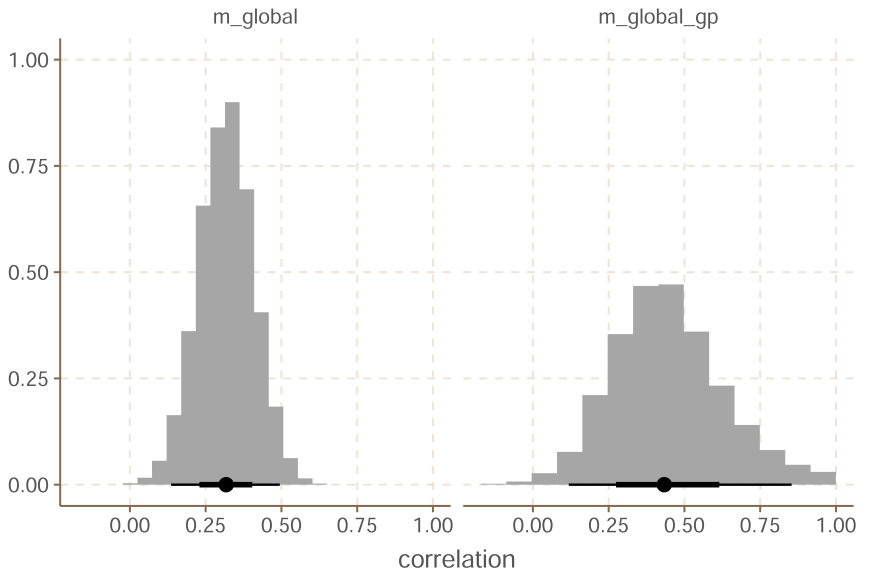

Figure 7: Posterior correlation with and without spatial component.

verbal grammatical coding. As the next step, Shcherbakova et al. (2023) looked at the correlation between nominal and verbal metrics in individual families. Figure 8 shows our results for Indo-European (using two different trees), Austronesian and Sino-Tibetan.

When fitting models for individual families, we no longer find clear evidence for correlations between the nominal and verbal metrics, because all correlation values receive some amount of probability, meaning that we cannot exclude any values. This is very likely a consequence of using much smaller datasets than when including all languages from the sample. The models fitted on fewer data points simply have considerably less certainty about the correlation. Although we do not find clear evidence for correlations, Figure 8 does show a few interesting differences across families and models. In general, we see that the models without a GP for spatial control (left side) show a somewhat less equal distribution of the probability mass. For the Austronesian and the two Indo-European models, we find more probability associated with negative correlation scores. The Sino-Tibetan model, on the other hand, shows a higher probability for positive correlation scores, similarly to the findings of Shcherbakova et al. (2023). However, once we include a GP in our models to control for potential spatial effects, the plots on the right side in Figure 8 show that the weak tendencies disappear altogether. This suggests that most of the correlation

