Sociogeographic correlates of typological variation in northwestern Bantu gender systems

2022, *Language Dynamics and Change*

Annemarie Verkerk and Francesca Di Garbo

Supplementary Information 2: Details on analyses

1 Introduction

This appendix contains the results of additional GLMM results and other regression models that factor in phylogenetic trees. We structure this supplementary information by topic:

- Further details on GLMMs (Section 2)
- Control for genealogical relationships (Section 3)
- Testing the influence of languages with a large number of L1 speakers (Section 4)
- Testing the influence of languages with only animacy-based gender or no gender (Section 5)
- Testing the influence of the 'sharing a border with Ubangi/Central Sudanic' predictor (Section 6)
- Including random slopes for relevant predictors (Section 7)
- Typologies of gender systems (dependent measures) (Section 8)

We include analyses on three types of measures: the binary typology (described in the main text, Section 3.1), the number of targets that agree syntactically/semantically (described in the main text, Section 3.1), and the two first and most important dimensions of a Multiple Correspondence Analysis (MCA) conducted on the full data set (described in Section 8). We also use various different ways to account for the genealogical relationships between the languages of the sample. These are listed in Table 1 and elaborated upon below in section 3, short-hands for the different models are introduced in Table 1.

2 Details on GLMMs

In this section we describe our generalized linear mixed effects models (GLMMs), focusing on aspects that were done in the same way for each dependent measure. We conducted GLMM analyses using the package brms (Bürkner 2017), available in R (R Core Team 2017). brms allows the user to fit Bayesian multilevel models in R using the probabilistic programming language Stan (Carpenter et al. 2017). In order to avoid problems with quasi-separation in the data, all independent measures (including the binary predictors, ancestor in rainforest and border with

Ubangi/Central Sudanic) were scaled using the methodology described by Gelman & Hill (2007: 56-57) and Gelman et al. (2008: 1380) such that the transformed measures have mean 0 and standard deviation 0.5. See section 4 on dealing with outliers in the number of L1 speakers. We used the recommended weakly informative prior described by Gelman et al. (2008: 1380) and Gelman (2019), a Cauchy distribution with 3 degrees of freedom, mu = 0, and scale = 2.5 (brms/R: (student_t(3, 0, 2.5)) for the main intercept and predictors. Other priors were not specified and thus received the default brms prior, which is a flat prior (see documentation on brms's set_prior() function). Chain convergence for all parts of the model (coefficients, random effects) was assessed with the help of brms report values (Rhat, Bulk ESS, Tail ESS) and was also visually assessed. Various parameters were tweaked in order to have chain convergence for all models, following warning messages from brms. The families and links used are:

- Binary typology: logistic regression, Bernoulli family, logit link
- Target counts: binomial family, logit link (15 trials)
- MCA dimensions: continuous measure, Gaussian family, identity link

As random intercepts for genealogical control, we used 1) the 'Glottolog' grouping variable based on the classifications of the Bantu language family by Glottolog (Hammarström et al. 2018) and 2) the MCC tree and full tree sets by Koile et al. (submitted) (see below). We also constructed models with random slopes from the 'Glottolog' grouping variable on those predictors that we found to be relevant in the models with only a random intercept; these are compared in section 7 of this Appendix. Since we are primarily interested in identifying the factors that impact NWB gender systems, we do not report the estimates for random effects (intercepts and slopes, when included) below.

We found that both random intercepts and random slopes were significant contributors in all GLMMs in which they were included, suggesting that these are needed to obtain the best model fits for all types of dependent measures. This was true both for analyses using the 'Glottolog' grouping variable as well as for those using Koile et al. (submitted)'s MCC tree and full tree set.

For analyses conducted using brms, we construct figures displaying the mean of the posterior distribution of each coefficient, and its two-sided 95% Credible Intervals. We use the 95% Credible Interval (95%CI) in order to interpret the relevance of each independent variable. When the 95% CI excludes zero this suggests that the evidence for contribution of the independent variable is strong enough and sufficient by standard decision rules (based on the data and applied model). In the Figures below (and in the main text), we highlight the independent variables whose 95% CI excludes zero by a preceding asterisk (*). The data and code to run the various models is added to the paper as Supplementary Information 3.

Table 1: Overview of GLMMs and other analyses summarized in this Appendix. When no reference is made to the function/package, the brms package (Bürkner 2017) is used. Random slopes were added one a one-by-one basis for predictors found to be relevant in the random intercept models, see the shorthand labels.

Genealogical control	**binary typology**	**syntactic agreement target counts**	**animacy-based agreement target counts**
none	bin_none	syn_counts_none	ani_counts_none
INT Glottolog grouping	bin_Glot_int (see main text)	syn_counts_Glot_int (see main text)	ani_counts_Glot_int (see main text)
INT Glottolog + slope	bin_Glot_int_slope_border_U_S	syn_counts_Glot_int_slope_border_U_S	ani_counts_Glot_int_slope_burder_U_S
INT Glottolog + slope	bin_Glot_int_slope_RF_overlap		
INT Glottolog + slope		syn_counts_Glot_int_slope_L1	
INT Glottolog + slope			ani_counts_Glot_int_slope_longitude
INT Glottolog + slope			ani_counts_Glot_int_slope_latitude
INT K. et al. MCC tree	bin_Koile_et_al_MCC	syn_counts_Koile_et_al_MCC	ani_counts_Koile_et_al_MCC
K. et al. tree set	bin_phylolm_Koile_et_al_treeset	syn_counts_pgls_Koile_et_al_treeset	ani_counts_pgls_Koile_et_al_treeset
K. et al. MCC tree	bin_phylolm_Koile_et_al_MCC	syn_counts_pgls_Koile_et_al_MCC	ani_counts_pgls_Koile_et_al_MCC

3 Control for genealogical relationships

We include in the analyses different ways to account for the genealogical relationships between the languages of the sample. These are:

- Grouping factor: main subgroups of Bantu distinguished by Glottolog (Hammarström et al. 2018);
- Tree set from Koile et al. (submitted);
- Maximum Clade Credibility (MCC) tree of tree set from Koile et al. (submitted).

In the main text, we report on analyses using the Glottolog grouping factor. It consists of the seven groups of Narrow Bantu as listed on Glottolog: Ababuan (22 languages in our sample), Bantu A-B10-B20-B30 (70), Bube (1), Central-Western Bantu (130), East Bantu (12), Lebonya (6), and Mbam (14). Bube is a genealogical grouping containing a single language. As this is not a workable grouping in a GLMM, we add Bube to Mbam as Mbam has been proposed as its closest affiliation. Our sample does not include all Central-Western and all East Bantu languages, as many of these are spoken in different Guthrie zones. We included only those from Guthrie zones A, B, C, D, and H. The Glottolog Bantu family tree is based on Bostoen & Gregoire (2007) which in turn is based on Bastin et al. (1999).

The phylogenetic trees we used are taken from Koile et al. (submitted). This is follow-up work on the route of the Bantu expansion and uses data from Grollemund et al. (2015), the latest phylogenetic analysis of the Bantu languages. The advantage of using the phylogenetic trees generated by Koile et al. (submitted) rather than those generated by Grollemund et al. (2015) is that the former includes more languages than the latter, with added languages being assigned to genealogical groupings on the basis of Glottolog (Hammarström et al. 2018, i.e. no new lexical data is used for phylogenetic estimation). Using these more recent analyses enables us to include almost all of the languages in our sample in the GLMMs, thus increasing their statistical power. The tree set includes 400 trees, the Maximum Clade Credibility (MCC) tree is a summary of the full sample of 400 trees.

In previous versions of this work, we considered a second grouping factor, the main subgroups of (NW) Bantu distinguished by Grollemund et al. (2015): North-Western Cameroon (48 languages in our sample), North-Western Gabon (38), Central-Western (89), and West-Western (55). There are also some H languages that belong to their South-Western clade (8), and some D languages that belong to their Eastern clade (17), so in total there are 6 groups. Bostoen (2019: end of sect. 4) suggests that these groups have gained wider acceptance, and names them as follows: "Mbam-Bubi (A40-A60) straddling Narrow and Wide Bantu; North-West (rest of zone A + B10-30); Congo-Basin (mainly zone C); West-Coastal (mainly B40-80, zone H); South-West (zones K and R + parts of zone L); East (remainder apart from some hard-to-classify zone D languages in North-East-Congo)." We did not use this grouping factor as we compared model fit of the models reported on in text (bin_Glot_int, syn_counts_Glot_int, and ani_counts_Glot_int) with parallel models using the Grollemund et al. (2015) groupings using the R package 'loo' (Vehtari et al. 2017), and found that the Glottolog grouping factor outperformed the Grollemund et al. (2015) grouping factor in all three models.

The GLMMs using the MCC tree by Koile et al. (submitted) are compared to the models reported on in text in Figure 1. The first thing that can be observed is the fit of the binary typology model (bin_Koile_et_al_MCC) vs. the two models using target counts; the posterior distributions of the binary typology model are very wide, resulting in no significant effects at all. For the model with the number of targets agreeing syntactically as the dependent variable, number of L1 speakers is a significant negative predictor. For the model with the number of targets taking animacy-based gender marking as the dependent variable, sharing a border with Ubangi/Central Sudanic is a significant positive predictor.

