Appendix of "Replication and methodological robustness in quantitative typology"

A Results of Dryer (2018)

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)

B Indo-European indexes in Seržant (2021)

	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)

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

C.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 (1) and (2), respectively.

(1) 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'

(2) 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. For the sake of consistency, 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. 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.[1] 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 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.

C.2 Original results

Out of the 12 possible combinations of gender marking, 8 types are attested in Berg's dataset. Table 3, 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)	%	genus sample N(287)	%
+	+	+	+	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 3: 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 4, 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

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

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 4: 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 5, 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 5: 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 5 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 5 (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 5, 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 3] 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 5 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).

C.3 Models 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 3, 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 ??). 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.

C.4 Results of the replication study

Before turning to the model results, Figure 1 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.[2]

Figure 2 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 3 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 2, the four model estimates are fairly similar for each of the four categories;

[2]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.

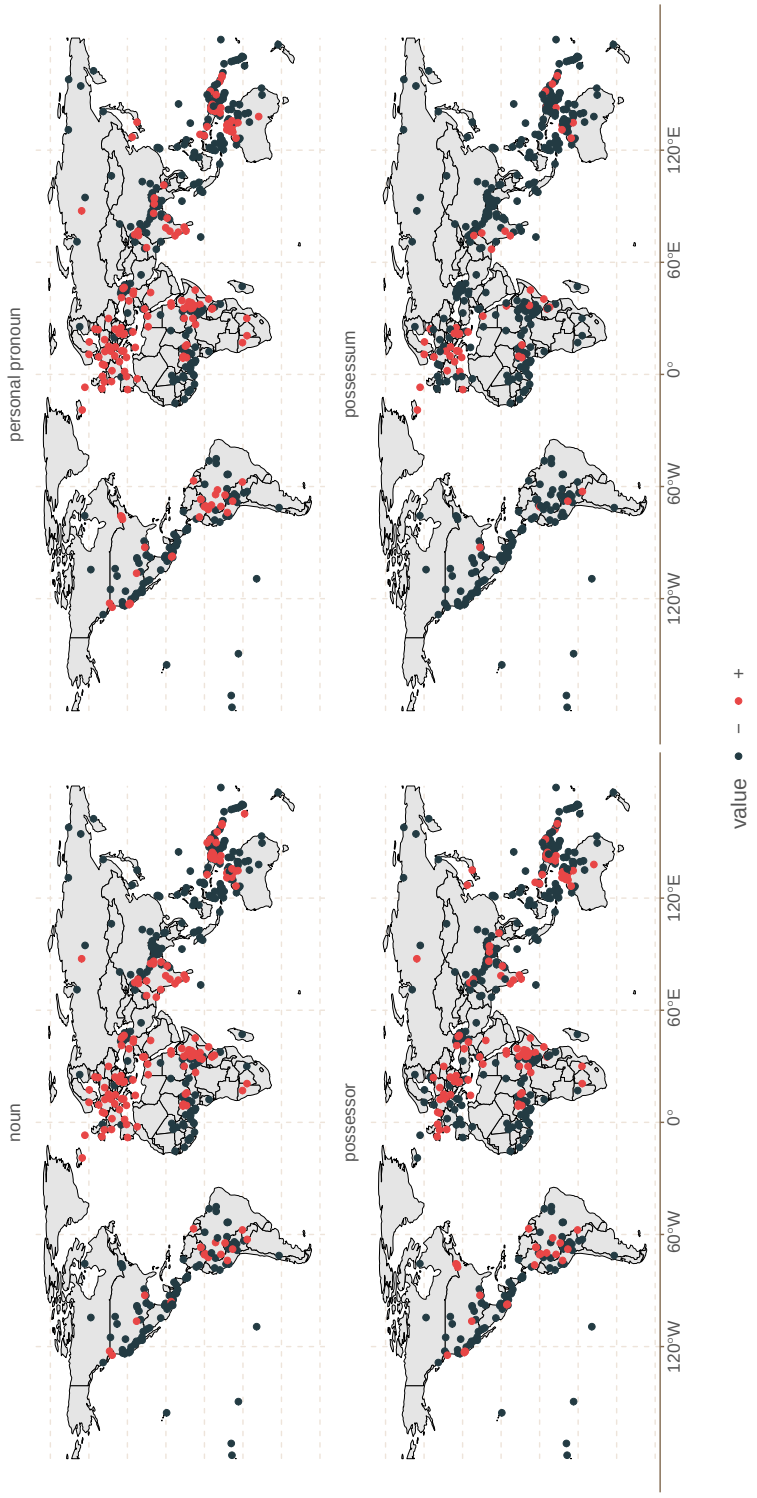

Figure 1: Areal distribution of gender marking based on Berg's (2020) sample

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 2 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