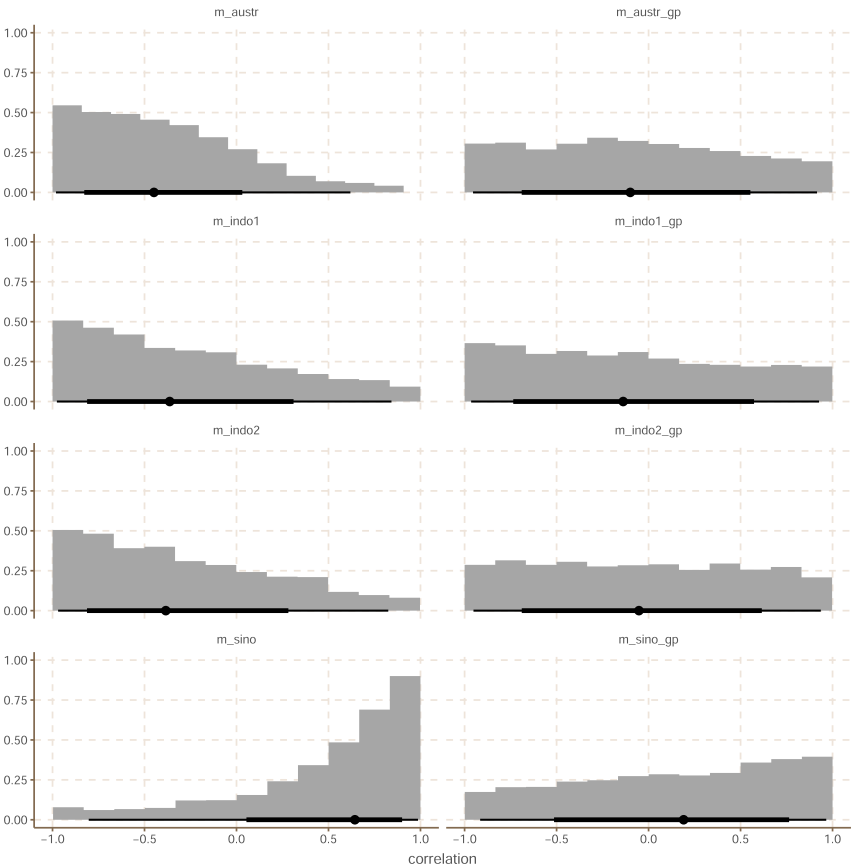

Figure 8: Posterior correlation with and without spatial component for individual families.

between nominal and verbal grammatical marking metrics can be accounted for by contact between languages.

We can therefore only partially replicate the original results of Shcherbakova et al. (2023), who find no (conclusive) evidence for a correlation between nominal and verbal metrics for Austronesian and Indo-European, but do find evidence for a positive correlation for Sino-Tibetan.

The second part of the original analysis by Shcherbakova et al. (2023) consists of testing for correlations between individual feature groups from the nominal and verbal domains (cf. Table 2). As explained in Section 6.3, our approach differs from the original one in an important way. While Shcherbakova et al. (2023) built multiple models for pair-wise comparisons, we fit a single model including all feature groups and we calculate a correlation matrix across all feature groups. By including all

feature groups, the model has more information about how the different feature groups are associated with each other and it can produce more reliable estimates.

Figure 9 shows the posterior correlations for all pairs of feature groups and compares them to the original findings from Shcherbakova et al. (2023). Our model results with spatial control (`m_13_gp`) are given in dark blue; all estimates additionally include 50 % (bold) and 95 % (light) uncertainty intervals. We see the correlations found in the original study in brown.[29] The ten feature pairs for which Shcherbakova et al. (2023) find sufficient evidence for co-evolution (BF > 2, cf. Table 2) are signaled by triangles, and the remaining correlations that the original study tested for are shown as points.[30]

Going into details with regard to the linguistic interpretation of the correlations between feature groups would surpass the scope of the present paper, especially because the feature groups do not correspond to single linguistic features. Here, we only focus on whether or not we can replicate the original findings when using different statistical methods for the analysis. The most important point that we see in Figure 9 concerns the ten feature pairs for which Shcherbakova et al. (2023) found evidence for positive or negative correlations, i.e. co-evolution. Although our results generally agree with the original findings in the direction of the correlation, our models do not fully replicate the original findings. Out of the ten feature pairs with evidence for co-evolution in the original study, we only find sufficient evidence for positive correlations for possession – core arguments, and gender – core arguments. For the other pairs of feature groups (gender – tense, case – core arguments, possession – tense, case – mood, articles – negation, possession – negation, possession – aspect), our results show much more uncertainty about the probability distribution of the correlation coefficient including 0. Additionally, some of our mean estimates, even if pointing into the same direction, are quite different from the correlation values reported by Shcherbakova et al. (2023), i.e. possession – core arguments, gender – core arguments, gender – non-core arguments, and possession – aspect. Besides the pairs tested in the original study, we do find clear evidence for a correlation between case – possession, core arguments – non-core arguments, transitivity – non-core arguments and number – gender.

29 The results from our model without spatial control (`m_13`) are very similar to the ones shown here. For a comparison between `m_13` and `m_13_gp`, see `shcherbakova/code/plots/p-cor-compare.pdf` in the Supplementary Materials.