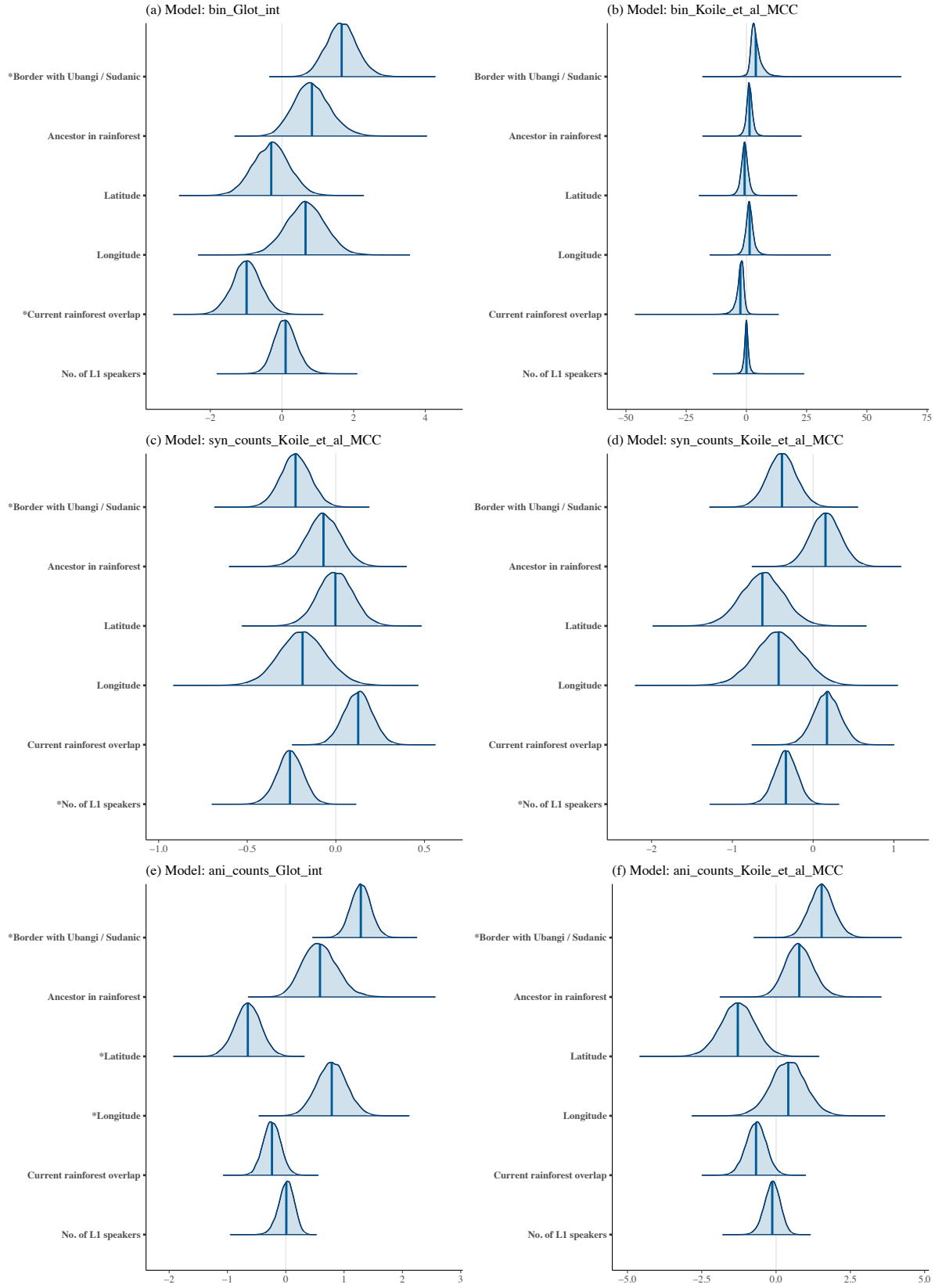

Figure 1: Comparison of posterior distribution of fixed effects of models reported on in the main text, using a random intercept using Glottolog groupings (left), and correlate models with a random intercept using the co-variance matrix from the MCC tree by Koile et al. (submitted) (right)

In addition to GLMMs in brms, we used phylogenetic comparative methods specialized in controlling for genealogy while running regression analyses. For the dependent measures where the number of targets with a particular type of gender marking is modelled, we used phylogenetic generalized least squares (PGLS); for the binary typology, we used a phylogenetic linear regression model.

We conducted phylogenetic generalized linear models to model the binary typology using phylolm (Ho & Ane 2014) in R (R Core Team 2017), reported on in Tables 2 and 3. The first of these analyses uses the MCC tree, the second the full treeset of 400 phylogenetic trees. phylolm uses not a Bayesian but rather a maximum likelihood algorithm. For those models, statistical relevance (or significance) of the independent measures is assessed using the reported p-value. We find four relevant predictors in these models, and both models give more or less the same results: current rainforest overlap is a significant negative predictor; longitude, ancestor in rainforest and sharing a border with Ubangi/Central Sudanic are significant predictors.

Table 2: Relevance of coefficients for bin_phylolm_Koile_et_al_MCC

	Estimate	Std. Error	z value	p value
Intercept	-3.58	1.03	-3.49	0.0005
No. of L1 speakers	-0.03	0.35	-0.08	0.94
*Current rainforest overlap	-1.12	0.44	-2.53	0.02
*Longitude	1.13	0.56	2.03	0.042
Latitude	-1.05	0.56	-1.88	0.06
*Ancestor in rainforest	2.61	0.98	2.66	0.008
*Border with Ubangi / Sudanic	1.73	0.14	3.06	0.003

Table 3: Relevance of coefficients for bin_phylolm_Koile_et_al_treeset

	mean Estimate	sd Estimate	mean z value	mean p value
Intercept	-3.63	0.31	-3.95	0.0007
No. of L1 speakers	0.03	0.05	0.08	0.89
*Current rainforest overlap	-1.04	0.09	-2.38	0.02
*Longitude	1.35	0.33	2.66	0.048
Latitude	-0.85	0.19	-1.66	0.11
*Ancestor in rainforest	2.58	0.29	2.78	0.008
*Border with Ubangi / Sudanic	1.65	0.57	2.88	0.004

The number of targets with a particular type of gender marking was modelled using phylogenetic generalized least squares (PGLS) implemented in the caper (Orme 2011) R package (R Core Team 2017), reported on in Tables 4 and 5 (syntactic agreement) through 6 and 7 (animacy-based agreement). PGLS is used to model continuous dependent measures, and using it for count data is thus technically not appropriate. However, we use it here in order to contrast the results with those using brms and to assess the relevance of the phylogenetic component (see below). We scaled the count data using the methodology described by Gelman & Hill (2007: 56-57) and Gelman et al. (2008: 1380) such that the transformed measures have mean 0 and standard deviation 0.5. PGLS as implemented in caper does not use a Bayesian but rather a maximum likelihood algorithm. For those models, statistical relevance (or significance) of the independent measures is assessed using the reported p-value.

The first of these analyses uses the MCC tree, the second the full tree set of 400 phylogenetic trees. These analyses are similar to the analyses reported on in the main text (syn_counts_Glot_int and ani_counts_Glot_int, see Figs. 7 and 8 in the main texts). The models using the number of targets that agree syntactically (syn_counts_pgls_Koile_et_al_MCC and syn_counts_pgls_Koile_et_al_treeset) have two relevant predictors, number of L1 speakers and sharing a border with Ubangi/Sudanic

(the former effect is not significant for the analyses using the full tree sample, p = 0.09). The direction of the effects is identical too. The models using the number of targets that display animacy-based agreement (ani_counts_pgls_Koile_et_al_MCC and ani_counts_pgls_Koile_et_al_treeset) have a single relevant predictor, sharing a border with Ubangi/Sudanic. In the corresponding brms model (model ani_counts_Glot_int, see Figure 8 in the main text) longitude and latitude are also relevant predictors, especially in Table 7 we do see longitude approaching significance.

Table 4: Relevance of coefficients for syn_counts_pgls_Koile_et_al_MCC

	Estimate	Std. Error	t value	p value
Intercept	0.02	0.19	0.08	0.93
*No. of L1 speakers	-0.15	0.07	-2.02	0.045
Current rainforest overlap	0.08	0.09	0.93	0.35
Longitude	-0.12	0.14	-0.84	0.40
Latitude	-0.23	0.13	-1.73	0.09
Ancestor in rainforest	0.08	0.16	0.47	0.64
*Border with Ubangi / Sudanic	-0.25	0.12	-2.06	0.04

Table 5: Relevance of coefficients for syn_counts_pgls_Koile_et_al_treeset

	mean Estimate	sd Estimate	mean t value	mean p value
Intercept	0.08	0.016	0.65	0.52
No. of L1 speakers	-0.13	0.005	-1.69	0.09
Current rainforest overlap	0.09	0.006	1.09	0.28
Longitude	-0.03	0.016	-0.33	0.74
Latitude	-0.13	0.031	-1.39	0.17
Ancestor in rainforest	-0.03	0.03	-0.20	0.80
*Border with Ubangi / Sudanic	-0.30	0.02	-2.59	0.01

Table 6: Relevance of coefficients for ani_counts_pgls_Koile_et_al_MCC

	Estimate	Std. Error	t value	p value
Intercept	-0.19	0.21	-0.94	0.34
No. of L1 speakers	-0.02	0.07	-0.37	0.71
Current rainforest overlap	-0.13	0.08	-1.57	0.12
Longitude	0.17	0.14	1.16	0.25
Latitude	-0.11	0.13	-0.88	0.38
Ancestor in rainforest	0.12	0.16	0.75	0.45
*Border with Ubangi / Sudanic	0.34	0.12	2.92	0.004

