Replication and methodological robustness in typology

1 Introduction

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. If so, then, more certainty can be attributed to the 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. This study addresses the issue of replication in typology in a different way. We use the original datasets of three previous studies to show how statistical modelling can be used to test the replicability of typological studies. This type of replication, i.e. using the original datasets with other methods, is necessary to assess to what extent the results are method-dependent. The fact that there is no single best statistical approach to analysing 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 thus 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.

The present paper is thus an exercise in replication using statistical techniques on the original datasets from studies using more traditional methods. We selected the following three test cases: Dryer (2018) on the order of elements in the noun phrase, Seržant (2021) on contact effects in Slavic morphosyntax and Berg (2020) on the association between gender marking on nouns and different types of pronouns.[1] 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.

For all three case studies it could be argued that the data should be collected and annotated in a different manner. This, however, will not be at issue in the present paper, as its 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 & Becker (2022) and Verkerk & Di Garbo (2022) in using phylogenetic regression to control for genetic effects and a Gaussian Process to control for contact and areal effects. As we will show, 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.

[1] 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.

The structure of the paper is as follows. Section 2 defines the relevant notions and gives a brief overview of replication in typology. Section 3 introduces the current approach, including the statistical methods used here to evaluate the three case studies. We then present and discuss the three case studies (Berg 2020, Dryer 2018, Seržant 2021) in Sections 4, 5 and 6, respectively. In Section 7, we discuss the three case studies in the light of replicability and transparency in typology, focusing on methodological robustness and statistical bias control in typology. Finally, Section 8 concludes.

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.[2] Before turning to replication and replicability, we need to clarify what we mean by robust findings. We will define robustness following Goodman, Fanelli & Ioannidis (2016) as shown in (1). Applied to typology, robust findings then need to hold across (i) different language samples, (ii) different types of grammatical categorizations 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, Fanelli & Ioannidis 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 Gould & Kolb (1964) and Schmidt (2009: 91), including the idea of uncertainty (cf. Gelman 2018, Vasishth & 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. Doing so, it can be used to establish a piece of knowledge of our world.

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.[3] 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, annotation as well as the analysis are identical to the original ones. While

[2]For more details on different types and uses of replication and replicability, cf. Donoho (e.g. 2010), Gawne & Berez-Kroeker (2018), Goodman, Fanelli & Ioannidis (2016), Hüffmeier, Mazei & Schultze (2016), Machery (2020) and references therein.

[3]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. Similarly, it is not clear a priori which results are (more) correct in case they differ between the original and the replication study.

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**[4]
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, Fanelli & Ioannidis 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 & Berez-Kroeker 2018, Grieve 2021, Harris, Hyman & Staros 2006, Himmelmann 1998, Kobrock & 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. A number of typological studies fall into 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, 2013, Foster & Hofling 1987, Siewierska & Bakker 1996, Sinnemäki 2010, 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 thematic 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 & Siegmund (2006: 74) make a more concrete proposal as to how replication can apply to typological work. Updating their classification gives us the distinction of 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 **categorization & 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,

[4]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".

we will not discuss replicability of levels (a) and (b) further.[5] 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, Barnes & Peterson (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 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 of the *LT* issue 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 & Siegmund 2006, Maddieson 2006, Widmann & 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. It is our impression that it is often implicitly assumed that quantitative typological studies are robust in terms of methods used. The purpose of the present study thus is to draw attention to the fact that we must not make this assumption, but replicate previous studies in order 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) reported 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.[6] Chen (2013) found 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, Winters & Chen 2015, Van Tuyl & Pereltsvaig 2012) revealed that the effects found in the original studies disappeared 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 established an association between an environmental factor and a linguistic property. Everett (2017) reported a relation between ambient humidity and the vowel-consonant ratio, concluding that languages in drier cli-

[5] Nevertheless, we acknowledge that data robustness on those two levels is a 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 & Berez-Kroeker 2018, Himmelmann 1998).

[6]

mates 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.

We are only aware of one replication study for methodological robustness which could confirm the original findings. Everett, Blasí & Roberts (2016), replicating Everett, Blasi & Roberts (2015), found phonemic tones to be more likely to develop in warmer climates than in colder or dessicated ones. In the replication study, Everett, Blasí & Roberts (2016) reacted to methodological criticism from Hammarström (2016) and adjusted their statistical model. According to the authors, Everett, Blasí & Roberts (2016) could replicate their original results.

Thus, examples of replication studies that independently test other typological studies for methodological robustness are relatively rare. 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, though, 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.[7]

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

[7] This last point applies especially to Atkinson (2011), which was published as a target article, with the goal of sparking a discussion in the linguistic community. Cf. Cysouw, Dediu & Moran (2012) and Wang et al. (2012) for more comments on Atkinson (2011) with methodological and data-related criticism.

(Verkerk & Di Garbo 2022) and Gaussian Processes for areal controls (Guzmán Naranjo & Becker 2022), as well as other types of phylogenetic models (Jäger & Wahle 2021). Despite the abundance of different available techniques, there is very little work comparing how robust results are across these different 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. Especially more recent studies which make use of statistical modeling no longer 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 & Di Garbo (2022), Guzmán Naranjo & Becker (2022) and Guzmán Naranjo & Mertner (2022), as well as the tutorials in the supplementary materials for the concrete computational implementations. All models were coded using Stan (Carpenter et al. 2017) and in some cases also the *brms* package (Bürkner 2017) in R 4.3 (R Core Team 2023). We will discuss more details regarding the models used for each of the three case studies in Sections 4.3, 5.3 and 6.3, respectively.

3.2.1 Phylogenetic regression

To control for phylogenetic effects we make use of a method called phylogenetic regression.[8] 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 & Becker 2022, Verkerk & Di Garbo 2022). It could be shown to be able to deal with bias resulting from multiple related languages in a sample.

[8] An exhaustive mathematical description of phylogenetic effects can be found in (de Villemereuil & Nakagawa 2014).

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).[9] In this study, we use a Gaussian Process (GP) to control for contact bias.[10] 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 about a large number of languages. In principle, a GP be used with other distance metrics as well that capture the spatial properties of languages in a more realistic way.[11] For other examples of GPs used to control for spatial effects in typological studies see Guzmán Naranjo & Becker (2022) and Guzmán Naranjo & Mertner (2022).

Thus, the advantage of using a GP to model areal or contact effects over other methods that rely on sampling 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 of the original study

Dryer (2018) surveys different word orders of elements in the nominal domain across a sample of 576 languages. The elements examined are the demonstrative (Dem), numeral (Num), adjective (A) and noun (N). Two examples to illustrate different word orders in the nominal domain are given in (5) and (6).

(5) tshɔ́hà jɔmỳ xhǿ njì ɣà
person.**N** good.**A** those.**Dem** two.**Num** CLF
'those two good persons'
Akha (Dryer 2018: 800)

(6) mi ran^{n}gaḷu tin fot
this.**Dem** good.**A** three.**Num** book.**N**
'these three good books'
Dhivehi (Dryer 2018: 800)

Dryer (2018) provides a dataset with 1096 languages, but out of these only 593 are coded for word order. The distribution of languages and features values is shown in Figure 1.

[9] There are a few other, more principled approaches to control for contact bias. See the overview in Guzmán Naranjo & Becker (2022: 22-26) for more details.

[10] For a discussion of the mathematics behind GPs, see (Rasmussen 2003) and (Williams & Rasmussen 2006).

