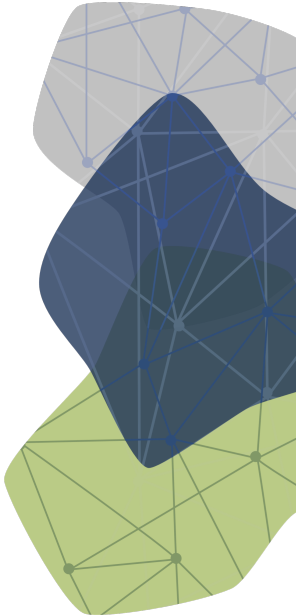
DGfS 2026 AG7: More than just noise – Detecting patterns in acceptability judgment data

Lectal Variants as Linear Predictors: A Case Study

Simon Masloch



LDSL
Linguistic Data Science Lab
RUHR UNIVERSITÄT BOCHUM
RUB

25 February, 2026, Trier

Introduction

- Participants in linguistic experiments differ.
- Differences may be interesting, e.g.,
  - … in themselves: In the most extreme case one may need different grammars for the phenomenon at hand.
  - … because properties of the items not directly under investigation may lead to a rejection by some participants, obscuring the patterns one is mainly interested in.
- Proposal: Components for statistical models in which potential differences between participants are directly reflected in assuming that fundamentally different linear predictors are relevant for participants belonging to different groups.
- Purposes:
  - Get a clearer picture of patterns where some participants may reject items for independent reasons.
  - Measure share of participants with different grammar (or at least acceptability pattern)

Example Data

The main example data for this talk will come from (Masloch & Poppek & Kiss 2025, henceforth MPK)

Acceptability judgment study on reflexive binding into the subjects of German experiencer-object (EO) verbs:

(3) (Masloch & Poppek & Kiss 2025: 206)

a. Es ist offensichtlich, dass [das Gerücht [über sich$_1$]]$_S$ [den Professor$_1$]$_O$ genervt hat.
it is obvious that the.NOM rumour about REFL the.ACC professor annoyed has

b. Es ist offensichtlich, dass [den Professor$_1$]$_O$ [das Gerücht [über sich$_1$]]$_S$ genervt hat.
'It is obvious that the rumour about himself annoyed the professor.'

MPK

2 × 2 Acceptability / ‘naturalness’ rating study (5-point scale)

- Order: SO, OS
- Case: of object. Accusative or dative.

Every item in both order conditions, but participants saw one only.
Each items contained either dative or accusative EO verb, participants saw both kinds of verbs.

Materials:

- 8 test items containing accusative-object EO verb, 8 containing dative EO verb
- 64 (related and unrelated) filler items:
  - 6 calibration items
  - 16 control items
  - 10 attention items

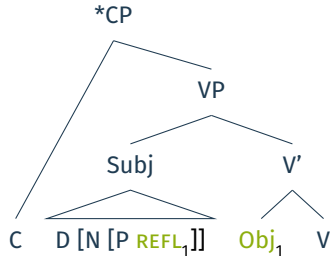
Hypotheses and Predictions MPK

- Reflexive embedded in subject, object only possible antecedent.

(4) Main Hypothesis MPK (p. 209)
In the German midfield, the object of an experiencer-object verb cannot bind a reflexive embedded in a subject preceding it.

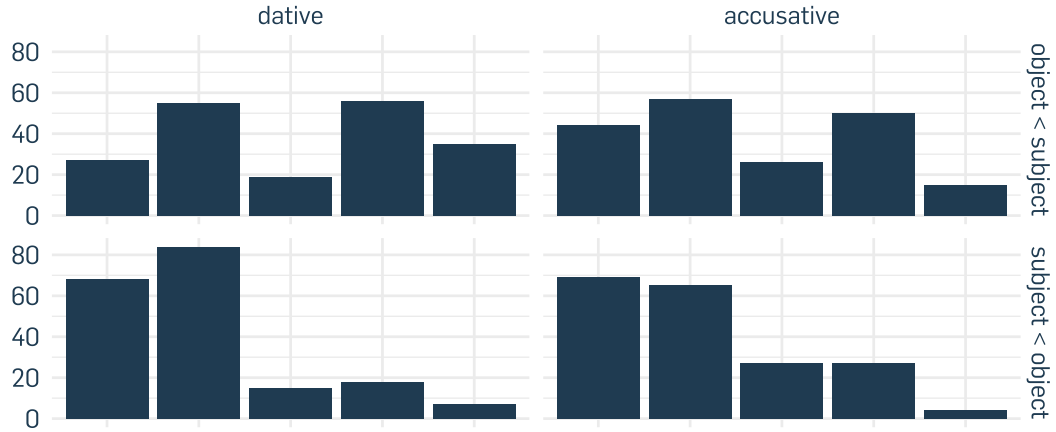
- Depending on views on German clausal syntax, the syntax of EO verbs and reflexive binding, various different predictions possible, but not relevant for this talk (see MPK on that).