Table 7: Relevance of coefficients for ani_counts_pgls_Koile_et_al_treeset

	mean Estimate	sd Estimate	mean t value	mean p value
Intercept	-0.30	0.032	-2.30	0.05
No. of L1 speakers	-0.005	0.010	-0.07	0.95
Current rainforest overlap	-0.09	0.009	-1.19	0.24
Longitude	0.18	0.030	1.91	0.09
Latitude	-0.10	0.016	-1.05	0.30
Ancestor in rainforest	0.23	0.046	1.74	0.11
*Border with Ubangi / Sudanic	0.42	0.031	3.81	0.0007

PGLS implemented in caper (Orme 2011) includes a test of phylogenetic signal called lambda (λ) Pagel (1999). λ is a branch length scaling parameter and can be used to measure the extent of the dependency of the data on the model of Brownian evolution given the structure of the phylogenetic tree. The optimized λ value indicates to what extent the data under consideration have been influenced by genealogy: scores that are zero or close to zero indicate that data does not pattern according to genealogy; scores of one or close to one (> 0.8) imply an affect of shared history on the data. High λ scores imply that closely related languages behave similarly. The distributions of λ scores (n = 400 phylogenetic trees) of the PGLS analyses for syn_counts_pgls_Koile_et_al_treeset and ani_counts_pgls_Koile_et_al_treeset are included in Figure 2. This Figure shows that neither analysis displays high λ scores, implying that there is no strong relationship between the data on the number of targets that agree syntactically or those that agree semantically, and the genealogical relationships between the NWB languages.

4 Models testing the influence of languages with a large number of L1 speakers

In the models reported so far we have not dealt with outlier languages in terms of number of L1 speakers (see Figure 10 in the main text). We have chosen not to log population figures prior to scaling, as we felt that 1) the variable would then become too far removed from the facts and 2) logging produced outliers on the other end of the scale, that is, for very small languages. In Figure 3, we compare the three main models reported on in the main text (left) with models that exclude the 15 languages with 400.000 speakers or more (right). This cut-off point was chosen after assessing a QQ-plot of the scaled number of L1 speakers.

The Figure shows that overall, the exclusion of these fifteen languages impacts the results quite minimally, only the model of the binary typology is affected such that the effect for current rainforest disappears, and an effect for ancestor in rainforest becomes relevant. However, the effect of number of L1 speakers turns around. While we find that a smaller number of targets agreeing syntactically is associated with a higher number of speakers in model syn_counts_Glot_int, a bigger number of targets agreeing syntactically is associated with a higher number of speakers

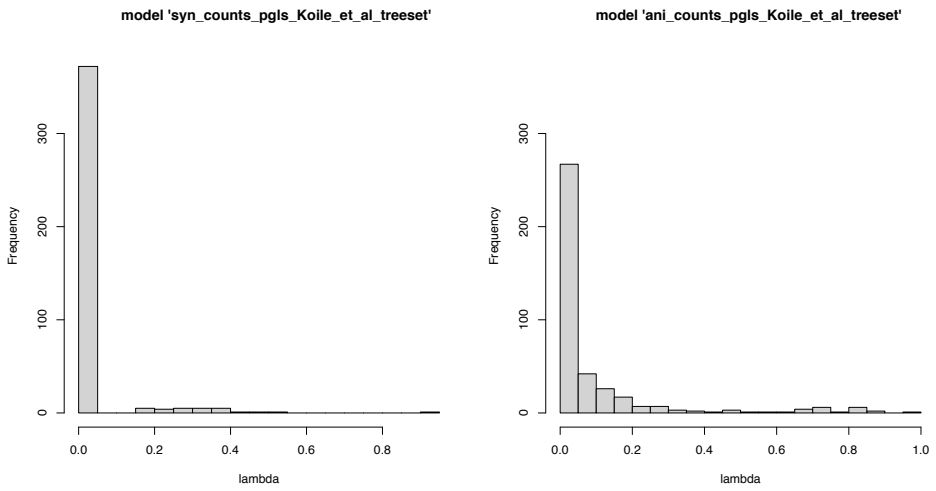

Figure 2: Distribution of λ scores for PGLS model syn_counts_pgls_Koile_et_al_treeset, for analyses modelling the number of targets that agree syntactically (left), and model ani_counts_pgls_Koile_et_al_treeset, for the number of targets that agree semantically (right)

in model syn_counts_Glot_int_no_L1_outl. Figure 10 in the main text shows these two opposite interactions. On the left, we can only see some data points (around 15) clearly; these all have a number of L1 speakers $>$ 400.000.[1] On the left side of Figure 10 in the main text, we can see a negative effect, with a smaller number of targets agreeing syntactically being associated with a higher number of speakers. On the right, we are looking only at data points from languages with 400.000 speakers and lower, and there we can observe a positive effect, i.e., a bigger number of targets taking syntactic agreement is associated with a higher number of speakers.

How to explain this seemingly discrepant result? We believe that, at least in this NWB data set, the effect of the number of speakers is only meaningful above a certain cutoff point, that is, for the bigger languages, and not for smaller languages (the cutoff point need not necessarily lay at exactly 400.000 speakers). In our data set, the distribution of the number of L1 speakers variable is skewed in both directions. We have 68 languages with 10.000 speakers or less, and 139 (including these 68) with 100.000 speakers or less. In contrast, we have 15 languages with 400.000 speakers or more.

The number of L1 speakers has been used as a proxy for language contact and within-group communication with adult learners. A large proportion of L1 speakers may lead to 'simplification' of language structures (e.g., reduction of inflectional morphology Lupyan & Dale 2010; Bentz & Winter 2013; Sinnemäki & Di Garbo 2018). Few languages of our sample qualify as languages of wider communication (e.g. Kituba and Kinshasa Lingala) and it is thus not surprising that the expected 'simplification effect' in the domain of gender marking only appears in the models where these languages are included. Conversely, our sample includes many languages that are (extremely) localized, i.e. used as languages that outsiders do not learn. It seems to us that the number of L1 speakers is thus less meaningful in this context, as those dynamics of language restructuring that the number of L1 speakers variable attempts to capture in big languages occurs relatively infrequently in our data set. In our sample, the effect that language contact may have on the typology

[1] In our sample, five languages out of 179 have a population size higher than one million speakers. These are, in ascending order of population size as documented by Ethnologue (Lewis et al. 2016): Fang (1.071.900 L1 speakers), the Congo variety of Kituba (1.160.000 L1 speakers), Kimbundu (1.700.000 L1 speakers), Kinshasa Lingala (2.040.000 L1 speakers), the Democratic Republic of Congo variety of Kituba (4.200.000 L1 speakers), South-Central Kikongo (5.016.500 L1 speakers). Out of these five languages, only Fang has a gender system with only syntactic agreement.

of gender is more directly and adequately captured by the other sociogeographic measures we employ.

We do not have an immediate explanation for the effect we find in model syn_counts_Glot_int_no_L1_outl (Figure 3d), where a bigger number of targets taking syntactic agreement is associated with a higher number of speakers. We can observe several languages with a reasonably large number of speakers and many targets that agree syntactically; the two languages with syntactic agreement on 14 targets, Akoose and Mbala at the far right of Figure 10 in the main text, are spoken by 100.000 and 200.000 speakers, respectively. Languages with a large number of speakers and many targets that agree syntactically may 'resist' restructuring on the basis of being surrounded by similar languages.