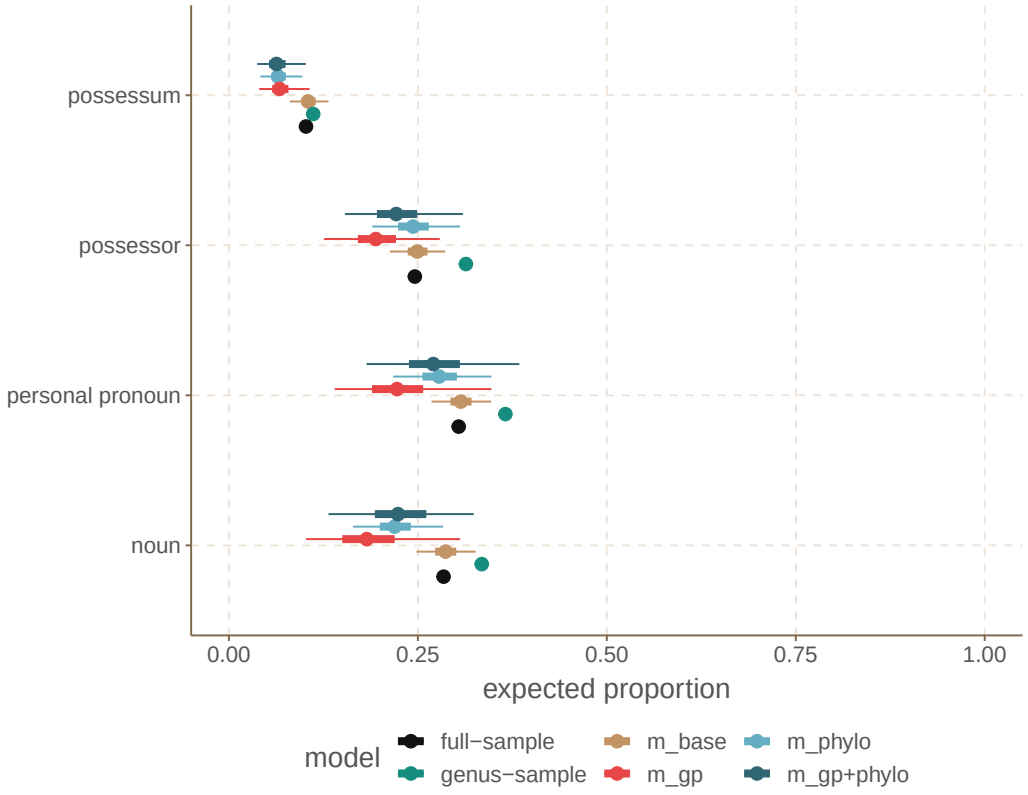

Figure 2: Expected proportions - all languages

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 1. 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 2 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 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 2 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 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`.[3] 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

[3]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.

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 marking categories. As described in Section C.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 3 for the predicted probabilities of the presence (+) and absence (−) across the rows for the four categories.