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

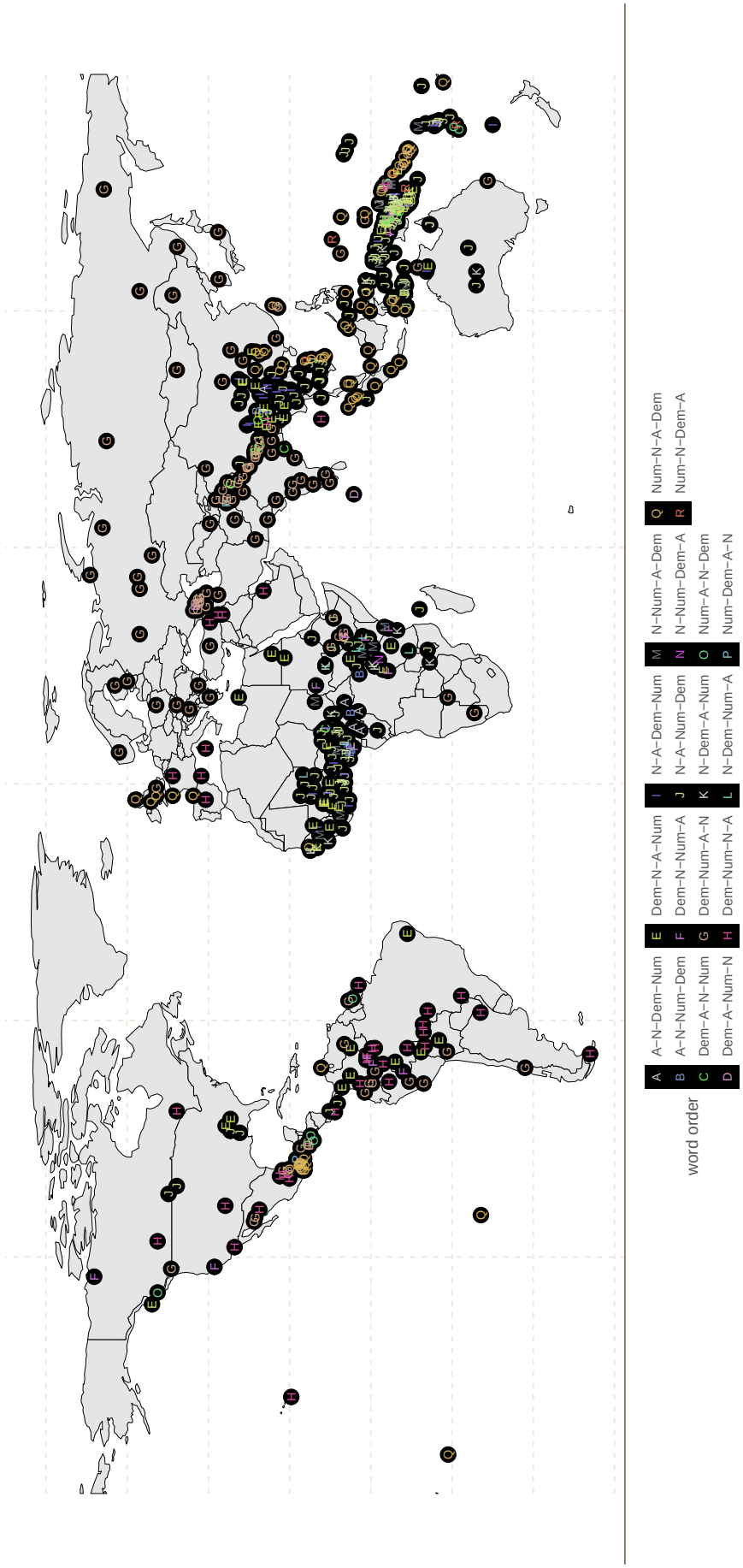

Figure 1: Distribution of word orders

Certain areal patterns start to become apparent from this figure 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 an adjusted frequency. The method is described as follows:

> The method used here takes random samples of languages of a given type such that each pair of languages in the sample satisfies two criteria: (i) the two languages are not in the same genus, and (ii) there are at least ten languages between the two languages, where a language X is said to be between a language Y and a language Z if the distance between X and Y and the distance between X and Z are both less than the distance between Y and Z. Languages are randomly added to each sample as long as they conform to these two criteria, until no additional languages can be added that conform to the criteria. I use the mean size of such samples over 10,000 trials as a measure of the frequency of that type. I refer to this metric below as the ADJUSTED FREQUENCY. (Dryer 2018: 803)

The present replication study will focus on this adjusted frequency count. The question is whether a statistical technique for contact and genetic bias control (as described in Section 3.2) can replicate the results obtained by Dryer (2018).

4.2 Original results

There are 24 logically possible orders between demonstrative, numeral, adjective and noun. Out of those, 18 are attested in Dryer's data. Table 1 summarizes the original results, showing the number of languages, genera, and the adjusted frequencies of all the 24 orders. We will not discuss the theoretical implications of Dryer further but return to the findings in Section 4.4.

4.3 Model 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. Bickel 2011, Guzmán Naranjo & 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 of how common a given value is across languages. Suppose we want to explore the proportion of languages which are predominantly OV vs. languages which have VO word order. We can code our data with OV = 1 and VO = 0, and fit a logistic regression to it including a phylogenetic term and GP. The intercept of the model corresponds to the expected value when all other predictors are set to 0 (i.e. when we ignore their effect). This expected value is effectively the expected proportion of 1s in the data, after we have accounted for phylogentic and areal effects.

This type of counting model can be extended to multiple outcomes by replacing the logistic model with a categorical model. In that case, we can estimate the expected proportion of each category after having controlled for areal and genetic correlations.[12] 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). The intercepts of the categorical model can be used to estimate the expected proportion of each predicted category after we have controlled for

[12]See the supplementary materials for the implementation.

word order	N languages	N genera	adjusted frequency
A–Dem–N–Num	0	0	0
A–Dem–Num–N	0	0	0
A–N–Dem–Num	5	3	2.50
A–N–Num–Dem	5	3	3.00
A–Num–Dem–N	0	0	0
A–Num–N–Dem	0	0	0
Dem–A–N–Num	12	7	5.34
Dem–A–Num–N	3	2	2.00
Dem–N–A–Num	53	40	29.95
Dem–N–Num–A	12	10	9.75
Dem–Num–A–N	113	57	35.56
Dem–Num–N–A	40	32	22.12
N–A–Dem–Num	36	19	14.80
N–A–Num–Dem	182	85	44.17
N–Dem–A–Num	13	11	9.00
N–Dem–Num–A	8	6	5.67
N–Num–A–Dem	11	9	9.00
N–Num–Dem–A	1	1	1.0
Num–A–Dem–N	0	0	0
Num–A–N–Dem	8	5	4.0
Num–Dem–A–N	2	2	2.0
Num–Dem–N–A	0	0	0
Num–N–A–Dem	67	27	14.54
Num–N–Dem–A	5	3	3.00

Table 1: Results of Dryer (2018)

phylogenetic and areal bias. In order to compare the methodological robustness of the results in an even more detailed way, we fitted the following four models:

1. a model without any controls for potential biases (`m_base`)
2. a model with controls for contact effects (`m_gp`)
3. a model with controls for phylogenetic effects (`m_phylo`)
4. a model with controls for both contact and phylogenetic effects (`m_gp+phylo`)

4.4 Results of the replication study

The results of the first replication study are given 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 95% (bold) and 50% (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.[13] There are four important observations that we can take from Figure 2. First, the estimates of `m_base` are essentially the same as the observed values, although including some uncertainty. This works as a sanity check

[13]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.

that the model is performing as expected.