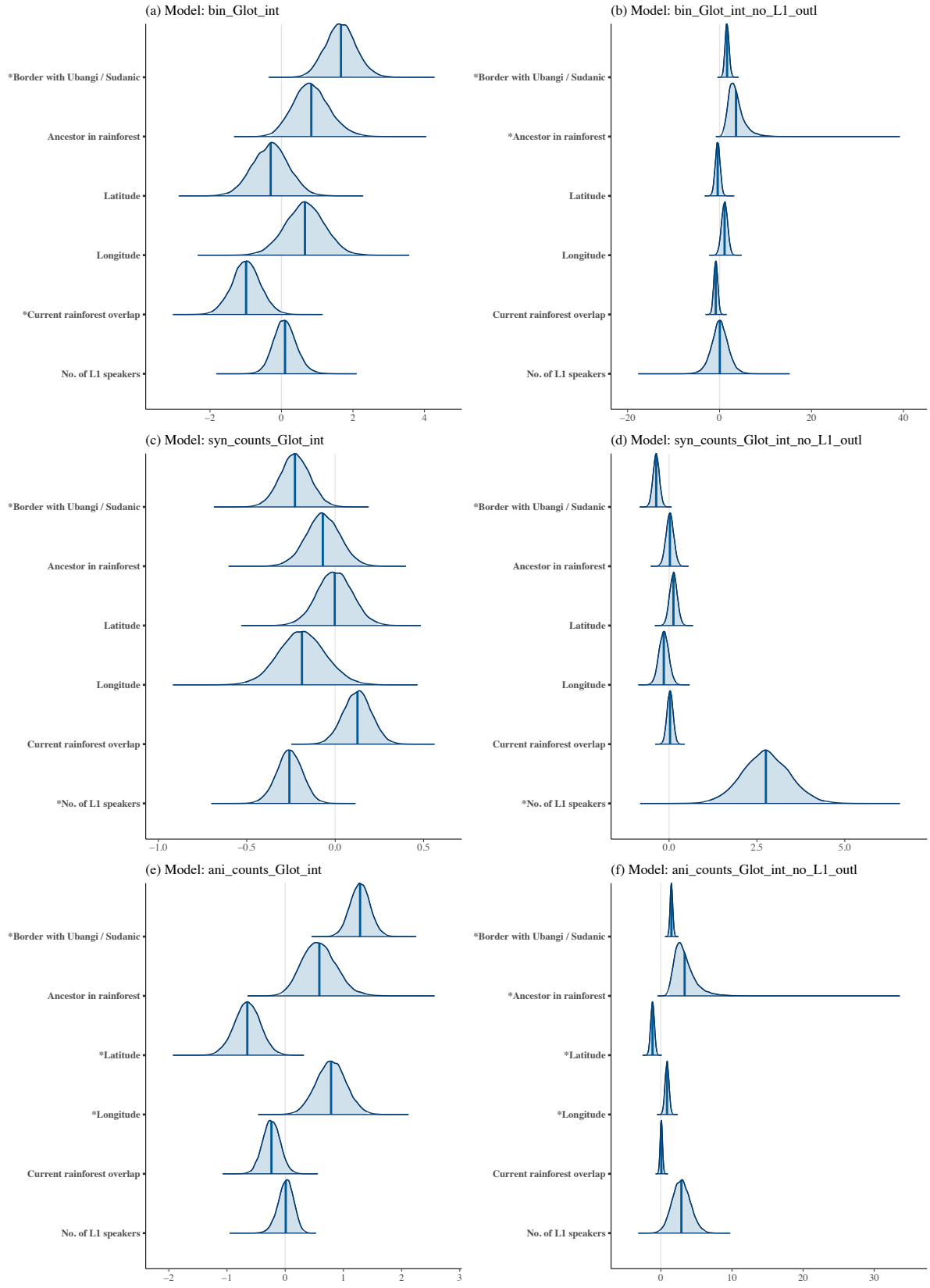

Figure 3: Comparison of posterior distribution of fixed effects of models reported on in the main text (left) and correlate models without the 15 languages with 400.000 speakers or more (right)

5 Models testing the influence of languages with only animacy-based gender or no gender

Figure 4 is a comparison of the three main models reported on in the main text (left) and correlate models without the 17 languages with only animacy-based agreement or no gender (right). It shows that, with these 17 languages excluded from the data set, 1) for the binary typology, the effect for current rainforest overlap disappears; 2) for the number of targets that agree syntactically, the effects from sharing a border with Ubangi / Central Sudanic and number of L1 speakers disappear, instead there are effects from latitude and current rainforest overlap; 3) for the number of targets marked for animacy-based gender, the effect for longitude disappears.

This suggests that the 17 languages with only animacy-based agreement or no gender are important for the sociolinguistic typology of gender systems in NW Bantu, especially regarding the number of L1 speakers variable, which does not appear in any of the models excluding these languages.

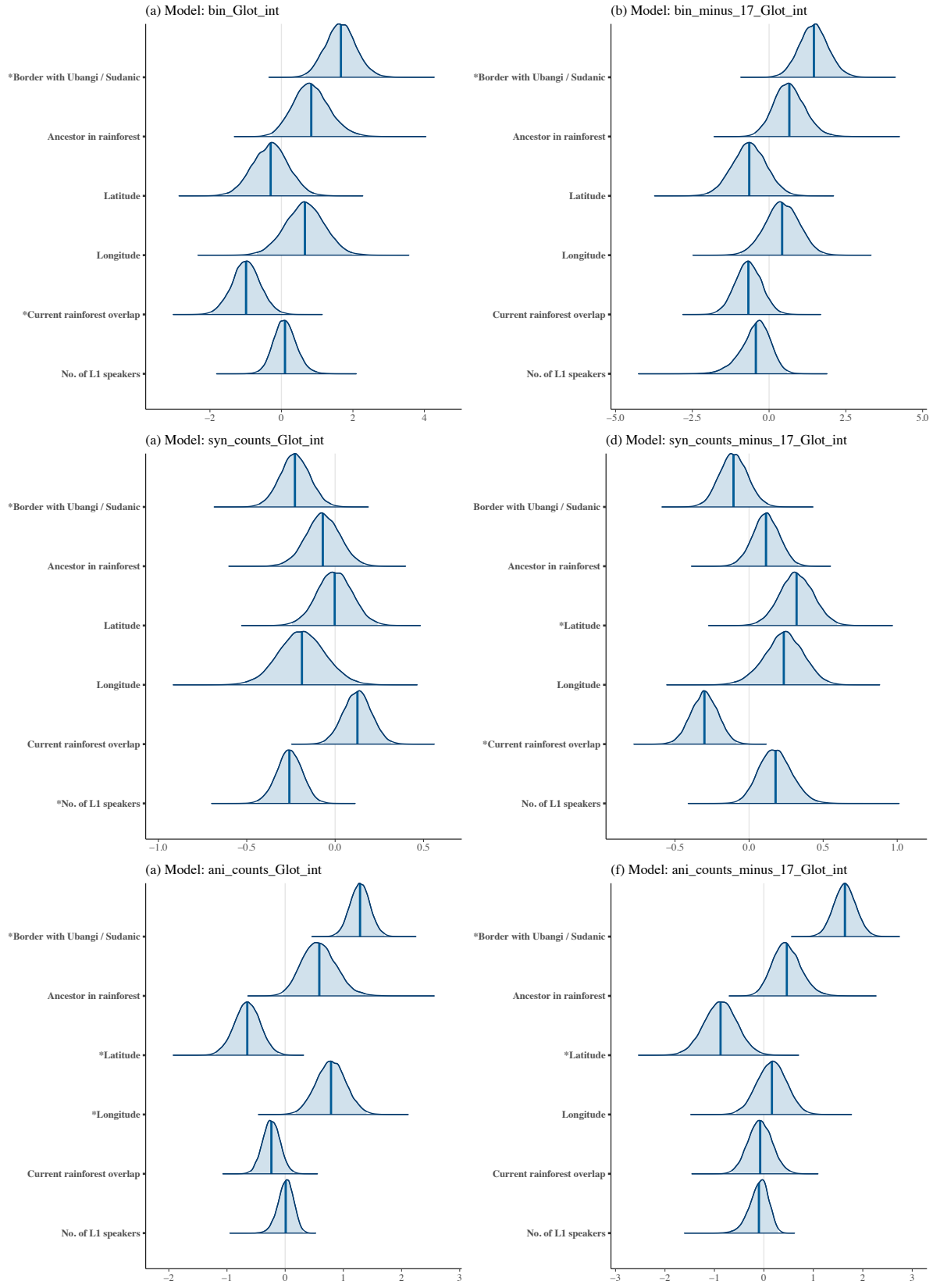

Figure 4: Comparison of posterior distribution of fixed effects of models reported on in the main text (left) and correlate models without the 17 languages with only animacy-based agreement or no gender (right)

6 Models testing the influence of the predictor sharing a border with Ubangi/Central Sudanic

Figure 5 is a comparison of the three main models reported on in the main text (left) and correlate models without including the predictor sharing a border with Ubangi/Central Sudanic. We suspected that if this relevant and sometimes quite strong predictor is excluded, other geographical predictors (longitude, latitude, current rainforest overlap, and ancestor in rainforest) might become relevant predictors. After all, these all speak towards contact-induced change in the northern Bantu borderlands in one way or another.

However, our suspicion was not justified. 1) For the binary typology, the other significant effect in model bin_Glot_int, current rainforest overlap, remains significant; 2) for the number of targets that agree syntactically, the effect of number of L1 speakers remains significant too; 3) for the number of targets marked for animacy-based gender, the effect for latitude disappears. No new relevant effects turn up, suggesting that these measures speak to different factors (see also Figure 9 in the main text).