30 In the original study, Shcherbakova et al. (2023) only test correlations for all feature groups from the nominal domain with all feature groups from the verbal domain. They do not analyze the correlation between all pairs of feature groups within the nominal and verbal domains. Because we test the correlations of all pairs of feature groups together in our model, it would not have been possible to add this restriction. This results in a number of feature pairs shown in Figure 9 which do not have any corresponding result in the original study.

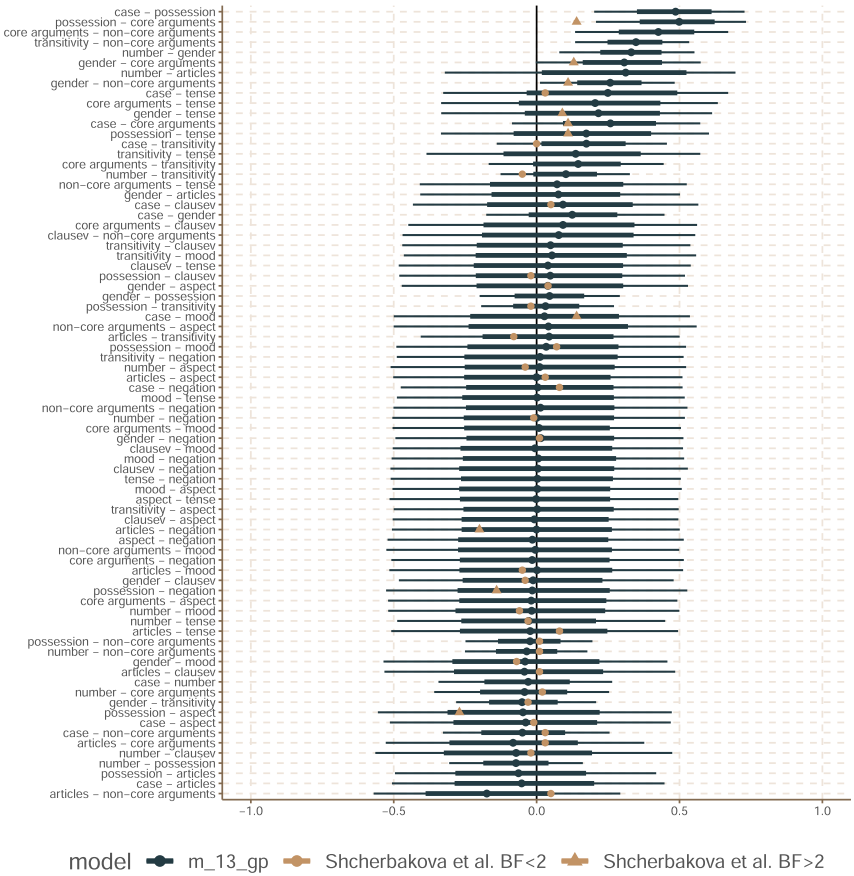

Figure 9: Posterior correlation between individual feature groups without (`m_13`) and with spatial component (`m_13_gp`).

6.5 Taking stock

The main finding from this section is that different statistical analyses of the same data can lead to very different results. While Shcherbakova et al. (2023) find no evidence for a correlation between nominal and verbal metrics of grammatical coding, we find clear evidence for a weak positive correlation (around 0.3) between nominal and verbal metrics in the global dataset. In opposition to the original study, we no longer find evidence for a correlation when fitting the models for individual language families. For our approach, it makes sense that we find a higher degree of uncertainty in the posterior correlation (i.e. no longer clear evidence) when fitting

the model on a smaller dataset. For the second part, testing for correlations between individual feature groups, our results suggested correlations in the same direction as in the original study, but often with widely different values and degrees of certainty. Overall, we could not replicate the findings in that many of the ten pairs of feature groups with evidence for co-evolution in Shcherbakova et al. (2023) did not show evidence for a correlation with our models. In addition, we found evidence for correlations that the original study did not find. While we could not fully replicate Shcherbakova et al.'s results, we did not find entirely opposite trends and patterns either. This discrepancy between the original findings and ours could therefore be a consequence of the relatively small size and high complexity of the dataset. It is possible that more data would be helpful for clearer cross-linguistic tendencies and more methodologically robust patterns in this case. Therefore, this replication study highlights the importance of not taking a single statistical analysis as an absolute result and of considering that the methods used can impact findings in a significant way. This also means that any result taken from statistical models calls for careful theoretic evaluation by the linguist.

