1 Appendix

1.1 Areal effects

Visualizing the uncertainty in the geographical effects within macroareas is very difficult. In this section we try to give an intuition to the amount of uncertainty in our estimates. The independent effects of longitude and latitude by macroarea are shown in Figures 1 and 2, respectively. In these figures each line represents the probability of each outcome. The ribbons represent the uncertainty intervals.[1] The x-axis represents the change in longitude and latitude, respectively. The longitude and latitude values are centered to make them comparable across macroareas. The value of 0 corresponds to the center point of each macroarea. Both Figure 1 and 2 reveal a very high degree of uncertainty in most of the estimates. Although the model can pick up some clear trends in most macro-areas, some of these trends do not have enough evidence for the model to be confident about them. This is especially clear in North America, where there is a large overlap in the uncertainty intervals.

It is important to reiterate that these are not absolute predictions, but rather predictions given that all other covariates are held constant. This means that the predictions for longitude are done for a single value of latitude, and the predictions of latitude are done for a single value of longitude. The implication is that we cannot conclude that the effects are insignificant, or do not matter,

1 We chose to use ribbons instead of using posterior samples, since the latter would make the plots too cluttered.

since these plots do not show the interaction between latitude and longitude. What these plots do show, is that on towards the regions with few languages (i.e. the edges of the plots), the uncertainty is much greater than towards the center. This is important because in the main paper the plots show predictions for areas without languages. Predictions outside areas with a high density of languages are highly uncertain.

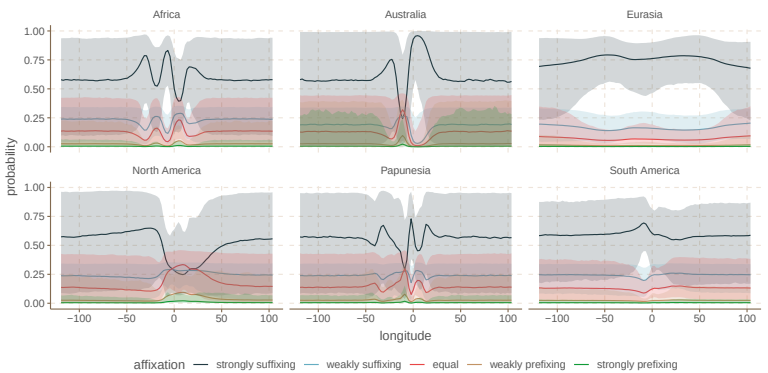

Fig. 1: Longitude by macroarea

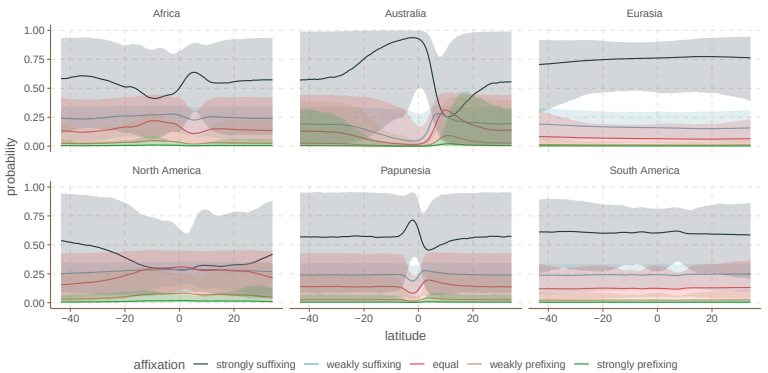

Fig. 2: Latitude by macroarea

1.2 Model checks

In order to argue for our interpretation of the results of the main model, we illustrate a situation in which we want to predict the effect of z on y, but the underlying, real cause of both y and z is an additional variable x. This mimicks the situation of phylogenetic and areal effects, likely influencing both word order and affixation patterns, which appear to be associated if the biases are not controlled for. We assume that the real distributions of these three variables is as follows:[2]

$$x \sim Uniform(0, 10) \quad (1)$$

$$z = sin(x) + \epsilon_1 \quad (2)$$

$$y = sin(x) * 2 + \epsilon_2 \quad (3)$$

$$\epsilon_1 \sim N(0, 0.2) \quad (4)$$

$$\epsilon_2 \sim N(0, 0.3) \quad (5)$$

In other words, both y and z are a function of x plus some normally distributed noise. With this data distribution, we could try to build a model (Model 1) in which we believe that z is the cause of y as: y $\sim$ z.[3] The conditional effects of Model 1 are shown in Figure 3. As can be observed, the model estimates that z has a clear linear effect on y, with very narrow uncertainty intervals.

2 This is the same data we used for Figure **??**.

3 We use formula notation here for simplicity. The full model specification including priors is in the supplementary materials.

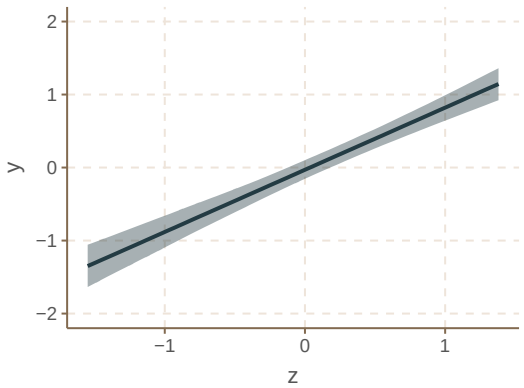

Fig. 3: Conditional effects of in Model 1

However, if we include x in our model, the effect of z on y changes considerably. Model 2 is defined as follows: y ∼ z + gp(x). The conditional effects of Model 2 are shown in Figures 4 and **??**. After adding x as a non-linear predictor, the model still estimates a positive correlation between z and y, but now the uncertainty intervals in Figure 4 are very large compared to the ones in Model 1, shown in Figure 3. Thus, we would conclude that there is no strong evidence for a real effect of z on y.

What these toy models show is that collinear predictors can easily lead to situations in which a model has high uncertainty about the effects of the predictors, even if the sampler has no trouble with the estimates. In this case, it is hard for the model to disentangle the fact that z is produced by x, and thus, should have no effect. As a consequence, we see large uncertainty intervals for the effect of z. While we cannot be completely certain, this is also a likely explanation for what we see in our models: areal and phylogenetic effects are likely associated with both word order and affixation preferences. Thus, having

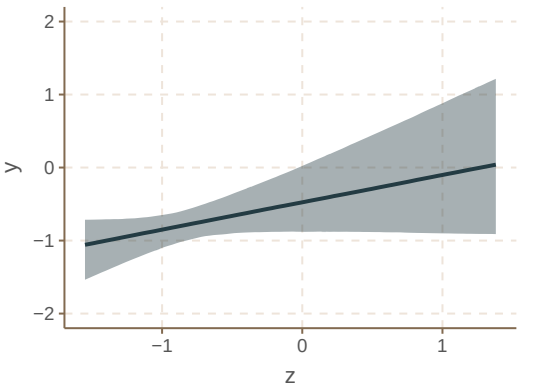

Fig. 4: Conditional effects of in Model 2

large uncertainty intervals when predicting affixation from word order does not necessarily mean that the model struggles to estimate the affixation preferences from the information given. Rather, it struggles to disentangle the effect of word order on affixation preference, because both are associated with areal and phylogenetic effects in a similar way.