(5)



Results MPK

90 participants, 48 surveys entered analysis



The Standard Model

- In MPK, we mainly used a Bayesian cumulative generalised linear mixed model with logit link and flexible thresholds:

```
ANSWER ~ case * order +
(1 + case * order | participant) +
(1 + order | item)
```

- Factors sum-coded: *dative* and *OS* as 1, *accusative* and *SO* as –1:
  - β_{ORDER}: overall effect of ORDER
  - β_{CASE}: overall difference between accusative and dative
  - $\beta_{\text{ORDER×CASE}}$: positive value would correspond to preference for normal order irrespective of other factors including binding constraints (see MPK)

CASE \ ORDER	*OS*	*SO*
dat	$\beta_{\text{ORDER}} + \beta_{\text{CASE}} + \beta_{\text{ORDER×CASE}}$	$-\beta_{\text{ORDER}} + \beta_{\text{CASE}} - \beta_{\text{ORDER×CASE}}$
acc	$\beta_{\text{ORDER}} - \beta_{\text{CASE}} - \beta_{\text{ORDER×CASE}}$	$-\beta_{\text{ORDER}} - \beta_{\text{CASE}} + \beta_{\text{ORDER×CASE}}$

Standard Model: $M_{standard}$

$$\text{RATING}_n \sim Categorical(\xi_{n,1}, \ldots, \xi_{n,5}) \tag{1}$$

$$\xi_{n,i} = \begin{cases} logit^{-1}(\tau_i - \eta_n) & \text{if } i = 1 \\ logit^{-1}(\tau_i - \eta_n) - logit^{-1}(\tau_{i-1} - \eta_n) & \text{if } 1 < i < 5 \\ 1 - logit^{-1}(\tau_{i-1} - \eta_n) & \text{if } i = 5 \end{cases}$$

$$\eta_n = u_{part[n],1} + w_{item[n],1} + \text{ORDER}_n \times (\beta_{\text{ORDER}} + u_{part[n],2} + w_{item[n],2}) +$$
$$\text{CASE}_n \times (\beta_{\text{CASE}} + u_{part[n],3}) + \text{CASE}_n \times \text{ORDER}_n \times (\beta_{\text{CASE:ORDER}} + u_{part[n],4})$$

n: data point number, $part[n]$, $item[n]$: participant / item of data point

$M_{standard}$: Prior

- Bayesian Model: Parameters are random variables ⇒ One can talk about the credibility of different values
- Prior reflects how likely the parameter values are considered to be before taking data into account
- Posterior captures one's updated beliefs after seeing the data. Described here by estimate (mean, $\hat{\cdot}$) and 95 % credible interval.

$$
\begin{aligned}
\tau_1 \sim N(-3, 7.5), \tau_2 \sim N(-1, 7.5), \\
\tau_3 \sim N(1, 7.5), \tau_4 \sim N(3, 7.5) \qquad (2) \\
\beta_{\text{ORDER}},\ \beta_{\text{CASE}}, \beta_{\text{CASE:ORDER}} \sim N(0, 4) \\
\begin{pmatrix} w_{i,1} \\ w_{i,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_w \right) \\
\Sigma_w = \begin{pmatrix} \sigma^2_{w_1} & \rho_w \sigma_{w_1} \sigma_{w_2} \\ \rho_w \sigma_{w_1} \sigma_{w_2} & \sigma^2_{w_2} \end{pmatrix} \\
\sigma_{w_1},\ \sigma_{w_2} \sim Student\_t_+(3, 0, 2.5) \\
\begin{bmatrix} 1 & \rho_w \\ \rho_w & 1 \end{bmatrix} \sim LKJcorr(1)
\end{aligned}
$$