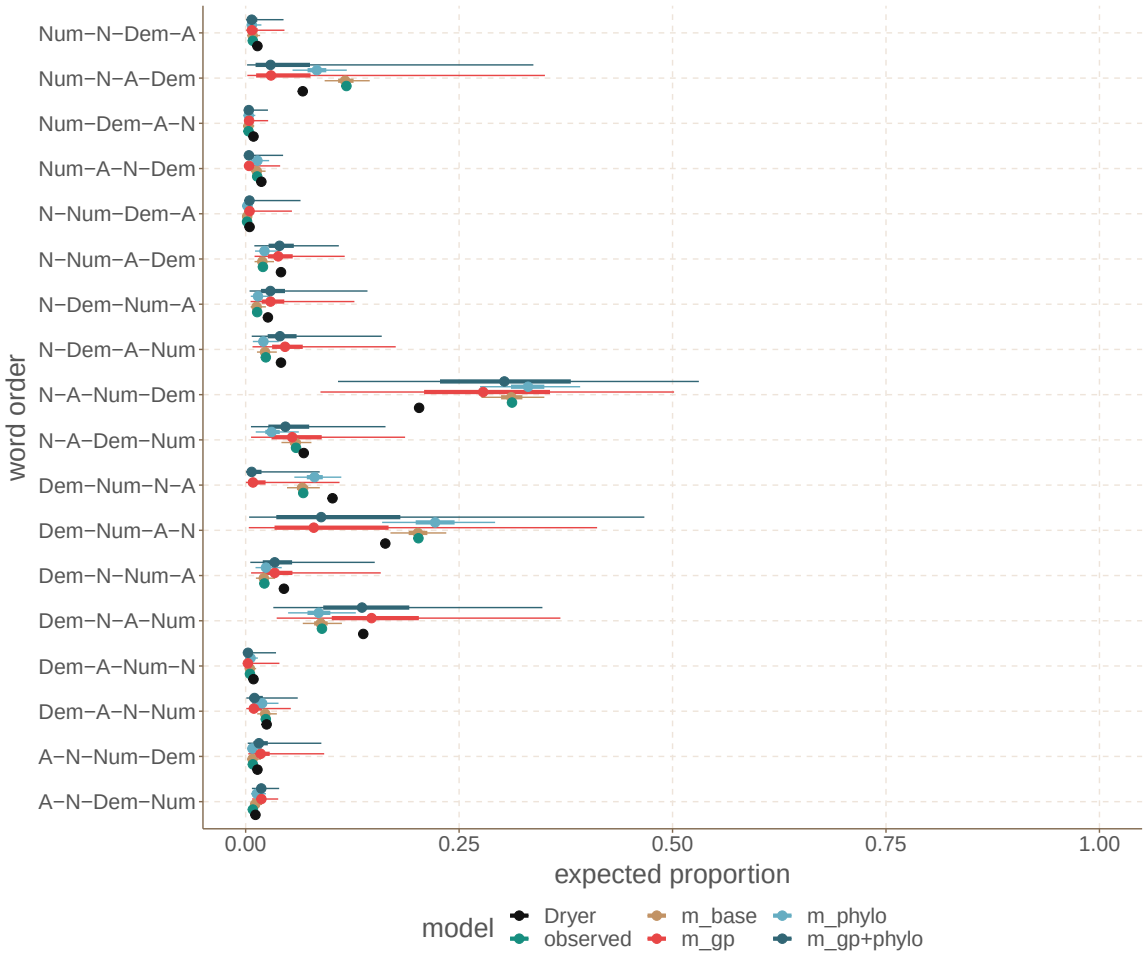

Figure 2: Original and replication results

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.

Third, if we compare the estimates of the full model `m_phylo+gp` with the observed proportions, we see that the observed proportions seem to be slightly biased, particularly for the orders Dem–Num–N–A, Num–N–A–Dem and Dem–N–A–Num. In both 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 these two cases, the observed proportions likely underestimate the actual proportions, as the estimates produced by `m_phylo+gp` are higher. 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 clearly 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. In other words, larger uncertainty intervals do not mean that the models are "worse" than the ones with smaller uncertainty intervals.[14] 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. This should rather be interpreted as follows: 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 for contact bias control 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 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 as value "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 (2018)'s 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

[14] 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 material for the result of the model comparisons, and see Section 5 for a more detailed description of the technique used.

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.

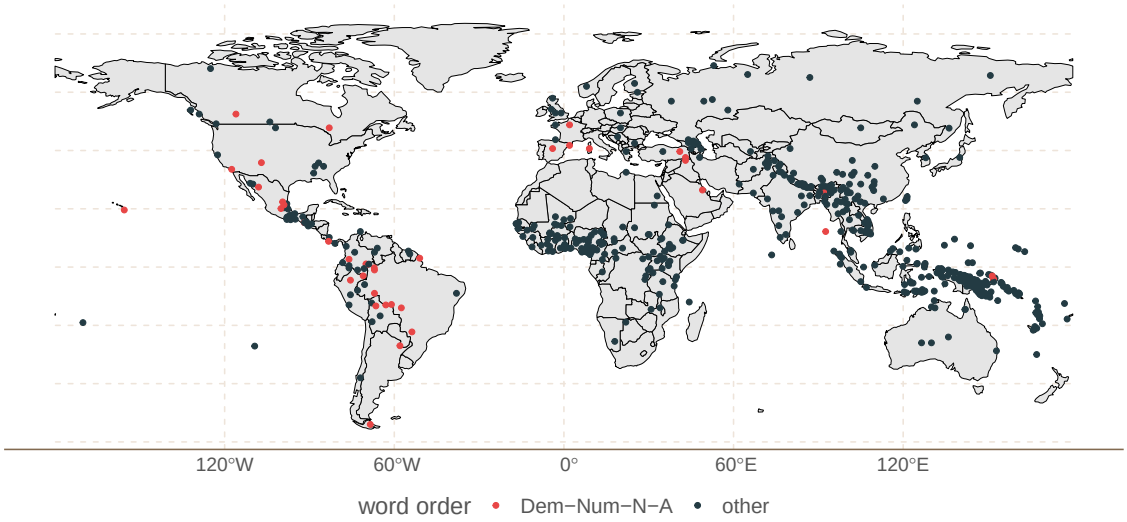

Figure 3: Distribution of Dem–Num–N–A vs. other word orders

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

4.5 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` that we fitted in this 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 in particular, and about the robustness of the results in general.

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 retension 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.[15] In total, his sample includes 150 languages from those six families. Table 2, a reproduction of Table 2 in Seržant (2021: 69), illustrates the data of the study with person-number indexes from Indo-European languages together with the reconstructed ones for Proto-Indo-European.

	1SG	2SG	3SG	1PL	2PL	3PL	decay
Proto-Indo-European	*oh$_2$	*e-s-i	*e-t-i	*o-m-es	*e-th$_2$-e	*o-nt-i	
Persian	am	i	ad	im	id	and	0.13
Greek	o	is	i	ume	ete	un	0.14
Macedonian	em	eš	e	eme	ete	at	0.11
Upper Sorbian	u	eš	e	emy	eće	u / ja	0.15
Slovenian	em	i	-	mo	te	so	0.24
German	e	st	t	en	t	en	0.30
Dutch	-	t	t	en	en	en	0.41
English	-	-	s	-	-	-	0.91
French	-	-	-	ǫ	e	-	0.77

Table 2: Indo-European indexes, reproduced from Seržant (2021: 69)

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.[16] 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). The decay measures obtained this way are shown in Table 2 for a number of Indo-European languages.

Although the operationalization of the decay measure proposed by Seržant (2021) could certainly be subject to discussion from both theoretical and methodological perspectives, we simply used the measures of decay as they are for the purposes of the present study. However, we needed to adjust and correct some of the latitude and longitude information in the dataset, as some values were

[15] 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).

[16] Crucially, if a phonological change has no impact on the contrasts in the paradigm, then it is not counted as leading to decay.

missing or conflated (e.g. all Kannada varieties were placed in the same location). We corrected these issues using the information from Glottolog (Hammarström et al. 2022).

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), 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.

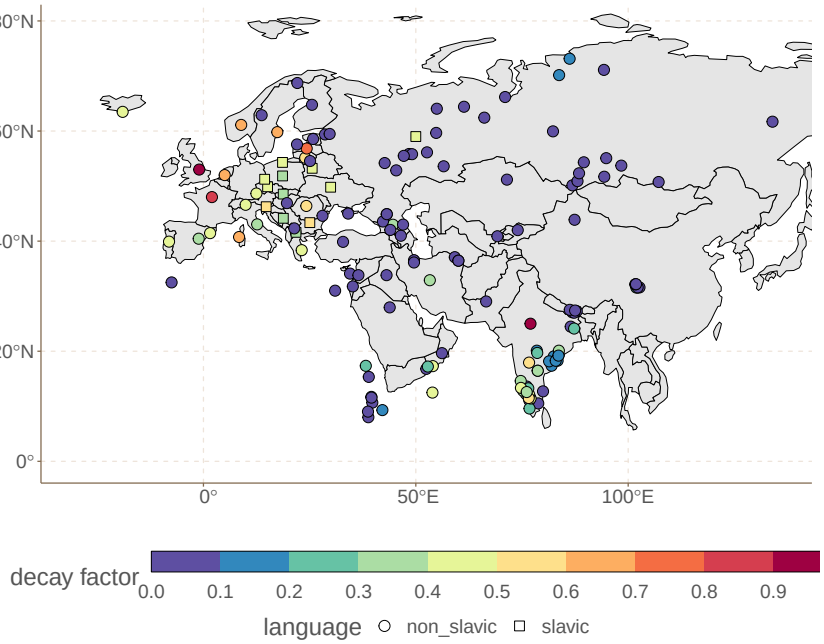

Figure 4: Paradigm decay factor across Eurasian languages

Seržant (2021: 72) mentions two hotbeds of decay, namely Northwestern Europe and India. These can be seen in Figure 4. He then argues that the second 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 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 argues that language contact is an important component of explaining why Slavic languages have preserved so much of their paradigm structure in contrast to other modern Indo-European languages in West and Central Europe:

> While the East-West cline roughly explains the intermediate position of the Slavic languages, it fails to explain why the Slavic decay factor (< Ø 0.15) strongly gravitates towards the Transitional area (Ø 0.12) as well as the languages of the conservative Northeastern Eurasia (Ø 0.05) and is considerably more distant from the Northwestern Europe (Ø 0.61). […] I argue that the particular contact configuration of Slavic is responsible for this skewing: Slavic languages were in much more intensive contact with languages of Northeastern Eurasia and the Transitional area than with the languages of Northwestern Europe. (Seržant 2021: 75-76)

Although Seržant makes a compelling argument regarding the role of contact, he does not perform any systematic statistical tests (or applies other types of controls). It is therefore difficult to evaluate how accurate the analysis really is. Most importantly, there is no principled way of disentangling genetic and areal patterns by visual inspection of a map as in Figure 4. Thus, we cannot exclude that what Seržant analyses as an areal effect to be simply the result of historical developments within these language families, contact playing a negligible role.