7 Discussion

7.1 Towards better replicability in typology

The four case studies presented in this paper (including Appendix C) have shown that some but not all results can be replicated when using the same data but (other) statistical techniques for analysis. There is no good a priori indicator for which findings are and which are not robust. Therefore, it is necessary to include regular, systematic replication in standard typological practices.

Replication is only possible if the original study is fully transparent in its description and documentation of the data, the annotation and analysis process. However, not much has been proposed as guidelines for transparency and replicability in typology. The only concrete proposal we are aware of comes from Harris et al. (2006). Adapting their proposal, we identify four main levels that require full transparency in order to allow for the replication or independent verification of a typological study and parts thereof:

(7) *Transparency requirement of typological studies*
   a. primary data collection
   b. secondary data collection / language sample
   c. data analysis & annotation
   d. (statistical) methods

As discussed above, we do not address primary data collection here, as it is often less relevant for large-scale, quantitative typological studies such as the ones presented here. Furthermore, there is substantial work from the language documentation literature that addresses transparency standards and provides guidelines for primary data collection (cf. Gawne and Berez-Kroeker 2018; Himmelmann 1998; Maxwell 2012). Since such a discussion is still lacking for the other three levels shown in (7), we provide concrete suggestions in Appendix D for best practices for full transparency and replicability in (quantitative) typology.

7.2 Evaluating methodological robustness

The main purpose of the three replication studies reported here was to evaluate the methodological robustness of previous typological studies. Although p-hacking and other poor statistical practices have been problematized in the linguistic literature (cf. Sönning and Werner 2021), the evaluation of methodological robustness in linguistics, including typology, has not received the attention it deserves. We will therefore discuss it in more detail in the remainder of this section.

7.2.1 The need for systematic methodological evaluation

Our results showed a variegated picture in that some of the original results could be replicated using more advanced statistical modeling, while others could not be replicated. Dryer (2018) used sampling methods in order to control for phylogenetic and contact effects. Our results generally replicate the results of Dryer (2018), which means that we can be somewhat more confident in the results on the one hand and in the sampling method on the other. As is shown in Appendix C, Berg (2020) also used manual sampling. He reported results that appeared to overestimate the linguistic effect in contrast to our results, which may be due to the specific sampling method employed in the original study. Seržant (2021) carried out an areal typological study, which is why he did not require a balanced sample but the maximum obtainable coverage of a region. His conclusions were mainly based on visual inspection of the spatial distribution of the relevant patterns. We showed how a statistical modeling analysis could be performed on the original dataset, and how it led to insights that go beyond Seržant's original conclusions.

Our replication of Shcherbakova et al.'s (2023) study shows that two different quantitative analyses, both based on statistical modeling and using the same data, can lead to different results. This has two important consequences. First, it shows how important it is that we as linguists are careful when interpreting findings from statistical models, and that we acknowledge the degree of uncertainty that comes

with every analysis. Second, on a more general and methodological level, our results highlight the importance of careful and systematic model evaluation and validation for the statistical techniques used in typology.

As mentioned in Section 2.3, most previous replication studies in typology that focused on methodological robustness dealt with the influence of ecological factors on a given linguistic property and with strong and controversial claims. The aim of the present study was to draw attention to the general need for evaluating methodological robustness, regardless of the research question or the conclusions.