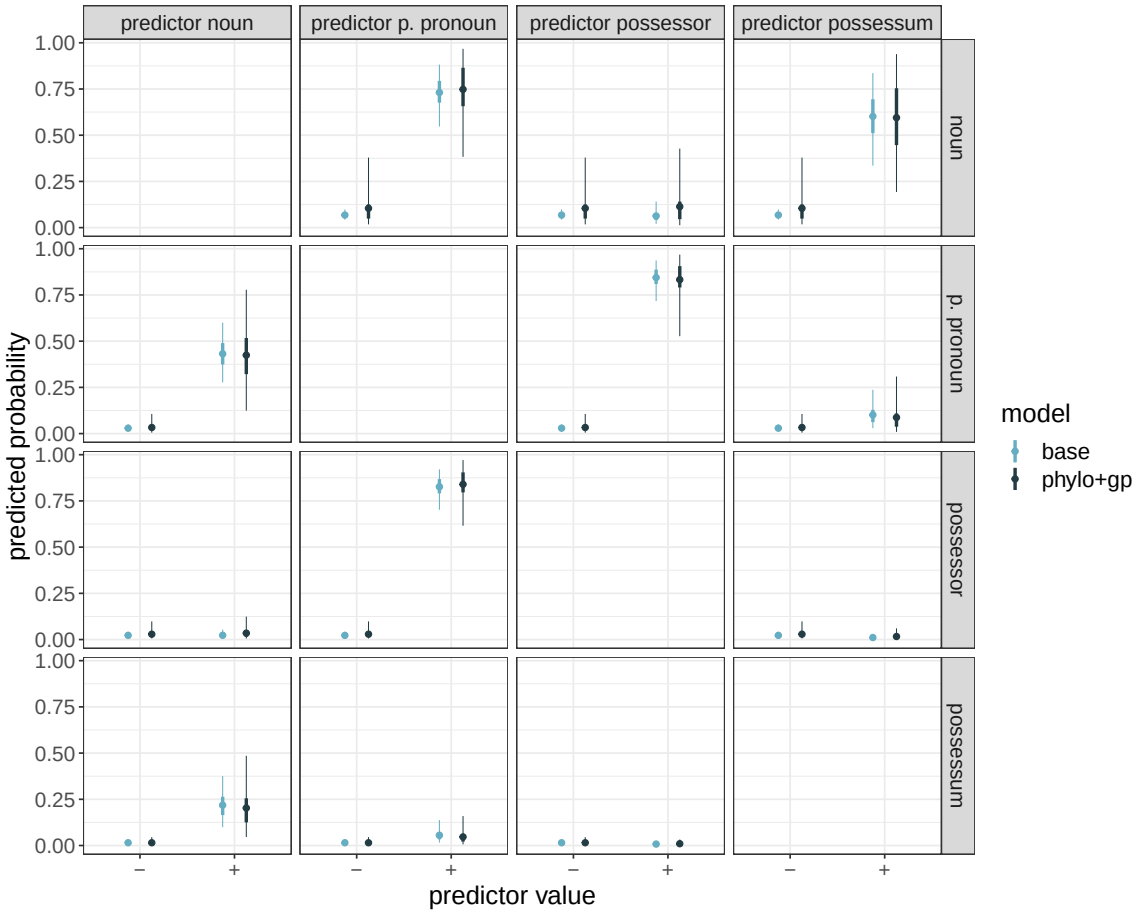

Figure 3: Predictive relations between categories in the full$_{457}$ sample

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 3 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 3 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 6 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 6: Replication of predictive relations between gender marking categories

As Table 6 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.

C.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.

D Guidelines for better replicability in typology

D.1 Transparent documentation of the language sample

In full agreement with Harris, Hyman & Staros (2006) (and others, for that matter), we think that providing the full sample including all relevant annotations and references is key for transparency and replicability. We wish to emphasize that the full sample needs to be provided; in case later subsampling is carried out in the original study as Dryer (2018) or Berg (2020), access to the original full sample is necessary to understand and evaluate all of the later sampling decisions. As for the annotations, it may also be helpful to provide an extra document with explanations for each of the variables in case the annotations are more complex.

Moreover, full transparency about the language varieties in the sample is crucial for replicability. It is not sufficient to identify a variety as, e.g., "Otomi", since this is a cover term for a number of varieties that may differ in the relevant linguistic features (Glottolog identifies 7 child varieties under "Otomi"). A simple solution is to include glottocodes as identifiers for each language of the dataset. One may argue that some of the nomenclature and phylogeny decisions in Glottolog are questionable. Yet, it is the best classification of language varieties that we currently have, being maintained and updated according to suggestions from experts in the community. Another advantage of the inclusion of glottocodes is that it makes it easy to include additional information about the language when replicating the study or re-using the dataset. In the present study, for instance, we added geographic information on the languages in the datasets using the glottocodes. Another example would be the addition of information from WALS (Dryer & Haspelmath 2013), APiCS (Michaelis et al. 2013) or Grambank (Skirgård et al. 2023), all of which is facilitated when glottocodes are provided in the original dataset.

The last point to be addressed is the format in which the dataset, i.e. the sample with annotations and references, is made available. In theory, the CLDF format (Forkel et al. 2018) is currently viewed as the standard of dataset formats in quantitative typology. This is certainly the case for larger typological databases that have been compiled in collaborative projects over the course of years and that have been published as databases to be used by the typological community. However, we doubt that this format is used much in smaller typological studies that involve a sample built by individual researchers for a particular study. In that case, it is more practical to simply provide the dataset (sample, annotations and references) in a tabulated format, e.g. as a ".csv" or ".tsv" file.[4] Such files are stored as comma-separated values (csv) or tab-separated values (tsv) and do not depend on specific operating systems or spreadsheet software such as Microsoft Excel. Tabulated formats are

[4]This does not mean that we argue against publishing data using the CLDF format; rather, we propose that a tabulated format should be included in addition to allow quick and barrier-free access to the dataset.

practical in two more ways. They are easier to access for quick visual inspection, and they can easily be imported to programming environments such as R, which is probably the environment used by most typologists for statistical analysis.[5] This last point relates to another advantage that sharing datasets in a tabular format has over sharing word or pdf files, which is still done in publications as recent as 2023. Tabular formats ensure that when re-using the dataset for another study, additional manual steps are kept to a minimum, which avoids potential errors in the otherwise manual transfer process.