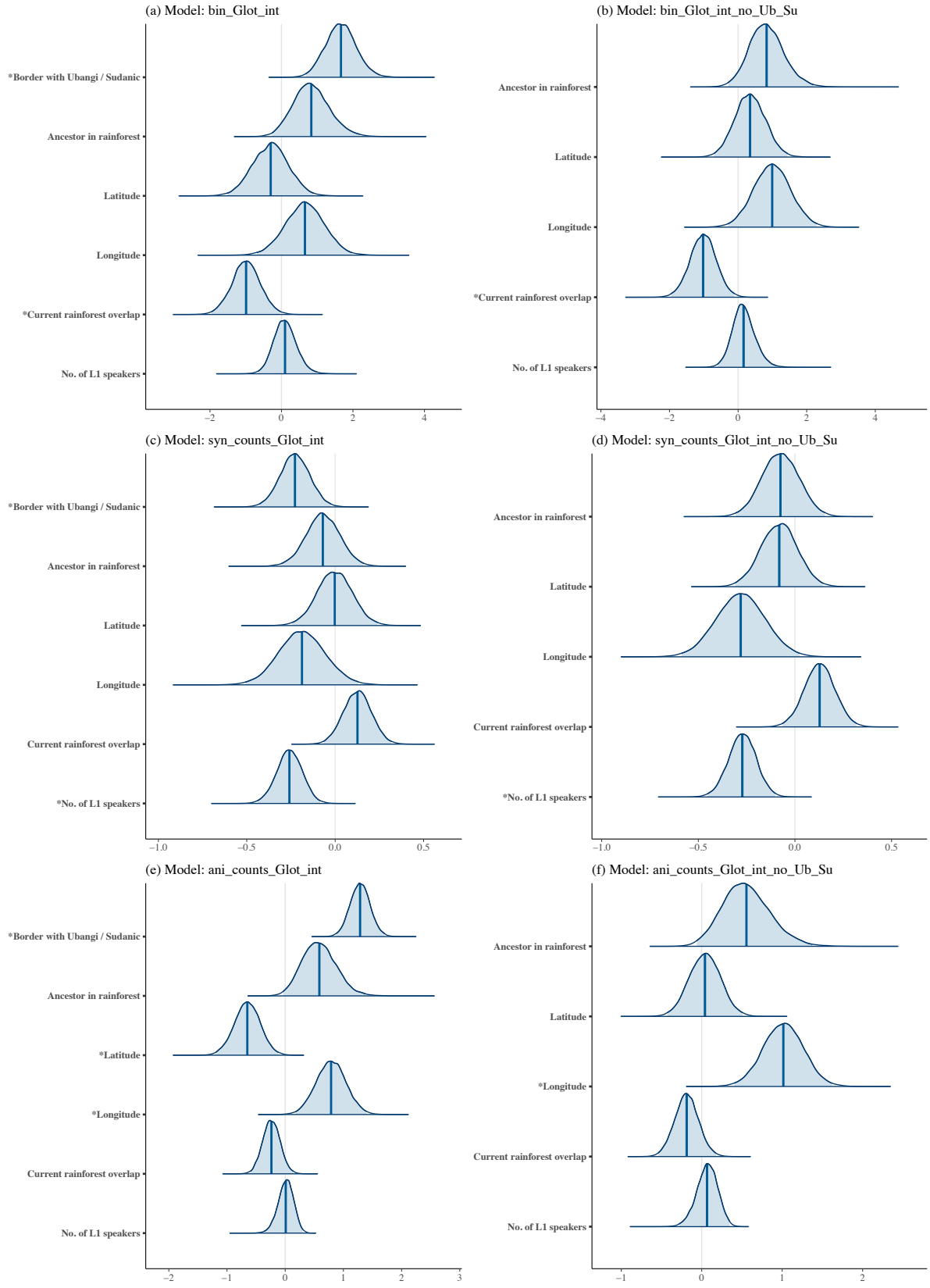

Figure 5: Comparison of posterior distribution of fixed effects of models reported on in the main text (left) and correlate models without border with Ubangi/Central Sudanic as predictor (right)

7 Models including random slopes for relevant predictors

Figures 6, 7, and 8 give the posterior distribution of fixed effects of the models reported on in the main text (bin_Glot_int, syn_counts_Glot_int, and ani_counts_Glot_int) in comparison with models to which random slopes are added for each of the significant effects reported on in the main text.

In all cases, the effect disappears when a random slope for that predictor is added, while the rest of the model stays similar to the model without the random slope (which is hard to observe because of the wide posterior distribution of variables with random slopes). The only exception is model ani_counts_Glot_int_slope_longitude, which includes a random slope on longitude; in this model, the effect for latitude also disappears.

In all of these models, the random slope on the various predictors is a significant component judged by the 95% confidence interval; all of these exclude zero. However, these models were hard to fit because plots of the MCMC posterior distributions looked capped. We intended to report on formal model comparison using the R package 'loo' (Vehtari et al. 2017), but were unable to as we could not estimate 'loo' or WAIC criterions for these models. We suspect this may be due to poor fit of these models, regardless of our efforts to improve their fit (following suggestions by brms, such as manipulating brms' settings such as 'adapt_delta' and 'max_treedepth', and running the chains for more iterations).

Why are random slopes important for GLMMs? Because they are the only factor that can model differences in the relationship between the predictor and the response variable across genealogical groups (genealogical groups in our study are the major subgroups as identified by Glottolog (Hammarström et al. 2018) or various genealogical subgroups captured by Koile et al.'s (submitted) phylogenetic trees). Imagine the following example: in the middle of an extended area, we have a cluster of languages with no gender, surrounded by languages with gender. A genealogical border runs right trough the middle of this cluster of genderless languages, dividing the area in two genealogical groups ('Western' and 'Eastern'). The slope of longitude as an explanatory variable on type of gender system will be positive in one genealogical group and negative in the other. Such differences can only be captured by random slopes.

Most statisticians (Barr et al. 2013 is the go-to citation) argue for 'full' models in which non-independence in the data is accounted for using both random intercepts and random slopes. However, there are also researchers that argue for a more nuanced approach (Matuschek et al. 2017; Coupe 2018: Winter 2020). We side with the latter group. There are many situations in which random slopes are needed, such as accounting for per-subject and per-item differences in psycholinguistics (Barr et al. 2013; Seedorff et al. 2019) and for accounting for genealogical and geographical non-independence in sociolinguistic typological studies that have a world-wide sample (Lupyan & Dale 2010; Atkinson 2011; Jaeger et al. 2011; Coupe 2018). However, as noted by Matuschek et al. (2017), random slopes that are not supported by sufficient data can reduce power, and it can be very difficult to tell whether the disappearance of observed effects when adding random slopes reflects a lack of power or a successful attempt to deal with type I errors (in our case, a type I error would be concluding that there is a correlation between NWB gender systems and demographic/geographic variables, while such a correlation is in fact due to shared descent).

We suspect that, in our case, the disappearance of observed effects (as reported in Figures 6, 7, and 8 in the main text) when adding random slopes (as reported in Figures 6, 7, and 8) is at least in part due to lack of power. We generally have problems in fitting these models despite obliging to brms' (Bürkner 2017) warnings, and could not conduct model comparison with loo (Vehtari et al. 2017).[2] How could we attempt to increase statistical power?

Statistical power depends on effect size, variability, and sample size (Winter 2020: 174) as well as their interaction. There is an immense literature on this topic, because experimental and survey-based research can to some extent manipulate all of these three factors, and researchers have

[2]However, since the random slopes were relevant components, we suspect model comparison would be inconclusive, i.e. small/inconclusive amounts of support for those models that include random slopes.

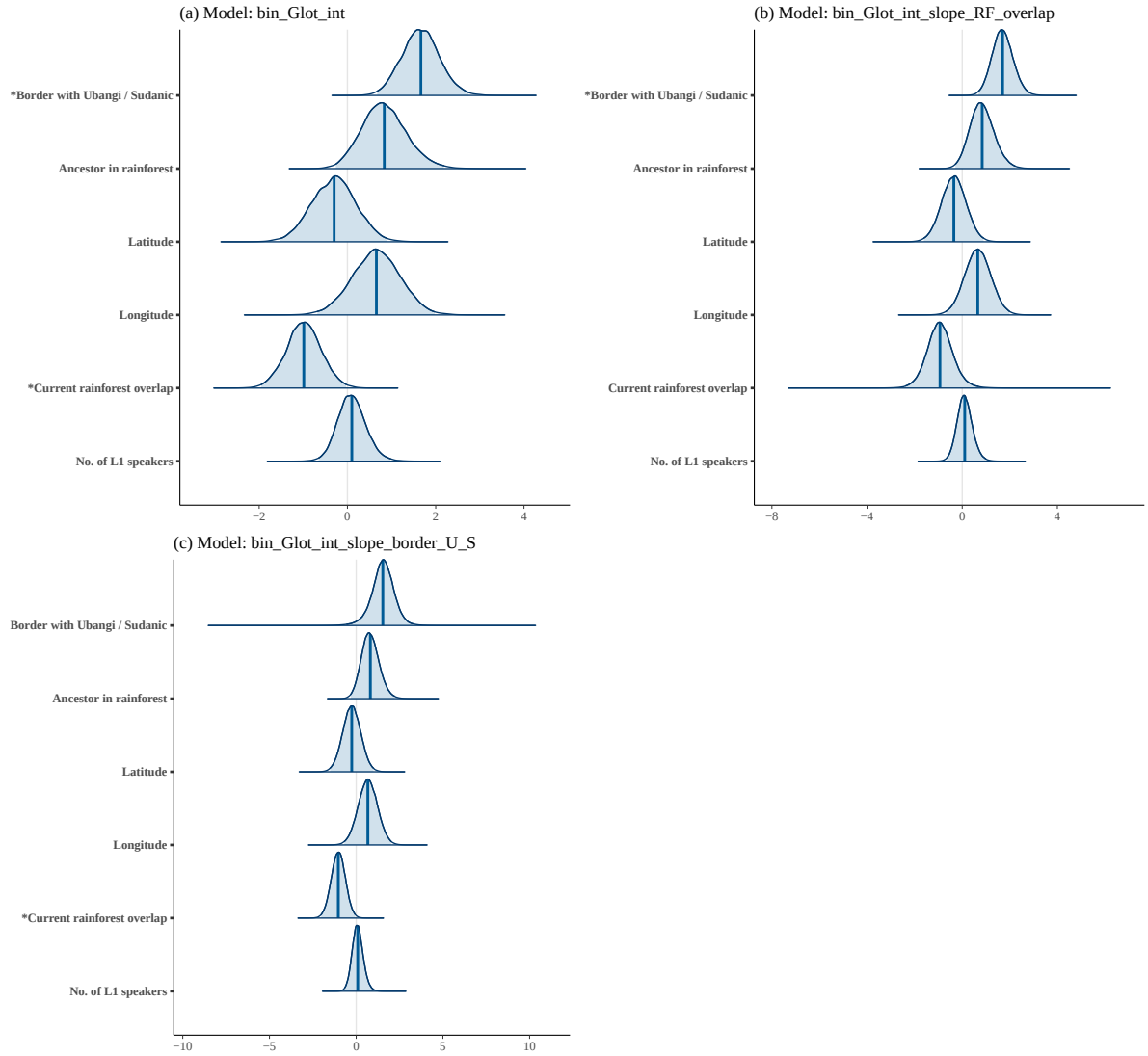

Figure 6: Comparison of posterior distribution of fixed effects of one of the models reported on in the main text (bin_Glot_int) and correlate models with random slopes added for relevant effects in bin_Glot_int, with a random slope for current rainforest overlap (bin_Glot_int_slope_RF_overlap) and a random slope for sharing border with Ubangi / Central Sudanic (bin_Glot_int_slope_border_U_S)