7.2.2 The advantage of statistical bias control

The present study also showed the advantages of statistical models over statistical tests and, of course, using no statistics at all for a quantitative analysis. This is mostly based on the fact that a statistical test is not able to capture dependencies between observations and can therefore only be applied in very specific situations. There are two possible ways that this has been dealt with in previous research, neither of which is ideal. Either the research question and dataset have to be adapted to meet the criteria of statistical tests, which may lead to a much more simplified view of the linguistic reality at hand. Or, the research question and dataset are not adapted, the test is applied nevertheless, and unwarranted conclusions are drawn. Issues related to the use of statistical tests under the wrong circumstances and to the wrongful interpretation of their results have been raised by various researchers from different disciplines for a long time.[31] Also in linguistics, a number of studies from different research areas have argued against the use of statistical tests and for the use of statistical modeling (often mixed effect regression) instead. Examples are Baayen et al. (2008); Jaeger (2008) and Vasishth et al. (2018) for psycholinguistics, Gries (2015) and Paquot and Plonsky (2017) for corpus linguistics, Aguilar-Sánchez (2014) and Tagliamonte and Baayen (2012) for socio-linguistics and Roettger (2019) and Roettger et al. (2019) for phonetics. Moreover, Winter and Grice (2021) offer a recent and detailed discussion of non-independent observations and their consequences in linguistics in general. Finally, Coupé (2018) describes different types of complex statistical models that are useful to account for dependencies in linguistic data.

Zooming in on typology, we find much less discussion on using statistical modeling instead of statistical tests in the literature. Some of these methodological considerations were part of replication studies criticizing the methodology used in previous studies (cf. Section 2.3). Examples are Hartmann (2022); Jaeger et al. (2011) and Roberts et al. (2015), who showed that including some form of statistical control

31 Cf. Berkson (1942); Cohen (1994); Cumming (2012); Kline (2013) and Ziliak and McCloskey (2008) for more details.

for phylogenetic and/or contact relations between languages results in a much weaker effect than the one found in the original studies, or in no effect at all. The results of the present study are very much in line with this trend. When using advanced statistical techniques for bias control in the sample, we found smaller effects. This reflects the general insight that disregarding non-independencies between observations likely leads to false positives or type 1 errors (cf. Winter and Grice 2021).

In particular, we found that the manual sampling method for phylogenetic bias used by Berg (2020) likely overestimated the proportions of gender marking languages (cf. Appendix C). As for Dryer (2018), we could replicate most of the original results. The automated repeated sampling from a larger sample he used is therefore likely to be a more suitable sampling method. Still, some minor differences between Dryer's original and our results are likely related to a number of areal effects that Dryer's sampling method does not account for. This suggests that sampling as a form of bias control (phylogenetic or contact) may not be ideal, and that statistical bias control in the form of a phylogenetic regression term and a Gaussian Process are able to represent the dependencies between languages in a sample more accurately. Besides, statistical bias control has the advantage of allowing to keep all datapoints in, and the model can make use of the information about dependencies between languages.[32]

The case study replicating Seržant (2021) emphasized how insightful the Gaussian Process is as a statistical tool to model contact and areal effects. Seržant (2021) carried out an areal typological study where no balanced sample but maximum obtainable coverage of an area was needed. The original study did not use statistical tools to control for phylogenetic or contact effects and mostly relied on visual inspection of the geographical patterns for the analysis. This led to the overestimation of linear areal effects, i.e. the East-West cline, in Seržant (2021). Our replication study showed that a model including a GP, which can capture non-linear spatial effects in the data, captures much more of the variation in the data than a model with a linear longitude predictor. This shows that a statistical tool to model contact or areal effects leads to insights that capture more of the complex interaction between languages in reality.

7.2.3 Accepting uncertainty

The other major insight from this study relates to the fairly high degree of uncertainty around some of the predicted means of our models. In the spirit of Gelman (2018) and Vasishth and Gelman (2021), we propose to accept uncertainty in statistical

32 See the Supplementary Materials (`model.R` in the folder `dryer`) for a thorough comparison.

analyses in typology. Uncertainty is at the core of any statistical analysis, since statistical tests and models serve to quantify the amount of uncertainty with respect to an observed effect.