D.2 Transparent linguistic analysis & annotation

Already Corbett (2005) raised an important point regarding the transparency of data annotation in typology, arguing that providing the sample together with the annotation values is not sufficient. He notes the following about how replicability is ensured in the Surrey Databases:

> [...] most important, the cells of the databases do not just contain values. They do not specify only that a language has this or that phenomenon. Rather they give a link to the data. This means that the user can reproduce the link from data to claim. If it is claimed in the database that language X has suppletion in situation Y, the user can see the data on which that claim is made. If s/he has reason to analyse the data differently, then that type can be eliminated from further analysis [...] the reasoning behind a key decision is made explicit so that the reader can make informed use of a typological resource. (Corbett 2005: 19-20)

We agree that full data transparency in typological studies should go beyond providing the final annotations together with the sample. Ideally, an independent researcher should be able to inspect the original sample and evaluate not only the annotations but also the decision leading to these annotations. In order to do that, a typological study would need to include at least one example for each of the annotation decisions made in the paper.

We are aware of the fact that this may not be realistic in every case. Still, we think that this should be the default goal for typological studies in general, with compromises made as necessary. Especially dissertations and other monographs that are built around a typological study with a language sample should aim at including examples of all the languages in the sample, e.g. as an appendix. This also forces the author to be more rigorous in the annotation process, i.e. to exclude annotation choices that cannot be backed up with a clear example from the literature. However, it is not yet the standard of large-scale typological studies to include an appendix with examples from all the languages of the sample, monographs included. Besides Corbett (2005), the few examples that we are aware of are Becker (2021), Haspelmath (1997), Hein (2020), which are three typological monographs that include examples from each language of the sample to justify the annotation decisions.[6] It remains to be seen to what extent this is also a realistic goal for studies of the size of journal articles, but the latter three examples show that it certainly is for monographs.

D.3 Transparent and accessible code for the statistical analysis

The last component necessary for transparency and replicability is the code of a statistical analysis, if carried out. While this point has not been brought up in earlier discussions, statistical tests and

[5] Tabulated formats like ".csv" or ".tsv" can easily be imported to Python as well, for that matter.

[6] Haspelmath (1997) provides complete examples for a 40 languages sub-sample of his study.

modeling are becoming increasingly common in typology. In order to be fully replicable, the code used for the statistical analysis needs to include:

- which operating system, including which version, was used,
- which software, including which version, was used,
- which other, additional software packages, including their respective versions, were used.

Information on the operating systems, software and versions thereof is crucial for replicability, as they can all influence how the code for the statistical analysis is executed. Specific functions of a software package may work well in one operating system but not in another, which could lead to code that cannot be evaluated by another researcher, with not much of a basis for trouble shooting. Functions can also do different things across operating systems and software versions, which can affect complete replication as well. With regards to this issue, Roberts (2018: 9) notes: "Roberts et al. (2015) found that results could differ substantially between running the stats on different operating systems, due to small bugs in the code for the *lme4* package (since fixed, see Roberts et al., 2015)."

There are several solutions to tackle this particular issue. R libraries like R-env or Packrat allow researchers to easily produce a project specific state (including libraries with fixed versions), which can be shared to facilitate replication.[7] Another alternative is to use a system like docker, which in addition to sharing the R libraries, allows researchers to share the exact operating system settings used to produce the analysis. Docker is a more complete system, but, in our experience, much harder to set up than R-env and Packrat.[8]

References

Becker, Laura. 2021. *Articles in the world's languages* (Linguistic Studies 577). Berlin: De Gruyter.

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

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

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

Dryer, Matthew & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(1). 180205.

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.

Harris, Alice, Larry Hyman & James Staros. 2006. What is reproducibility? *Linguistic Typology* 10(1). 69–73.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

[7] We provide lock files for our R-env state in the supplementary materials.

[8] We planned to provide a docker container but there seems to be a bug that leads to a memory leak when running Stan models, which is difficult to debug.

Hein, Johannes. 2020. *Verb doubling and dummy verb: Gap avoidance strategies in verbal fronting*. De Gruyter.

Michaelis, Susanne, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013. *APiCS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

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

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

Skirgård, Hedvig et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.