5.3 Model 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 five different models:

1. a model with only an intercept and no predictions (`m_base`)
2. a model with a linear effect of longitude, which represents the East-West cline as proposed by Seržant (`m_cline`)
3. a model with a phylogenetic term (`m_phylo`)
4. a model with a GP (`m_gp`)
5. a model with a phylogenetic term and a GP (`m_phylo+gp`)

5.4 Results of the replication study

We first focus on the spatial effects of models `m_cline`, `m_gp` and `m_phylo+gp`. To do so, Figure 5 shows the estimated areal effects of those three 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.

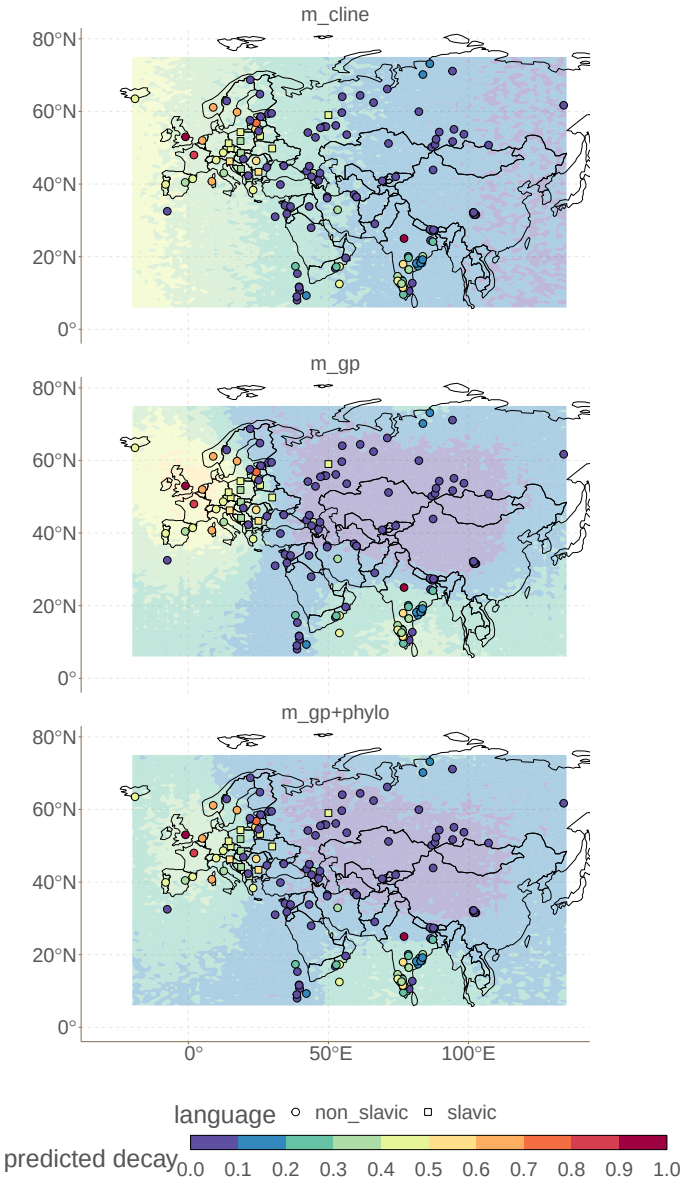

Figure 5: Predicted areal effects

The plot on top 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.

However, once we account for non-linear geographic effects with a GP as in the center and bottom plots in Figure 5, the picture changes substantially. In a way, the predictions of `m_gp` and `m_phylo+gp` 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_phylo+gp` 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_phylo+gp`, 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.

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 is not 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_phylo+gp`. 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_phylo+gp`. The fact that the predicted spatial distribution of decay does not fundamentally change between `m_gp` and `m_phylo+gp` 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_phylo+gp`) should have a better predictive performance than the model which only includes a phylogenetic term (`m_phylo`).

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, Gelman & Gabry (2017). The metric for the comparison is ELPD, the Expected Log pointwise Predictive Density. While it is 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 3 for the five different models. Here, the models are arranged according to their performance, from best at the top to worst at the bottom. Table 3 does not show absolute ELPD values but relative 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 this ELPD 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, Gelman & Gabry 2017).

	ELPD difference	standard error
m_phylo	0.0	0.0
m_phylo+gp	-2.7	3.0
m_gp	-16.3	6.2
m_cline	-58.1	6.4
m_base	-65.8	6.6

Table 3: Approximate LOO-CV

From Table 3, we can conclude that there is no clear difference between m_phylo and m_phylo+gp. 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.[17]

We do observe two important differences, however. First, adding phylogenetic effects to m_gp marks a clear improvement (ELDP difference of 16.3). This suggests that areal effects alone cannot account for the variation in the data. Second, m_cline, which assumes a linear longitudinal effect, performs much worse than the other models. Its performance is similar to the one of m_base, which did not include any predictors. This means that adding longitude information as a 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 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. A larger dataset allowing to include a global decay baseline in the model would be necessary to resolve this issue empirically.

[17] 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 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.

6 Case study: Berg (2020) on gender marking on nouns and pronouns

6.1 Overview of the original study

Berg (2020) studies the association between gender-sensitivity in nouns and gender marking on personal pronouns and possessive determiners. For the latter, there are two possible sources for gender agreement, namely the gender of the posssessor or the gender of the possessum. Two examples to illustrate possessor and possessum marking on a possessive determiner are given in (7) and (8), respectively.

(7) Gender marking of the possessor in Lithuanian (Berg 2020: 526)

a. **jo** širdis
3SG.**M**.POSS heart.F
'his heart'

b. **jos** širdis
3SG.**F**.POSS heart.F
'her heart'

(8) Gender marking of the possessum in Urdu (Berg 2020: 526)

a. us-**kā** **dost**
3SG-**M**.POSS friend.**M**
'his/her (male) friend'

b. us-**kī** **zakeli**
3SG-**F**.POSS friend.**F**
'his/her (female) friend'

For the purposes of his study, Berg distinguishes the following four gender marking categories: (i) gender as a grammatical category of nouns, (ii) gender marking on personal pronouns, (iii) possessor gender marking on the possessive and (iv) possessum gender marking on the possessive.[18] All four gender marking categories are annotated as binary categories by Berg (2020), either as gender marking being present (+) or absent (−). This means that there are 16 different logical patterns. Berg excludes four patterns on theoretical grounds, namely those with no gender on nouns but gender marking on the possessor, which leaves 12 possible patterns of gender marking.

Expanding on Greenberg's universal 43, Berg (2020) addresses several related research questions regarding the correlation between those four gender categories.[19] The empirical part consists of two main analyses, which are used to answer the following two research questions: (i) what is the crosslinguistic distribution of the four gender marking categories?, (ii) what are the predictive relations between the four gender marking categories?

To carry out his study, Berg (2020) built an initial sample, called "language sample", with 500 languages distributed across all 6 macroareas. Furthermore, Berg built a more reduced "genus-sample" of 287 languages by sub-sampling based on the genera classification, selecting one language per genus per gender coding type. Berg (2020: 534) called this the "each-type-once" strategy. Both the full language sample and the genus sample still include a number of languages with no gender marking in any of the four categories. In certain cases, the sample used for analysis excludes those

[18] For the sake of simplicity, we follow Berg in using the term "gender marking", even though nouns, as the controller of gender agreement relations, are required to have gender as a grammatical category (and given rise to gender agreement on other elements in the clause) rather than being marked for gender directly.

[19] Universal 43 states that "[i]f a language has gender categories in the noun, it has gender categories in the pronoun" (Greenberg 1963: 96).

languages without any gender marking. This reduces the language sample to 172 languages and the genus sample to 118 languages with gender marking in at least one of the four categories.

