

Reassessing scale effects on differential case marking
Karsten Schmidtke-Bode and Natalia Levshina (Leipzig University)

Supplementary material SM3: Technical information

This document contains additional details about the statistical models reported in the paper.

Software. To perform the analyses, we used *R*, version 3.4.0 (R Development Core Team 2017) with the add-on packages *lme4* (Bates et al. 2015) and *MuMIn* (Barton 2016).

Variable coding. All variables are categorical. However, Helmert contrasts were used to represent the *Rank* variable in the rank model. This means that the algorithm compared the effect of each rank on the scale (3) with the average effect of all preceding ones, i.e. the 3rd person with the 1st and 2nd, high-prominence nouns with all pronouns (1st, 2nd and 3rd persons) and low-prominence nouns with all the rest. All other variables were represented internally by treatment contrasts.

Variable selection. To decide whether or not main effects, interaction terms and random effects should be included in a model, we used the likelihood ratio test.

Model diagnostics. Both models were tested for overdispersion, with a negative result. In the rank model, there were a few observations with unusually high Pearson residuals (> 5), which belonged to the languages Finnish and Osage. To test their influence, these languages were removed and the model was fitted again. The difference between the original and new model was very small and non-significant, so the observations were kept in the model.

Goodness of fit. The type model's marginal R^2 , which is based on the fixed effects only (i.e. *MacroContinent* and *Fit*), was 0.33, whereas the conditional R^2 , which also takes the families into account, was 0.44. As for the rank model, the discriminating power of the model was almost perfect ($C = 0.98$). To compare, the discriminating power of the model without the random effects, which revealed very similar patterns, was also very high: $C = 0.904$.

Tables of regression coefficients.

Table 1. Type model (I): Poisson GLM

Term	b (log factor)	b (factor)	SE	p -value
Intercept	-0.127	0.881	0.336	0.706
Fitting = Fit	1.427	4.166	0.219	< 0.0001
Microcontinent = Africa	-1.966	0.14	0.406	< 0.0001
Microcontinent = Americas	-1.376	0.252	0.445	0.002
Microcontinent = Sahul	-0.891	0.41	0.481	0.064
Random intercepts for family	$\sigma^2 = 0.75, \sigma = 0.866.$			

Table 2. Type model (II): Logistic GLM (quasibinomial)

Term	b (log odds)	b (odds)	SE	p -value
Intercept	1.442	4.229	0.26	<0.0001

Table 3. Rank model

Term	b (log odds)	b (odds)	SE	p -value
Intercept	1.153	3.168	0.602	0.056
Rank = 2	0.385	1.47	0.244	0.114
Rank = 3	-0.296	0.744	0.137	0.03
Rank = 4	-2.54	0.079	0.322	< 0.0001
MacroCont = Africa	-1.15	0.317	0.84	0.171
MacroCont = Americas	-0.272	0.762	0.845	0.747
MacroCont = Sahul	-1.18	0.307	0.913	0.196
Number = SG	0.869	2.385	0.256	0.0007
Rank = 2 : Number = SG	-0.417	0.659	0.312	0.181
Rank = 3 : Number = SG	-0.435	0.647	0.157	0.006
Rank = 4 : Number = SG	-0.121	0.886	0.19	0.526
Rank = 2 : MacroCont = Africa	-0.597	0.55	0.426	0.161
Rank = 3 : MacroCont = Africa	-0.659	0.517	0.263	0.012
Rank = 4 : MacroCont = Africa	1.708	5.518	0.348	< 0.0001
Rank = 2 : MacroCont = Americas	-0.112	0.894	0.532	0.833
Rank = 3 : MacroCont = Americas	-0.529	0.589	0.262	0.043
Rank = 4 : MacroCont = Americas	0.969	2.635	0.357	0.007
Rank = 2 : MacroCont = Sahul	-0.746	0.474	0.367	0.042
Rank = 3 : MacroCont = Sahul	-1.067	0.344	0.211	< 0.0001
Rank = 4 : MacroCont = Sahul	0.833	2.3	0.362	0.021
Random intercepts for family	$\sigma^2 = 1.576, \sigma = 1.255.$			
Random intercepts for system	$\sigma^2 = 2.518, \sigma = 1.587.$			

References

- Barton, Kamil (2016). MuMIn: Multi-Model Inference. *R* package version 1.15.6. <<https://CRAN.R-project.org/package=MuMIn>>.
- Bates, Douglas, Martin Maechler, Ben Bolker and Steve Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67.11: 1–48. doi:10.18637/jss.v067.i01.