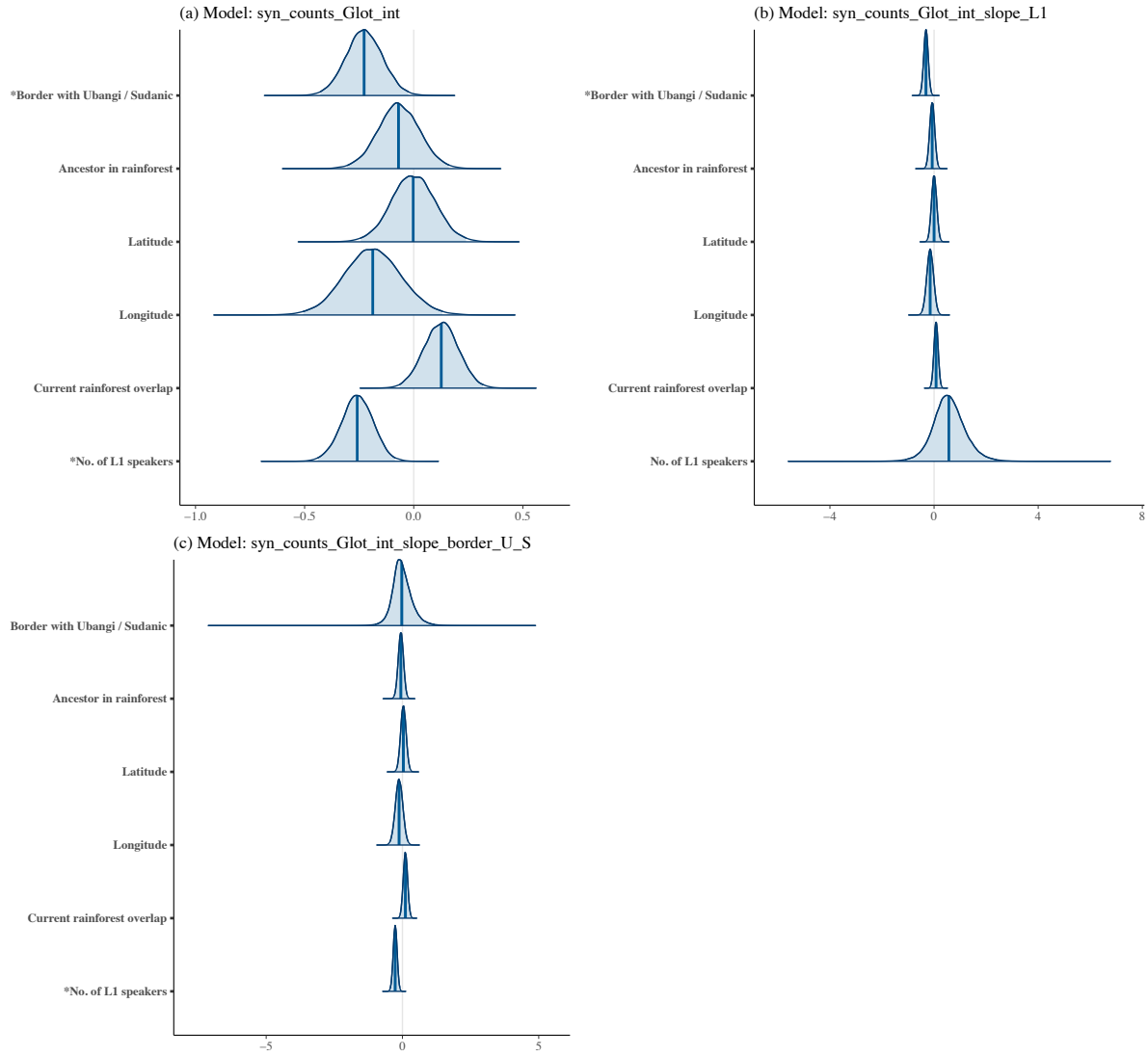

Figure 7: Comparison of posterior distribution of fixed effects of one of the models reported on in the main text (syn_counts_Glot_int) and correlate models with random slopes added for relevant effects in syn_counts_Glot_int, with a random slope for number of L1 speakers (syn_counts_Glot_int_slope_L1) and a random slope for sharing border with Ubangi / Central Sudanic (syn_counts_Glot_int_slope_border_U_S)

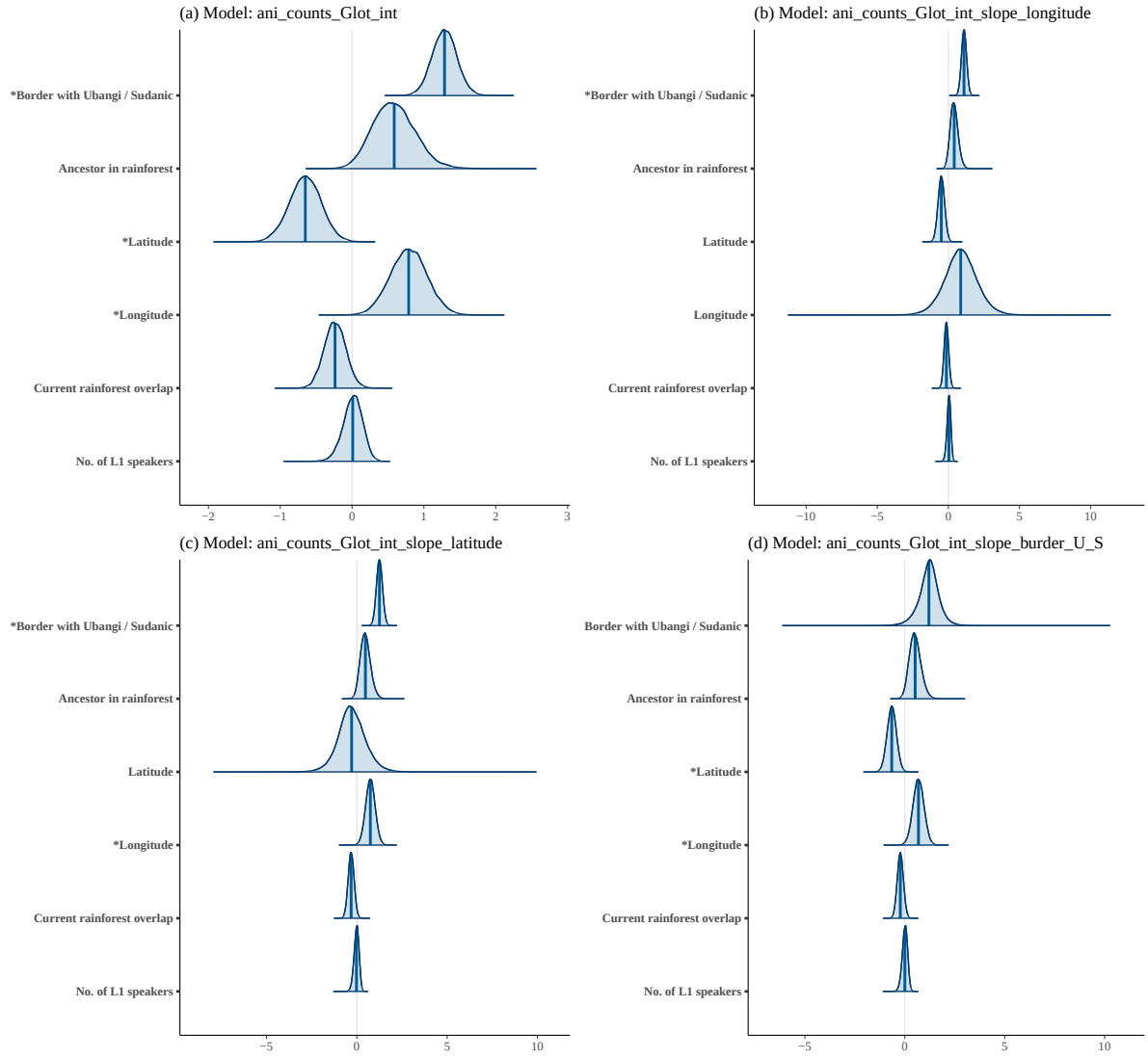

Figure 8: Comparison of posterior distribution of fixed effects of one of the models reported on in the main text (ani_counts_Glot_int) and correlate models with random slopes added for relevant effects in bin_Glot_int, with a random slope for longitude (ani_counts_Glot_int_slope_longitude), a random slop for latitude (ani_counts_Glot_int_slope_latitude) and a random slope for sharing border with Ubangi / Central Sudanic (ani_counts_Glot_int_slope_burder_U_S)

asked which manipulations lead to the required power - a luxury typologists do not have, as will become clear in the following. First of all, effect size. Effect size is a way of quantifying the size of the difference between two or more groups. We can manipulate it in our measurements, for instance by making a psycholinguistic task harder, potentially increasing the difference between test and control group. Large variability within groups, contrary to what one may first think, affects statistical power negatively because it implies a larger overlap in the measurements between groups; hence we would require a larger sample size in order to increase resolution. Sample size is the most straightforward one, the more data one has, the better (Winter 2020: 267). Simulation studies (Scherbaum & Ferreter 2009, Maas & Hox 2005) suggest that in multilevel models such as ours, the number of groups (in the current study: genealogical subgroups) is more important than the level of individuals (languages).