6.2 Original results

Out of the 12 possible combinations of gender marking, 8 types are attested in Berg's dataset. Table 4, adapted from Table 2 in Berg (2020: 540), shows the distribution of those types in the full language sample and the genus sample.

noun	personal pronoun	possessor	possessum	language sample N(500)	language sample %	genus sample N(287)	genus sample %
+	+	+	+	30	6.0	24	8.4
+	-	+	+	0	0	0	0
+	+	-	+	16	3.2	4	1.4
+	+	+	-	65	13.0	44	15.3
+	-	-	+	5	1.0	4	1.4
+	+	-	-	13	2.6	11	3.8
+	-	+	-	0	0	0	0
+	-	-	-	14	2.8	9	3.1
-	-	-	-	328	65.6	169	58.9
-	+	-	-	0	0	0	0
-	-	+	-	0	0	0	0
-	+	+	-	29	5.8	22	7.7

Table 4: Distribution of the different gender marking types in Berg (2020)

Berg (2020: 541) finds that 124 out 143 languages (87%) in the language sample and 83 out of 96 (86%) languages in the genus sample have nominal gender and also code gender in the pronouns. This confirms Greenberg's claim, although as a statistical, rather than an absolute universal.

Regarding the first question Table 5, reproduced from Table 3 in Berg (2020: 541), shows the probability of gender marking in the four different categories in the language sample (N=500, including languages without gender). Berg (2020: 541) performs several chi-square tests on this table and finds no statistical difference in proportions for nouns, personal pronouns and possessors. He does report on a statistically significant difference between gender marking of the possessum and all other categories.

category	probability
noun	0.334
personal pronoun	0.366
possessor	0.314
possessum	0.111

Table 5: Proportion of gender marking

To examine the second question about the predictive relations between the four gender marking categories in more detail, Berg proposes a method to calculate what he calls "rate of gender matches". This metric is based on matches and mismatches of gender marking between a pair of categories. Note that the order of the categories is relevant and indicated by a "$\rightarrow$" here. Matches are split into two types. In the case of plus matches ($+ \rightarrow +$), both categories have gender marking. In the case

of minus matches ($- \rightarrow -$), gender marking is missing in both categories. Mismatches correspond to $+ \rightarrow -$ and $- \rightarrow +$ patterns between a pair of categories.

The rate of gender matches is then calculated as the proportion of matches in gender marking divided by the number of mismatches of type $+ \rightarrow -$. Berg (2020) furthermore distinguishes two types of match rates: "plus match rates" based on matches with gender marking being present, and "minus match rates" with gender marking being absent. To give an example, on the basis of the genus sample the rate of plus matches from noun to personal pronoun corresponds to $N_{++}/(N_{++} + N_{+-}) = 83/(83 + 13) = 0.86$. The rate of minus matches from noun to personal pronoun then is $N_{--}/(N_{--} + N_{-+}) = 169/(169 + 22) = 0.86$. The rates of plus and minus matches for different pairs of categories are presented in Table 6, reproducing Tables 5 and 6 in Berg (2020: 543, 545).

categories			plus match rate	minus match rate
noun	$\rightarrow$	personal pronoun	0.86	0.88
noun	$\rightarrow$	possessor	0.71	0.88
noun	$\rightarrow$	possessum	0.33	/
personal pronoun	$\rightarrow$	possessor	0.82	1.0
personal pronoun	$\rightarrow$	possessum	0. 27	0.98
possessor	$\rightarrow$	possessum	0.27	0.96
possessum	$\rightarrow$	possessor	0.75	0.74
possessum	$\rightarrow$	personal pronoun	0.88	0.70
possessum	$\rightarrow$	noun	/	0.75
possessor	$\rightarrow$	personal pronoun	1.0	0.92
possessor	$\rightarrow$	noun	0.76	0.86
personal pronoun	$\rightarrow$	noun	0.79	0.93

Table 6: Match rates between pairs of gender marking categories (Berg 2020: 543)

The plus match rate from possessum to noun and the minus match rate from noun to possessum are excluded by definition, as gender marking cannot be present in the possessum category but absent in the noun category. Berg (2020) states the following as to how the numbers in Table 6 should be interpreted: "The values range from 0 to 1. In view of the binary distinction between gender marking and the lack thereof, chance is at 0.5. Values above 0.5 indicate a positive correlation (facilitation) while values below 0.5 indicate a negative correlation (inhibition)." These numbers are no correlations in the strict sense; they correspond to the percentage of languages that mark both categories out of the number of languages that mark the left category in Table 6 (in the case of the plus match rates). If the proportion is higher than 0.5, it shows that most of the languages marking the left category also mark the right category (since the mismatch count is low). A proportion lower than 0.5, on the other hand, indicates that less than half of the languages that mark gender in the left category also mark it in the right category.

For the plus matches of the upper half of Table 6, Berg (2020: 544) notes that the rate of matches decreases from top to bottom. The author concludes "that the power of gendered nouns to predict gender marking diminishes from personal pronouns to possessors to possessums. The values in rows 1 to 3 decrease monotonically as the distance between the gender categories in [Table 4] increases. The low value in row 3 shows that gendered nouns do not facilitate gendered possessums." For the

lower half, he interprets the overall high values as showing that gender marking on possessums and possessors generally predict gender marking on the other elements. He also concludes that gender marking on personal pronouns is a good predictor of gender-sensitivity in nouns.

For minus matches, the numbers in Table 6 show the proportions of languages in which both categories are absent out of all languages in which the left category is absent. Berg interprets this as follows: "the values for the minus matches hover at a consistently high level in all pairwise comparisons. Thus, a minus sign is strongly predictive of a minus sign elsewhere. To put it differently, if one category is gender neutral, it is highly unlikely for another to be gender-marked" (Berg 2020: 544).

6.3 Model of the replication study

An additional challenge of this study concerns the original data used by Berg. The paper includes the list of languages and values for the 172 languages with gender marking of Berg's original sample of 500 languages. The list, however, does not have glottocodes (or language names corresponding to the ones used in Glottolog). Since the modeling techniques we use in this study require the genetic and location information of the languages, we manually matched the languages with their glottocodes to add this information to the sample. During that process, we encountered an issue with the languages Ngankikurungkurr and Ngan'gityemerri. Glottolog treats them as two varieties of the same language with no separate location or ID information. For the sake of comparability with the original study, we opted for keeping both languages in and assigned them the same glottocode (nang1252) as well as location instead of removing one of the languages.

Another, and more serious, issue of the sample concerns the remaining 328 languages with no gender marking in Berg's original sample of 500 languages. These languages are not openly accessible. This is a problem, since Berg (2020) uses the genus sample including languages with no gender marking for the relevant analyses replicated in this study. After contacting the author, he kindly sent us a hand-written list of the non-gendered languages. However, this list only contained 285 languages with no gender marking instead of the total of 328 languages as shown in Table 4, making it incomplete, as far as we can tell. Combining both lists of languages with and without gender marking, we only have a total of 457 languages instead of 500. We decided to use Berg (2020) as a case study for replication nevertheless, as it shows how difficult full replicability can be in practice, despite the sample being openly accessible at first sight.

The first question that we address in our replication study concerns the crosslinguistic distribution of the four gender marking categories, once phylogenetic and contact biases are controlled for. Comparing the results of the expected proportions is straightforward and we can use a similar kind of model as in the first case study (cf. Section 4.3). In this case the model is simpler, though, as we are dealing with only two values, namely the presence vs. absence of gender marking. We can therefore use logistic regression, which estimates the probability of successes in a series of repeated single trials with a binary outcome (e.g. presence vs. absence of gender marking). We fitted the following series of four models to each gender marking category:

1. a model with an intercept and no other predictors (`m_base`)
2. a model with phylogenetic control (`m_phylo`)
3. a model with contact control (`m_gp`)
4. a model with phylogenetic and contact controls (`m_phylo+gp`)

We used the reconstructed full sample for our models (N=457), which we will refer to as the "full$_{457}$

sample". To ensure comparability between our results and the original results concerning the first question, we scaled the estimated proportions to the original full sample size of 500.

For the second question about the predictive relations between categories, we fitted a series of logistic regression models for pairwise prediction of the gender marking probability of one category from another category. We used two types of models: one with no additional controls (`m_base`) and one with both phylogenetic and contact controls (`m_phylo+gp`). While Berg's original analysis using the match rate metric is based on the genus sample including the languages with no gender marking, we do not think this choice translates to our models in an evident way. For the sake of better comparability, we opted for the solution of including languages without gender marking. The model series was thus fitted to the full_{457} sample.

