**Details about Bayesian models**

Models with four Markov chains were tried out in order to test convergence, which was acceptable. Because of some autocorrelation in the chains, for the final model we created one very long chain with 20,000 iterations (2,000 in the warm-up phase) and the thinning parameter of 3, which means that only every third posterior sample was considered. This produced the total of 6,000 posterior samples. Also, due to some divergent transitions, the adaptive parameter delta was increased to 0.99. This slowed down the Markov chain, but ensured that the posterior samples were valid. All R-hat values were 1.00, and the effective sample sizes were sufficiently large to produce reliable posterior distributions, which are summarised in Section 4.

Bayesian inference involves the use of priors, which inform the model of our expectations before we look at the data. In our case, we chose to use regularizing, or weakly informative Cauchy priors centred around 0 with the scale 2.5, which is common practice in logistic regression, where regression coefficients greater than 7 or smaller than -7 are uncommon (at least, in datasets of the size like ours). These priors were applied to fixed-effects coefficients and standard deviations of the random effects. This helps to improve inference by pushing the coefficients somewhat towards realistic values without imposing subjective expectations (Nicenboim & Vasishth, 2016).

Some additional models were fitted for diagnostics and sanity checks. Potential non-linear effects of the numeric variables were tested with the help of thin plate regression splines, but no convincing non-linearity was detected. Since some informativity measures could be unreliable due to low frequencies in the denominators of the formulas in (2) and (3), an additional model was fitted based on the data without all words on the left and all infinitives with frequency less than 5. Removing these observations did not result in any substantial differences.

The models fit the data well. The first model based on the small sample with complete observations has the following goodness-of-fit measures: the Bayesian pseudo-*R2*, which ranges from 0 (useless model) to 1 (perfect fit), was 0.48, with the 95% credible interval from 0.45 to 0.51. The *C*-index, which is traditionally used for evaluation of logistic models and which ranges from 0.5 to 1, is 0.909, with the 95% credible interval from 0.900 and 0.918. The second model based on the large dataset with imputed values fits the data somewhat better: Its Bayesian pseudo-*R2* is 0.54 (95% credible interval 0.53 – 0.56), and the *C*-index is 0.938 (95% credible interval 0.934 – 0.943).