Returning to our model predictions, high uncertainty around a predicted value does not mean that the model is "bad" or uninformative. In fact, in the replication of the Dryer (2018) study, the model that captured the variation in the data best also had the largest uncertainty intervals around the means of the predictions.[33] The high degree of uncertainty about the expected proportions in the replication studies of Dryer (2018) and Berg (2020) (cf. Section 4 and Appendix C) results from much of the variation in the data being accounted for by the phylogenetic and contact controls. If two closely related languages or languages spoken in close proximity to each other have the same linguistic feature, the model can attribute this to those relations. At the same time, the model takes these dependencies into account when estimating the real distribution of a linguistic feature once biases are controlled for. The expected values we reported thus represent what the model predicts on top of phylogenetic and contact relations.

This means that a model with such controls, outperforming a simpler model, is likely to make predictions that are less certain than a simpler model. The predictions of the simpler model may look more certain and confident and can appear "better" at first sight, but this is not the case. The simpler model is less able to represent real linguistic complexities. As it contains less (and simpler) information in the predictors, it provides more confident results. This is a common issue in science, where the application of statistics is often no longer used to estimate and then evaluate the degree of uncertainty of a result, but instead used to (erroneously) provide certainty, if not proof, about the existence of an effect. Besides testing for methodological robustness of previous results, our four replication studies also served as examples of how a statistical analysis in typology can focus more on estimation and on accepting uncertainty.

8 Concluding remarks

The present paper has called for more attention to replication in typology, since it is a valuable tool for evaluating the robustness of previous results. In particular, we focused on replication using the original data but applying a different statistical analysis to test for methodological robustness. We did so employing advanced

33 Cf. Guzmán Naranjo and Becker (2022) and Verkerk and Di Garbo (2022) for more information on these methods of statistical bias control.

statistical bias controls, namely phylogenetic regression for genetic effects and a Gaussian Process for contact effects. Our findings indicated that some of the original results could be replicated, but some could not. On the one hand, finding agreement between the main results is reassuring and allows for some confidence in them. On the other, this type of replication revealed important methodological insights. In line with previous work in typology, our comparisons showed that more advanced statistical techniques that can model the phylogenetic and contact relations between languages do pick up more complex patterns in the data than traditional sampling methods. The patterns may not always provide clearer answers and they may make the interpretation more difficult. Statistics helps us to quantify and evaluate the degree of uncertainty of our results. It should not be used as tool for certainty or proof, and we must remember that there is no single best way to analyze a given dataset. We showed that there still is much to learn about various linguistic questions when replicating previous studies and comparing results.

Acknowledgments: We wish to thank the participants of the Freiburg Linguistics reading group, Uta Reinöhl, Naomi Peck and Marvin Martiny, as well as three anonymous reviewers for their valuable comments on earlier versions of this study.
Author contributions: Both authors contributed equally to all aspects of the study and the paper.
Research funding: Matías Guzmán Naranjo received funding from the German Research Foundation (DFG, grant no. GU 2369/1-1, project number 504155622).

References

Aguilar-Sánchez, Jorge. 2014. Replicability of (socio) linguistics studies. *Journal of Research Design and Statistics in Linguistics and Communication Science* 1(1). 5–25.

Atkinson, Quentin. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346–349.

Baayen, Harald, Davidson Douglas & Bates Douglas. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS One* 10(6). e0128254.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Heston Tyler, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Berg, Thomas. 2020. Nominal and pronominal gender: Putting Greenberg's Universal 43 to the test. *Language Typology and Universals* 73(4). 525–574.

Berkson, Joseph. 1942. Tests of significance considered as evidence. *Journal of the American Statistical Association* 37(219). 325–335.

Bickel, Balthasar, Alena Witzlack-Makarevich & Taras Zakharko. 2015. Typological evidence against universal effects of referential scales on case alignment. In Ina Bornkessel-Schlesewsky, Andrej Malchukov & Marc Richards (eds.), *Scales and hierarchies: A cross-disciplinary perspective*. Berlin: De Gruyter.

Bisang, Walter. 2011. Variation and reproducibility in linguistics. In Peter Siemund (ed.), *Linguistic universals and language variation*, 237–263. Berlin: De Gruyter.