6.4 Results of the replication study

Before turning to the model results, Figure 6 shows the areal distribution of the presence (blue) and absence (red) of gender marking across the four categories in the full_{457} sample. Visual inspection suggests clear areal preferences for gender marking, which appears to be favored in Europe, India, Northeast Africa and Australia. For gender marking of the possessum, we see that it is most common in Europe and India.[20]

Figure 7 shows the model results concerning the proportions of gender marking across the four categories. We see the proportion of gender marking as predicted by `m_base` (beige), `m_phylo` (light blue), `m_gp` (red) and `m_phylo+gp` (dark blue). They are shown with 50% (bold) and 90% (light) uncertainty intervals. The model predictions are further compared to the proportions in Berg's full sample (black) and genus sample (green). In his original study, Berg only analyzes the proportions in the genus sample, as it includes a form of genetic bias control. We calculated the proportions of the full language sample based on the counts given in Table 4 for reasons of comparison. As mentioned above, our model results from the full_{457} sample are scaled to match a sample size of 500. This allows for a direct comparison of our model results to the raw proportions of the unbalanced full sample and to the proportions from the balanced genus sample analyzed in Berg (2020).
As can be seen in Figure 7, the four model estimates are fairly similar for each of the four categories; their 90% uncertainty intervals overlap to a great extent and generally include the observed proportion in the full sample. The mean estimate of `m_base` is very close to the proportions of the full sample in all categories, which works as a sanity check in that the model performs as expected.

Regarding the three models with controls, Figure 7 shows that they predict somewhat lower proportions than `m_base`. Especially `m_gp`, which only adds a contact control, makes comparatively low predictions for the proportions. The likely explanation for this is that the GP estimates that there is a heavy areal bias for the presence of gender marking in the relevant categories. In other words, contact is taken to account for much of the occurrence of gender marking. This is what appeared as an areal effect on the maps in Figure 6. The `m_phylo+gp` model, however, points to slightly higher proportions. As `m_phylo+gp` controls for phylogenetic effects as well, it could be that parts of the variation identified as an areal effect by `m_gp` may also be accounted for by phylogenetic structures.

Furthermore, we see in Figure 7 that the uncertainty intervals are fairly large for possessor, personal pronoun and noun, including the ones of `m_base`. This suggests that there simply is a high degree of variation with no straightforward pattern emerging. Gender marking of the possessum

[20] Of course, Europe and India are strongly associated with Indo-European and Northeast Africa with Cushitic, which means that these patterns may not be purely areal effects.

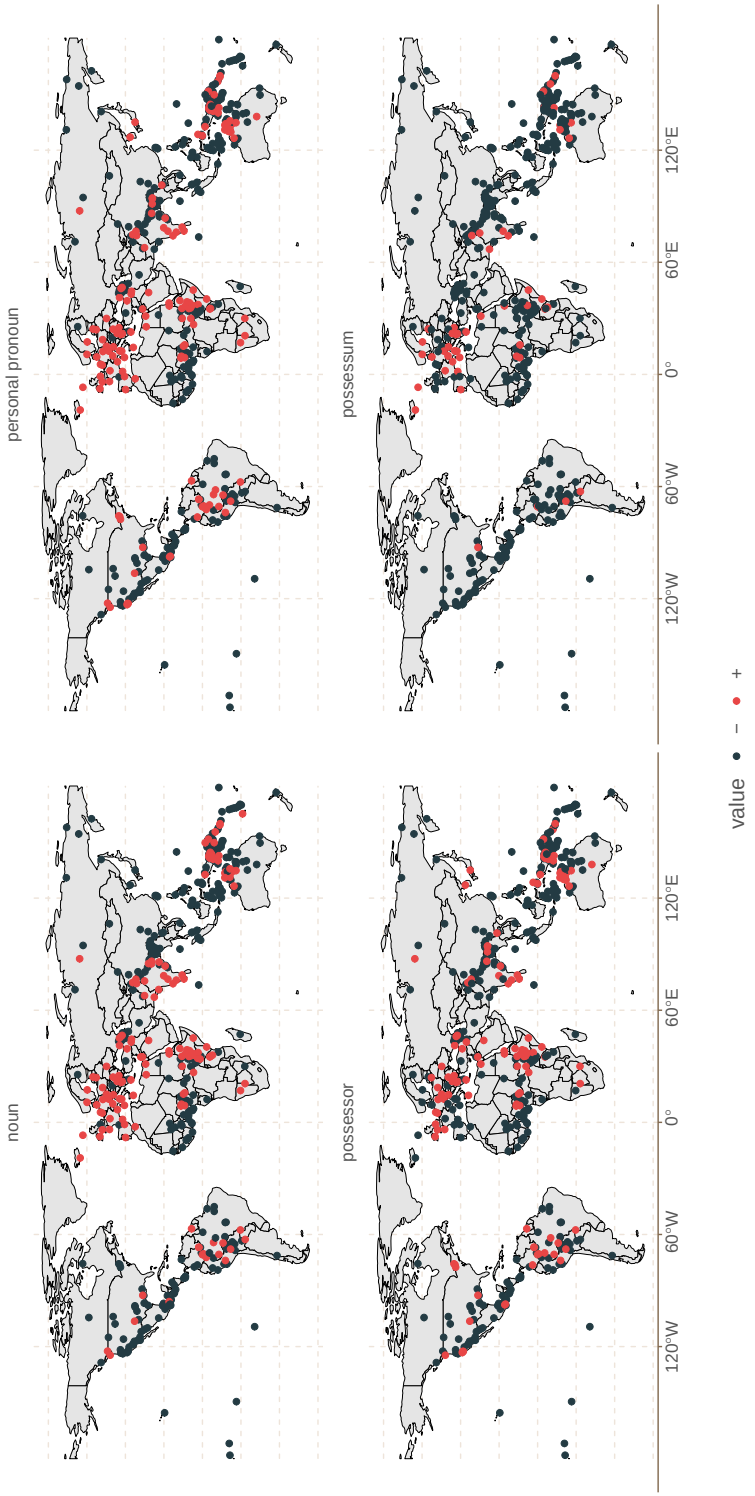

Figure 6: Areal distribution of gender marking

is estimated to be much less common, which may also be the reason for less uncertainty around its prediction.

The next step is to compare our models with controls with the original proportions reported in Berg (2020) based on the genus sample. Across all four categories, Figure 7 shows that the mean estimated proportions are corrected to a smaller value by the three models, while Berg's corrected proportions in the balanced genus sample are systematically corrected towards a higher value. Put