Typological measure design with an eye on increasing effect size should be possible. In fact, we can observe this in some sense also in the current study, as the counts of the number of targets that take syntactic/animacy-based agreement seem to be more sensitive measures than the binary typology. We may of course devise other measures in this light; we discuss this topic in Section 5 in the main text. Variability and sample size, however, are very difficult to manipulate in (sociolinguistic) typological studies (for similar considerations, see also Piantadosi & Gibson 2014). So far, sparse sampling has turned out to be a problem for typological studies with a world-wide sampling (see Atkinson 2011, Jaeger et al. 2011). Here we aimed for an exhaustive sample and still face power problems. We included 176 languages distributed over 6 genealogical groups (Glottolog, Hammarström et al. 2018), and some of these groups simply are very small: Abaduan: 22 languages; East Bantu: 12;[3] Lebonya: 6; Mbam-Bube: 15. Ideally, we would have more than six groups - but of course, the Bantu genealogical tree is a (relatively) fixed concept, we cannot conjure up more Bantu subgroups.[4] It is also far from clear that in a typological context, it is indeed the number of groups and not the number of individuals that needs to be increased. Jaeger et al. (2011: Ap. C) show how hard it is to show an effect between phonemic inventory and distance from West Africa (data from Atkinson 2011) including random slopes even with a minimum of 10 languages per group.

Additionally, we would argue that not all predictors need to be fitted using random slopes all the time, following Winter's (2020: 242) recommendation to reason about the most appropriate model, given the data and theory. We would argue that the number of L1 speakers in our study, i.e. within the NWB context, does not need a random slope for the following reason. We use random intercepts and slopes to account for genealogical relatedness. For a world-wide sample, we can imagine that the number of L1 speakers has different effects on typological variables across different families (see Bentz et al. 2015: Fig. 7 for an example of across-family differences in the relationship between lexical diversity and number of L2 speakers). But since we find very little difference in the number of L1 speakers across NW Bantu Glottolog groupings (see Figure 5 in the main text)[5], we cannot imagine that there would be an effect of genealogical relatedness for this variable, i.e. that the number of L1 speakers would have a different interaction with the type of gender systems across genealogical groupings. We leave the discussion regarding random slopes for sociolinguistic typological studies here and hope for fruitful discussion regarding these matters in future work.

[3] 12 East Bantu languages have been included in the current sample; the East Bantu subgroup includes 253 languages according to Glottolog Hammarström et al. 2018.

[4] We could of course extend our sample to the entire Bantu family, but this would not help necessarily - in Glottolog (Hammarström et al. 2018), Narrow Bantu has 559 languages in the same six subgroups featured in the current study (counting Mbam-Bube as a single subgroup as we did). Another solution might be to use genealogical groupings lower down in the phylogenetic tree; Central-Western Bantu for instance has 11 further subgroups in Glottolog. This issue needs further investigation.

[5] The small differences that we do observe may be argued to be due to the small size of some of the Glottolog groupings, such as Lebonya and Mbam.

8 MCA

We conducted analyses on three types of measures: the binary typology (described in the main text), the number of targets that receive syntactic agreement /animacy-based agreement (described in the main text), and the two first and most important dimensions of a Multiple Correspondence Analysis conducted on the full data set. This third dependent measure, MCA dimensions, was not introduced in the main text, but is based on Di Garbo & Verkerk (Accepted, 2021). MCA stands for multiple correspondence analysis, a method of data analysis which is very similar to the better known principal component analysis (PCA). Both methods are used to detect and represent structures in a data set by transforming potentially correlated variables into a smaller set of variables, called components or dimensions, which are no longer correlated and which best describe the variation attested in the data set. While PCA is used for continuous variables and is thus not applicable to our data set, MCA deals with categorical variables, like the ones we use in our questionnaire. Another type of factorical analysis that we could use is Multiple Factor Analysis (MFA), which would allows to include both continuous measures/counts (such as the questions on the number of noun class forms and agreement classes listed in Appendix A in the main text) and the categorical variables we can work with for MCA.

We conducted MCA analyses on the answers to all the questions included in our questionnaire (see Appendix A in the main text), that is both the binary questions on syntactic and animacy-based agreement, and the set of additional questions which concludes the coding sheet), using the package FactoMineR in R (Lê et al. 2008, R Core Team 2017). The results are presented in Figure 9, which displays the two first (most important) dimensions of the MCA analyses. The first and second dimensions together capture around 50% of the variability in the data set. The third, fourth, ..., nth dimension explain a lower and lower proportion and are not further considered here.

Figure 9 suggests that we can identify one main cluster of languages from the MCA analysis, that is a triangle-shaped cluster to the center-left. The rest of the data points are spread throughout the center-right of the typological space delimited by the first dimension. We attempted to link these results to the four-way typology discussed in the main text using color-coding. We can observe that the center-left triangle cluster contains languages with only syntactic agreement or a combination of syntactic and animacy-based agreement. The languages of these two types show greater similarity to each other as compared the two other types (languages with only animacy-based gender and languages with no gender). These latter two types are scattered throughout the remainder of the space and display no clustering. The second MCA dimension that is plotted onto the y-axis distinguishes between languages with only syntactic agreement (in black) as they have a negative loading on Dimension 2, and languages with both syntactic and animacy-based agreement (in blue) that load positively on Dimension 2. Since these two MCA Dimensions seem to capture variation in gender marking systems quite nicely, we include them here as an alternative dependent measure to report on.

Figure 10 displays the results of these analyses. The model using the first MCA dimension as independent measure has two relevant (positive) effects, longitude and current rainforest overlap; the model using the second MCA dimension as independent measure has very wide predictor posterior distributions, with no relevant predictors. When we add random slopes to model MCA_Dim1_Glot_int for longitude (MCA_Dim1_Glot_int_slope_longitude) and current rainforest overlap
(MCA_Dim1_Glot_int_slope_RF_overlap), the effects are no longer significant, matching the pattern that we find in section 7.

From this we conclude that the second MCA dimension does not carry enough signal regarding the type of gender system; as can be observed in Figure 9, many languages with only syntactic agreement and languages with both syntactic and animacy-based agreement are very condensed in a small part of the Figure. The first MCA dimension, however, is an interesting depen-

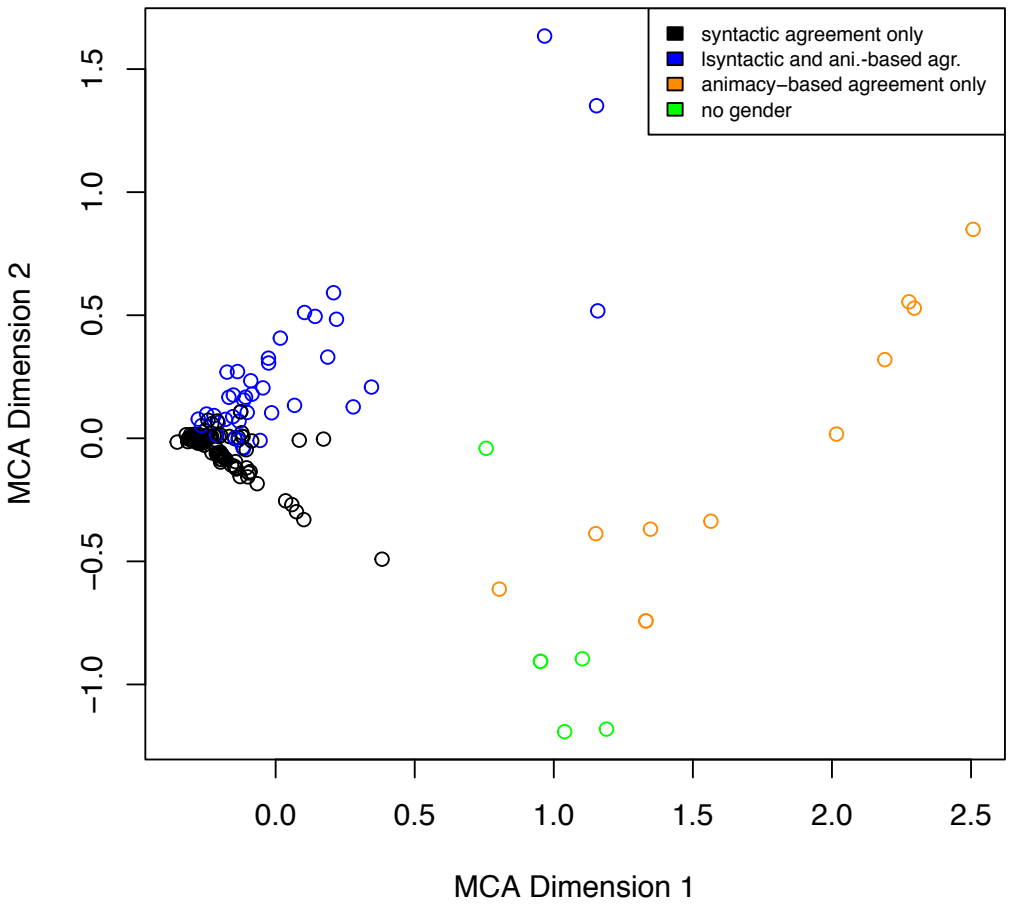

Figure 9: First two dimensions from multiple correspondence analysis (MCA) on the entire questionnaire including the additional questions. The first dimension (x-axis) captures 38% of the variance, the second dimension (y-axis) captures 12%. This Figure has been taken from: Di Garbo, Francesca, and Annemarie Verkerk. 2022. 'A Typology of Northwestern Bantu Gender Systems'. *Linguistics*.