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

Bürkner, Paul-Christian. 2017. Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28.

Burnham, Kenneth & David Anderson. 2002. *Model selection and multimodel inference*. New York: Springer.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Brubaker Marcus, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software Articles* 76(1). 1–32.

Chang, Will, David Hall, Chundra Cathcart & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Chen, Keith. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review* 103(2). 690–731.

Cohen, Jacob. 1994. The earth is round ($p < 0.05$). *American Psychologist* 49(12). 997–1003.

Corbett, Greville. 2005. Suppletion in personal pronouns: Theory versus practice, and the place of reproducibility in typology. *Linguistic Typology* 9(1). 1–23.

Coupé, Christophe. 2018. Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology* 9. 1–21.

Cumming, Geoff. 2012. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Donoho, David. 2010. An invitation to reproducible computational research. *Biostatistics* 11(3). 385–388.

Donohue, Mark. 2011. Stability of word order: Even simple questions need careful answers. *Linguistic Typology* 15(2). 381–391.

Dryer, Matthew. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292.

Dryer, Matthew. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.

Dryer, Matthew. 2011. The evidence for word order correlations. *Linguistic Typology* 15(2). 335–380.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798–833.

Everett, Caleb. 2017. Languages in drier climates use fewer vowels. *Frontiers in Psychology* 8. 1285.

Everett, Caleb, Damián Blasi & Seán Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences of the United States of America* 112(5). 1322–1327.

Everett, Caleb, Damián Blasi & Seán Roberts. 2016. Language evolution and climate: The case of desiccation and tone. *Journal of Language Evolution* 1(1). 33–46.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt & Andrew Gelman. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2). 389–402.

Gawne, Lauren & Andrea Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 22–32. Honolulu: University of Hawai'i Press.
Gelman, Andrew. 2018. Ethics in statistical practice and communication: Five recommendations. *Significance* 15(5). 40–43.
Goodman, Steven, Daniele Fanelli & John Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8(341). 341ps12.
Gries, Stefan. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93–117.
Grieve, Jack. 2021. Observation, experimentation, and replication in linguistics. *Linguistics* 59(5). 1343–1356.
Hammarström, Harald. 2016. Commentary: There is no demonstrable effect of desiccation. *Journal of Language Evolution* 1(1). 65–69.
Hammarström, Harald, Robert Forkel & Martin Haspelmath and Sebastian Bank. 2022. *Glottolog 4.7*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Harris, Alice, Larry Hyman & James Staros. 2006. What is reproducibility? *Linguistic Typology* 10(1). 69–73.
Hartmann, Frederik. 2022. Methodological problems in quantitative research on environmental effects in phonology. *Journal of Language Evolution* 7(1). 95–119.
Haspelmath, Martin & Sven Siegmund. 2006. Simulating the replication of some of Greenberg's word order generalizations. *Linguistic Typology* 10(1). 74–82.
Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196.
Jaeger, Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.
Jaeger, Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319.
Jäger, Gerhard & Johannes Wahle. 2021. Phylogenetic typology. *Frontiers in Psychology* 12. 2852.
Kline, Rex. 2013. *Beyond significance testing: Statistics reform in the behavioral sciences*. Washington, D.C.: American Psychological Association.
Kobrock, Kristina & Timo Roettger. 2023. Assessing the replication landscape in experimental linguistics. *Glossa Psycholinguistics* 2(1). 1–28.
Levshina, Natalia. 2022. *Communicative efficiency: Language structure and use*. Cambridge: Cambridge University Press.
Machery, Edouard. 2020. What is a replication? *Philosophy of Science* 87(4). 545–567.
Maddieson, Ian. 2006. Correlating phonological complexity: Data and validation. *Linguistic Typology* 10(1). 106–123.
Maddieson, Ian. 2018. Language adapts to environment: Sonority and temperature. *Frontiers in Communication* 3. 1–8.
Martin, Alexander, Theeraporn Ratitamkul, Klaus Abels, David Adger & Jennifer Culbertson. 2019. Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard* 5(1). 20180072.
Maxwell, Mike. 2012. Electronic grammars and reproducible research. In Sebastian Nordhoff (ed.), *Electronic grammaticography*, 207–235. Honolulu: University of Hawai'i Press.
Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.
Guzmán Naranjo, Matías & Miri Mertner. 2023. Estimating areal effects in typology: A case study of African phoneme inventories. *Linguistic Typology* 27. 455–480.
Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119.