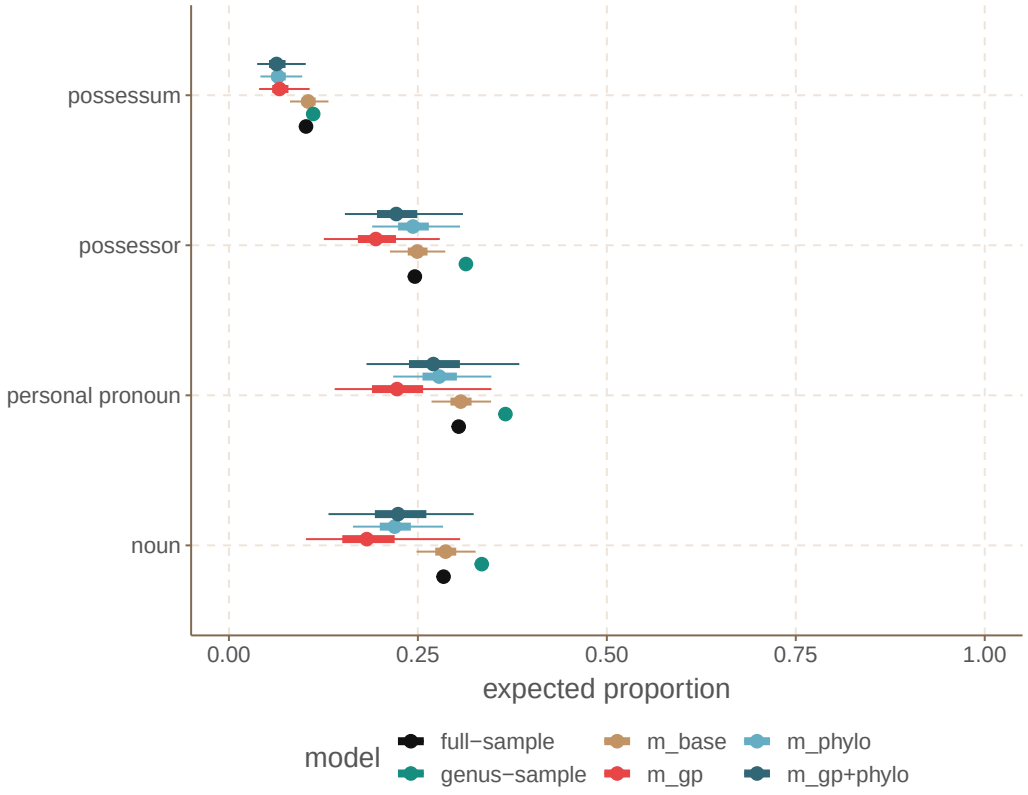

Figure 7: Expected proportions - all languages

differently, our method suggests expected proportions that are systematically lower than the ones reported in the original study. In the case of possessum, possessor and noun marking, Berg's genus sample proportions fall outside of the 95% uncertainty intervals of `m_gp+phylo`.[21] Berg's genus sample clearly overestimates the proportions of gender marking with respect to the proportions estimated by the models. We think a possible explanation for this lies in the fact that most of the languages in the full language sample lack gender marking all together (328 out of 500 languages, i.e. 66%). The second most frequent type is gender marking in the noun, personal pronoun and possessum in only 65 languages (13% of the full sample). Thus, the total absence of gender marking is very common in the sample, and we can assume that it is fairly consistent within genera. At the same time, we can also expect a small degree of variation within genera with gender marking across the four different categories. As mentioned above, Berg's genus sample includes more than one language per genus if they represent different types. Therefore, the sampling method used by Berg could have led to a higher rate of exclusion for languages with no gender marking as opposed to languages with some form of gender marking. This would explain why the proportions of gender marking are consistently higher in the genus sample than in the language sample. What is interesting is that our model results rather pattern with the full language sample and not the genus sample. This suggests that the genus sampling method as applied by Berg produces somewhat biased results in terms of absolute proportions of gender marking for the four categories. In terms of relative differences between categories, on the other hand, our method could replicate the overall pattern found by Berg (2020).

We now turn to the second question about the predictive relations between the four gender

[21] The category of possessum is likely not affected as much, because the languages with gender marking in this category are much less frequent than for the other categories.

marking categories. As described in Section 6.2, Berg uses his own metric of plus and minus match rates to address this question. For replication, we fitted a series of models for pairwise prediction of the probability of gender marking in one category based on another gender category. We used two types of models, one with no additional controls (`m_base`) and one with both phylogenetic and contact controls (`m_phylo+gp`). The results are shown in Figure 8 for the predicted probabilities of the presence (+) and absence (−) across the rows for the four categories.

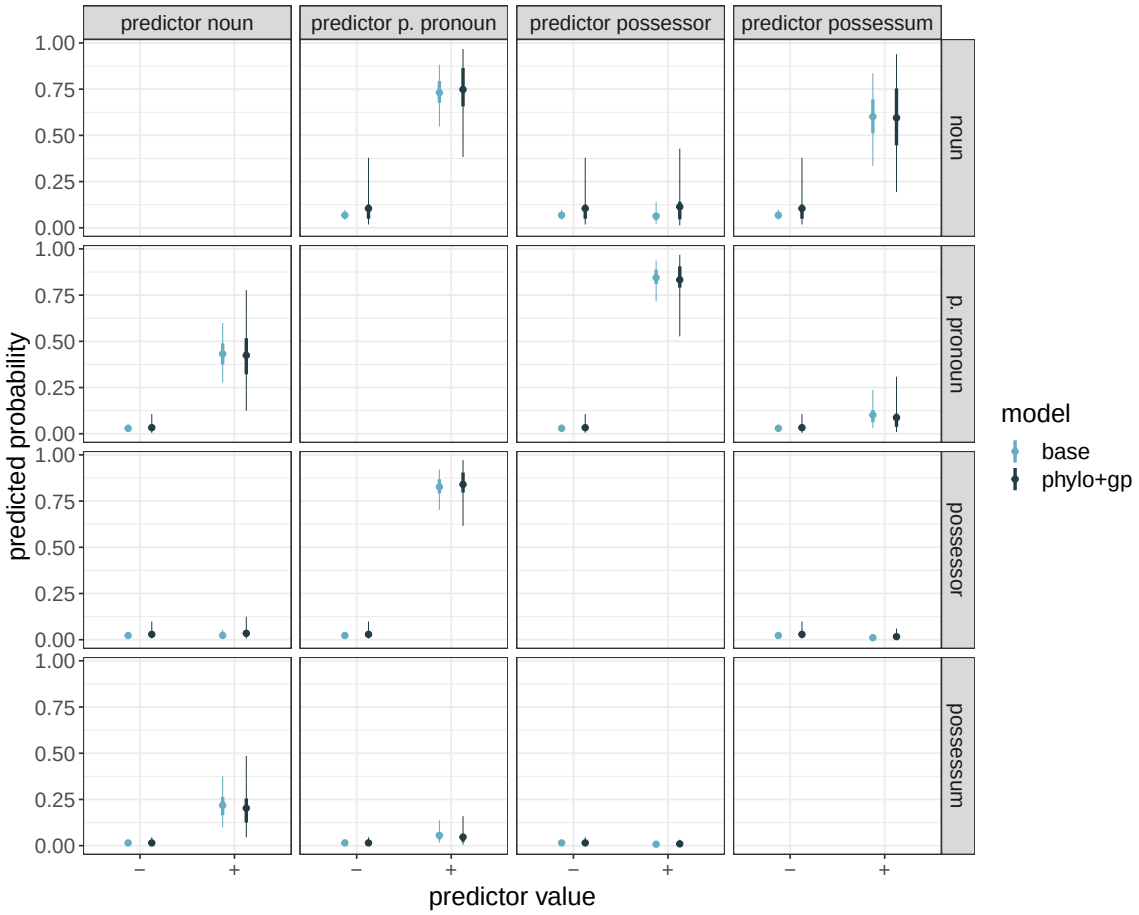

Figure 8: Predictive relations between categories in the full$_{457}$ sample

The columns represent the predictors. The results of `m_base` are marked in light blue, the results of `m_phylo+gp` in dark blue. The predictions are given together with their 50% (bold) and 90% (light) uncertainty intervals. Since the results in Figure 8 represent probabilities, they cannot be compared to the rates of matches as calculated by Berg (2020) in a direct way. Our models estimate the effect of the presence and absence of gender marking in the predictor category at the same time, and the interpretable result is the difference between their effects on the probability of gender marking in the dependent category. Because each pair of categories represents a separate model, we cannot compare them directly. Such comparisons, however, are performed by Berg (2020). Therefore, we will briefly summarize our results and compare them to the original results in a more conceptual way.

The first important observation in Figure 8 is that the mean estimates of `m_phylo+gp` are very similar to those of `m_base`. The important difference is that the former has larger uncertainty intervals. This means that some of the variation is accounted for by phylogenetic and contact relations, increasing our uncertainty about the real probabilities. The results of `m_phylo+gp` suggest that the

presence (as opposed to the absence) of gender marking in nouns increases the probability of gender marking in personal pronouns and the possessum (with a much smaller effect). The presence of gender marking in the personal pronoun has an effect on nouns and possessors, the latter of which is very strong and holds vice versa. Possessors do not seem to have an effect on any category other than personal pronouns, and possessums do not show any clear effect in `m_phylo+gp` at all.

To compare our results to the original ones in Berg, Table 7 summarizes the strong predicting relations as reported by Berg using the plus matches metric.

effects reported in Berg (2020)			replicated
p. pro	→	noun	✓
p. pro	→	possessor	✓
possessor	→	p. pro	✓
nouns	→	p. pro	(✓)
nouns	→	possessor	✗
possessum	→	possessor	✗
possessum	→	p. pro	✗
possessor	→	noun	✗

Table 7: Replication of predictive relations between gender marking categories

As Table 7 shows, Berg (2020) found 8 strong positive match rates in the pairwise comparisons, which he analyzed as strong predictive relations. Our models, however, only replicated clear effects for four of those relations, with the presence of gender marking nouns having a very small effect on personal pronouns.

6.5 Taking stock

This case study showed that data transparancy is essential. Although a dataset was published with the article by Berg (2020), a closer look revealed a number of inconsistencies which made it much harder to replicate the study and to interpret the robustness of the original results. Both parts of our replication study showed that using statistical techniques led to somewhat different results than in the original study. While we could replicate parts of the results in Berg (2020), there were a number of differences between the original and our results. As for the proportions of gender marking categories, the differences between the original and our results are likely due to the choices of building the balanced genus sampling in Berg (2020). We do not have a good explanation for the differences between Berg's results and ours for the second part, and we conclude that those findings should be subject to further, more detailed theoretical analysis. What can be taken away from this case study is the importance of evaluating the methodological robustness of typological studies, as it helps to estimate how confident we can be about results in the literature.