dent measure that correlates highly with the four-way typology we proposed in Section 2.3 in the main text. It is unsurprising therefore that the MCA_Dim1_Glot_int model shows an effect for current rainforest overlap, just as the bin_Glot_int reported on in Figure 6 in the main text. However, the languages with only animacy-based agreement or no gender at all seem to be driving this effect, and the MCA Dimensions do not capture differences between languages with only syntactic agreement or both syntactic and animacy-based agreement. The questionnaire (Appendix A in the main text) actually emphasizes differences between the former types because certain questions are only relevant for languages with only animacy-based gender marking or no gender at all, and thus the differences between these languages receive undue weight. While we report the results here to indicate that we are open to quantifying gender systems in various different ways, we also think that not all of these different ways may do justice to the nature of the data. In sum, using MCA to construct continuous measures of NWB gender marking seems useful for obtaining information about the differences between languages with heavily restructured gender systems, but not as a more comprehensive typology of NWB gender systems.

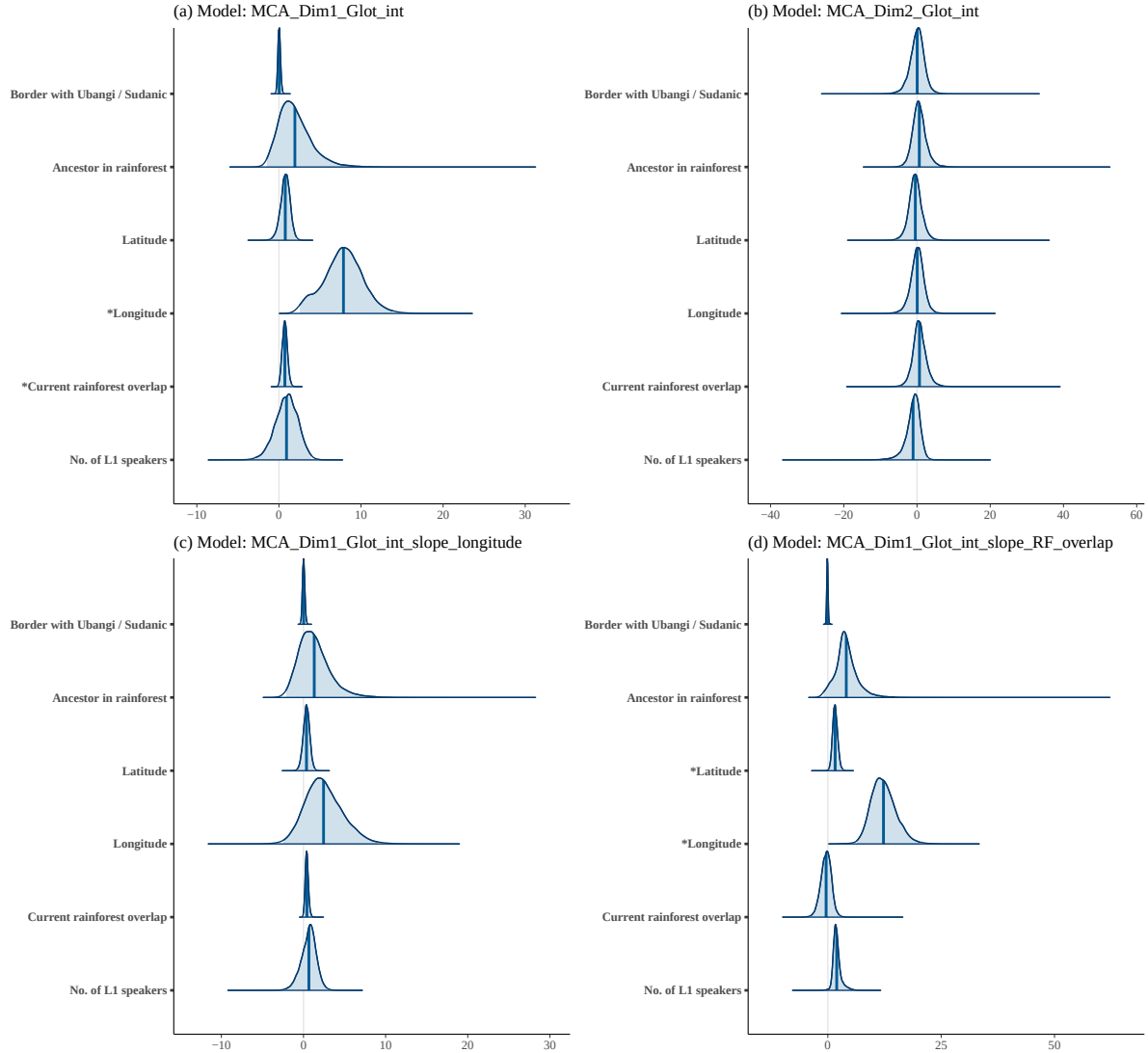

Figure 10: Comparison of posterior distribution of fixed effects of models using the first and second MCA Dimension, using a random intercept using Glottolog groupings (MCA_Dim1_Glot_int and MCA_Dim2_Glot_int), as well as an additional random slope on selected predictors (MCA_Dim1_Glot_int_slope_longitude and MCA_Dim1_Glot_int_slope_RF_overlap)

References

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

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68. 255–278. doi:https://doi.org/10.1016/j.jml.2012.11.001.

Bastin, Yvonne, A. Coupez & M. Mann. 1999. *Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study*. Tervuren: Royal Museum for Central Africa.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paul Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *Plos One* 10. 1–23.

Bentz, Christian & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3. 1–27.

Bostoen, Koen. 2019. Reconstructing Proto-Bantu. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu languages. 2nd edition*, 308–334. London: Routledge.

Bostoen, Koen & Claire Gregoire. 2007. La question Bantoue: bilan et perspectives. *Mémoires de la Société de linguistique de Paris* XV. 73–91.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28. doi:10.18637/jss.v080.i01.

Carpenter, Bob, Andrew Gelman, Matthew D. 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* 76(1).

Coupe, 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.

Di Garbo, Francesca & Annemarie Verkerk. Accepted, 2021. A typology of northwestern Bantu gender systems. To appear in *Linguistics*.

Gelman, Andrew. 2019. *Prior choice recommendations*. GitHub website.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau & Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4). 1360–1383.

Grollemund, Rebecca, Simon Brandford, Koen Bostoen, Andrew Meade, Chris Venditti & Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Science of the United States of America* 112. 13296–13301.

Hammarström, Harald, Robert Forkel & Martin Haspelmath (eds.). 2018. *Glottolog 3.3*. Jena: Max Planck Institute for the Science of Human History. (Available online at http://glottolog.org, Accessed on 2017-05-11.).

Ho, Lam Si Tung & Cecile Ane. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* 63. 397–408.

Jaeger, Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15. 281–320.

Koile, Ezequiel, Simon J. Greenhill, Damian E. Blasi, Remco Bouckaert, Tom Guldemann, Patrick Roberts & Russell D. Gray. submitted. Phylogeographic analysis of the Bantu language expansion supports a rainforest route .

Lê, Sébastien, Julie Josse & François Husson. 2008. Factominer: An r package for multivariate analysis. *Journal of Statistical Software* 25(1). 1–18.

Lewis, M. P., Gary F. Simons & Charles D. Fenning (eds.). 2016. *Ethnologue: Languages of the world, nineteenth edition*. Dallas: SIL International. Online version: http://www.ethnologue.com.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5(1). 1–10.

Maas, Cora J. M. & Joop J. Hox. 2005. Sufficient sample sizes for multilevel modeling. *Methodology* 1. 86–92.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates. 2017. Balancing type i error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

Orme, David. 2011. *The caper package: Comparative analysis of phylogenetics and evolution in r*.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401. 877–884.

Piantadosi, Steven T. & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38. 736–756. doi:https://doi.org/10.1111/cogs.12088.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Com-

puting Vienna, Austria. http://www.R-project.org/. ISBN 3-900051-07-0.
Scherbaum, Charles A. & Jennifer M. Ferreter. 2009. Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods* 12. 347–367. doi: https://doi.org/10.1177/1094428107308906.
Seedorff, Michael, Jacob Oleson & Bob McMurray. 2019. Maybe maximal: Good enough mixed models optimize power while controlling type i error. *PsyArXiv* doi:https://doi.org/10.31234/osf.io/xmhfr.
Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. doi:10.3389/fpsyg.2018.01141. https://www.frontiersin.org/article/10.3389/fpsyg.2018.01141.
Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing* 27. 1413–1432.
Winter, Bodo. 2020. *Statistics for linguists: An introduction using r.* New York: Routledge.