Nichols, Johanna, Jonathan Barnes & David Peterson. 2006. The robust bell curve of morphological complexity. *Linguistic Typology* 10(1). 96–106.

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

Paquot, Magali & Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1). 61–94.

R Core Team. 2023. *R: A Language and Environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, Carl Edward. 2004. Gaussian processes in machine learning. In Olivier Bousquet, Ulrike von Luxburg & Gunnar Rätsch (eds.), *Advanced lectures on machine learning*, 63–71. Berlin: Springer.

Roberts, Seán. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology* 9. 1–21.

Roberts, Seán, James Winters & Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. *PLoS One* 10(7). e0132145.

Roettger, Timo. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology* 10(1). 1–27.

Roettger, Timo, Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.

Saldana, Carmen, Yohei Oseki & Jennifer Culbertson. 2021. Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology. *Journal of Memory and Language* 118. 104204.

Schmidt, Stefan. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* 13(2). 90–100.

Schmidtke-Bode, Karsten & Natalia Levshina. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In Ilja Seržant & Alena Witzlack-Makarevich (eds.), *The diachronic typology of differential argument marking*, 509–537. Berlin: Language Science Press.

Seržant, Ilja. 2021. Slavic morphosyntax is primarily determined by its geographic location and contact configuration. *Scando-Slavica* 67(1). 65–90.

Shcherbakova, Olena, Volker Gast, Damián Blasi, Hedvig Skirgård, Russell Gray & Simon Greenhill. 2023. A quantitative global test of the complexity trade-off hypothesis: The case of nominal and verbal grammatical marking. *Linguistics Vanguard* 9(s1). 155–167.

Siewierska, Anna & Dik Bakker. 1996. The distribution of subject and object agreement and word order type. *Studies in Language* 20. 115–161.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). 20191010.

Song, Jae Jung. 2012. *Word order*. Cambridge: Cambridge University Press.

Sönning, Lukas & Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59(5). 1179–1206.

Steele, Susan. 1978. Word order variation: A typological survey. In Joseph Greenberg, Charles Albert Ferguson & Edith Moravcsik (eds.), *Universals of human language IV: Syntax*, 585–623. Stanford, CA: Stanford University Press.

Tagliamonte, Sali & Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Tal, Shira, Kenny Smith, Jennifer Culbertson, Eitan Grossman & Inbal Arnon. 2022. The impact of information structure on the emergence of differential object marking: An experimental study. *Cognitive Science* 46(3). e13119.
Tomlin, Russell. 1986. *Basic word order: Functional principles*. London: Croom Helm.
Van Tuyl, Rory & Asya Pereltsvaig. 2012. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335(6069). 657.
de Villemereuil, Pierre & Shinichi Nakagawa. 2014. *Modern phylogenetic comparative methods and their application in evolutionary biology*. Berlin: Springer.
Vasishth, Shravan & Andrew Gelman. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* 59(5). 1311–1342.
Vasishth, Shravan, Daniela Mertzen, Lena Jäger & Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.
Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.
Verkerk, Annemarie & Francesca Di Garbo. 2022. Sociogeographic correlates of typological variation in northwestern Bantu gender systems. *Language Dynamics and Change* 12(2). 155–223.
Wichmann, Søren & Taraka Rama. 2021. Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society B* 376(1824). 20200202.
Widmann, Thomas & Peter Bakker. 2006. Does sampling matter? A test in replicability concerning numerals. *Linguistic Typology* 10(1). 83–95.
Williams, Christopher & Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*, Vol. 2. Cambridge, MA: MIT Press.
Winter, Bodo & Martine Grice. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5). 1251–1277.
Ziliak, Stephen & Deirdre McCloskey. 2008. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/lingty-2023-0076).