7 Discussion

7.1 Towards better replicability in typology

The three case studies presented in this paper 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, Hyman & Staros (2006). Adaping 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:

(9) *Transparency requirement of typological studies*

a. primary data collection
b. secondary data collection / language sample
c. data analysis & annotation
d. (statistical) methods

We will not address primary data collection further for two reasons. First, it is often less relevant for large-scale, quantitative typological studies such as the ones presented here. Second, there is substantial work from the language documentation literature that addresses transparency standards and provides guidelines for primary data collection (cf. Gawne & Berez-Kroeker 2018, Himmelmann 1998, Maxwell 2012).[22] Since such a discussion is still lacking for the other three levels shown in (9), we provide concrete suggestions in the appendix for best practices for full transparency and replicability in typology. The guidelines in the appendix include a discussion on how the language sample can be documented in a more transparent way, what could be the gold standard of transparent linguistic analysis and annotation, as well as how the code for the statistical analysis can be shared in a transparent and accessible way.

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 similar poor statistical practices have been problematized in the linguistic literature (cf. Sönning & Werner 2021), we think that 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. Both Dryer (2018) and Berg (2020) used sampling methods in order to control for phylogenetic and, in the case of Dryer (2018), contact effects. Our results generally replicate the ones 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. Berg (2020) reported results that appeared to overestimate the linguistic effect (proportions of gender marking) in contrast to our results. We argued that this may be due to the specific sampling method employed by Berg (2020). 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

[22] Also see Berez-Kroeker et al. (2018) on data citation standards in linguistics in generally.

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.

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. In addition, most of the original studies sparking discussions and replications involved strong claims that do not fit in with the general expectations in the field. 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. In our three replication case studies, we showed that some results are replicable using a different statistical method, while others are not. Since there is no good way of knowing which results are methodologically robust and which are not, the methodological robustness of typological results should be evaluated more systematically in typology.

7.2.2 The advantage of statistical bias control

In addition to the general point that typological studies should be systematically evaluated for methodological robustness, the present study 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.[23] 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.[24] 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, Davidson & Bates (2008), Jaeger (2008), Vasishth et al. (2018) for psycholinguistics, Gries (2015), Paquot & Plonsky (2017) for corpus linguistics, Larson-Hall & Herrington (2010), Plonsky (2015) for second language acquisition, Aguilar-Sánchez (2014), Tagliamonte & Baayen (2012) for sociolinguistics and Roettger (2019), Roettger, Winter & Baayen (2019) for phonetics. Moreover, Winter & Grice (2021) offers 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 a previous studies (cf. Section 2.3). Examples are Hartmann (2022), Jaeger et al. (2011), Roberts, Winters & Chen (2015), who showed that including some form of statistical control 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 re-

[23] We do not mean to imply that researchers knowingly apply tests that are not defined for the situations used. This rather happens inadvertently because the assumptions of statistical tests can be very complex, and because of fairly lax methodological traditions, at least in linguistics, where many researchers do not receive proper statistical training.

[24] Cf. Berkson (1942), Cohen (1994), Cumming (2012), Greenland et al. (2016), Kline (2013), Meehl (1967), Nickerson (2000), Ziliak & McCloskey (2008) for more details.

sults 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 & Grice (2021) who discuss common levels of non-independence in different linguistic sub-disciplines).

In particular, we found that the manual sampling method used by Berg (2020) likely overestimated the proportions of gender marking languages, as he found a number of predictive relations between gender marking categories that we could not replicate with our models. It is important to note that Berg's sampling method only controlled for phylogenetic but not for contact bias. 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.

Those two case studies thus suggest 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 doing away with building smaller sub-samples. It allows to keep all datapoints in, and the model can make use of the information about dependencies between languages.[25]

The second 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. The results of the model with a GP do not show any clear East-West cline. It is therefore plausible that this cline is an artifact of a more complex non-linear contact effect. This shows that a statistical tool to control and 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 replication 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 & Gelman (2021), we propose to accept uncertainty in statistical 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 little informative. In fact, the model that captured the variation in the data best also had to largest uncertainty intervals around the means of the predictions.[26] The high

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

[26] See supplementary materials for a thorough comparison.

degree of uncertainty about the expected proportions in the first and third case study (cf. Sections 4 and 6) are the result of 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, performing better as 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 three replication studies also served as examples of how a statistical analysis in typology can focus more on estimation. We therefore fully agree with Vasishth & Gelman (2021), who note:

> The most difficult idea to digest in data analysis – and one that is rarely taught in linguistics and psychology – is that conclusions based on data are almost always uncertain, and this is regardless of whether the outcome of the statistical test is statistically significant or not. This uncertainty can and must be communicated when addressing questions of scientific interest. The perspective we take is that the focus in data analysis should be on estimation rather than (or only on) establishing statistical significance or the like [...] (Vasishth & Gelman 2021: 1320)

8 Concluding remarks

The present study 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 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, but we have shown that they capture more of the real relations between languages and their effects on linguistic structure. 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.

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

Baayen, Harald, Douglas Davidson & Douglas Bates. 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, Tyler Heston, 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. *STUF-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. 2011. Statistical modeling of language universals. *Linguistic Typology* 15(2). 401–413.

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

Bürkner, Paul-Christian. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1). 1–28.

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

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 ($p < .05$). *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. 513.

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

Cysouw, Michael, Dan Dediu & Steven Moran. 2012. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335(6069). 657–657.

de Villemereuil, Pierre & Shinichi Nakagawa. 2014. *Modern phylogenetic comparative methods and their application in evolutionary biology*. Berlin: Springer.

Donoho, David L. 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. International Journal sponsored by the Foundation "Foundations of 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. 2013. Order of object and verb. In Matthew Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

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 Blasí & Seán Roberts. 2016. Language evolution and climate: the case of desiccation and tone. *Journal of Language Evolution* 1(1). 33–46.

Foster, Joseph F. & Charles A. Hofling. 1987. Word order, case, and agreement. *Linguistics* 25(3). 475–500.

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 N., Daniele Fanelli & John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8(341). 341ps12.

Gould, Julius & William Kolb (eds.). 1964. *A dictionary of the social sciences*. London: Tavistock Publications.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg (ed.), *Universals of Language*, 73–113. Cambride, MA: MIT Press.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman & Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31. 337–350.

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.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.

Guzmán Naranjo, Matías & Miri Mertner. 2022. Estimating areal effects in typology: a case study of African phoneme inventories. *Linguistic Typology* aop.

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 & 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. Re-doing typology. *Linguistic Typology* 10(1). 74–82.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. 36(1). 161–196.

Hüffmeier, Joachim, Jens Mazei & Thomas Schultze. 2016. Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology* 66. 81–92.

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).

Larson-Hall, Jenifer & Richard Herrington. 2010. Improving data analysis in Second Language Acquisition by utilizing modern developments in applied statistics. *Applied Linguistics* 31(3). 368–390.

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.

Maxwell, Mike. 2012. Electronic grammars and reproducible research. In Sebastian Nordhoff (ed.), *Electronic grammaticography*, 207–235. Honolulu: University of Hawai'i Press.

Meehl, Paul E. 1967. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34(2). 103–115.

Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language*. 56–119.

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

Nickerson, R. S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* 5(2). 241–301.

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.

Plonsky, Luke (ed.). 2015. *Advancing quantitative methods in second language research*. New York: Routledge.

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

Rasmussen, Carl Edward. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

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

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).

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

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.

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

Siewierska, Anna & Dik Bakker. 1996. The distribution of subject and object agreement and word order type. *Studies in Language* 20. 115–61.

Sinnemäki, Kaius. 2010. Word order in zero-marking languages. *Studies in Language* 34(4). 869–912.

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 Harold Greenberg, Charles Albert Ferguson & Edith A. 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.

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). 657–657.

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* 1. 1–69.

Wang, Chuan-Chao, Qi-Liang Ding, Huan Tao & Hui Li. 2012. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335(6069). 657–657.

Widmann, Thomas & Peter Bakker. 2006. Does sampling matter? A test in replicability, concerning numerals. *Linguistic Typology* 10(1). 83–95.

Williams, Christopher KI & 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.