$M_{standard}$: Prior

$$\begin{pmatrix} u_{i,1} \\ u_{i,2} \\ u_{i,3} \\ u_{i,4} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_u\right)$$

$$\Sigma_u = \begin{pmatrix} \sigma^2_{u_1} & \rho_{u1,u2}\sigma_{u_1}\sigma_{u_2} & \rho_{u1,u3}\sigma_{u_1}\sigma_{u_3} & \rho_{u1,u4}\sigma_{u_1}\sigma_{u_4} \\ \rho_{u1,u2}\sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2} & \rho_{u2,u3}\sigma_{u_2}\sigma_{u_3} & \rho_{u2,u4}\sigma_{u_2}\sigma_{u_4} \\ \rho_{u1,u3}\sigma_{u_1}\sigma_{u_3} & \rho_{u2,u3}\sigma_{u_2}\sigma_{u_3} & \sigma^2_{u_3} & \rho_{u3,u4}\sigma_{u_3}\sigma_{u_4} \\ \rho_{u1,u4}\sigma_{u_1}\sigma_{u_4} & \rho_{u2,u4}\sigma_{u_2}\sigma_{u_4} & \rho_{u3,u4}\sigma_{u_3}\sigma_{u_4} & \sigma^2_{u_4} \end{pmatrix}$$

$$\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3}, \sigma_{u_4} \sim Student\_t_+(3, 0, 2.5)$$

$$\begin{bmatrix} 1 & \rho_{u1,u2} & \rho_{u1,u3} & \rho_{u1,u4} \\ \rho_{u1,u2} & 1 & \rho_{u2,u3} & \rho_{u2,u4} \\ \rho_{u1,u3} & \rho_{u2,u3} & 1 & \rho_{u3,u4} \\ \rho_{u1,u4} & \rho_{u2,u4} & \rho_{u3,u4} & 1 \end{bmatrix} \sim LKJcorr(1)$$

Results $M_{standard}$

Model fit with *Stan* (Stan Development Team 2024) via *brms* (Bürkner 2017) in R (R Core Team 2023) and sensitivity analyses etc. conducted in MPK. Now re-implemented in PyMC (Martin & Kumar & Lao 2021) for comparison with other models.

Bayes Factor (BF): $\frac{P(Data \mid Model_1)}{P(Data \mid Model_2)}$ Summary of evidence provided by data for one model over the other.
BF_{10}: BF of model containing effect vs. model where it is set to 0.

- Population-level effects:
  - $\hat{\beta}_{\text{ORDER}}$ = 0.79 [0.47, 1.12], BF_{10} = 126.7: **Ratings are better if binding not backward**
  - $\hat{\beta}_{\text{CASE}}$ = 0.16 [−0.24, 0.57], BF_{10} = 0.074: Evidence against effect of CASE.
  - $\hat{\beta}_{\text{ORDER×CASE}}$ = 0.23 [−0.06, 0.52], BF_{10} = 0.138. Evidence against interaction effect.
- **Participants differ strongly**: intercepts: $\hat{\sigma}_{u_1}$ = 1.57 [1.22, 2], ORDER: $\hat{\sigma}_{u_2}$ = 0.5 [0.26, 0.75],

For comparison: $\hat{\tau}_1$ = −1.6 [−2.23, −0.98], $\hat{\tau}_2$ = 0.66 [0.05, 1.28], $\hat{\tau}_3$ = 1.49 [0.87, 2.12], $\hat{\tau}_4$ = 3.74 [3.06, 4.46]

Unattractive Aspects of $M_{standard}$

- Variation between participants assumed effectively spans the whole scale
- Even in the condition we would expect to be grammatical, estimate is mediocre
- Explorative investigation shows: Participants with individual intercepts which would mean that they must consider everything bad do *not* reject all filler items

Possible reason: Test items contained [D [N [P REFL]]] structures (6), which are problematic in that they often do not sound very natural.

(6) das Gerücht über sich
the rumour about REFL

N P REFL

We took great care to use only N P REFL sequences sounding natural.

Still: May be a problem and some of the participants rejecting all test items reject the four fillers containing N P REFL sequences, too (e.g. (7))!

(7) Maryam hat behauptet, dass sie eine Tante von sich in Venedig getroffen hat.
Maryam has claimed that she a aunt of REFL in Venice met has
'Maryam claimed that she met an aunt of her's in Venice.'

Idea Alternative Model

What if there is lectal variation and some speakers simply reject [N [P REFL]] altogether?

Standard Model has means to account for this via the the group-level effects (⇒ Large estimates for their SDs!)

But we can build models that reflect this assumption directly: Each participant may either belong to a group rejecting all test items or to the 'normal' group.

Expected Advantage: Responses of participants rejecting everything do not influence main parameters of interest anymore.

Lectal Variants as Linear Predictors

- π: Overall rate of participants who consider all test items unnatural. We may assume an uninformative prior like $Beta(1, 1)$
- $b \sim Bernoulli(\pi)$: vector of Booleans, indicating for each participant which group they belong to
- $\eta_n = \begin{cases} \eta_{1,n} & \text{if } b_{part[n]} = 1 \\ \eta_{2,n} & \text{if } b_{part[n]} = 0 \end{cases}$
- $\eta_{1,n} = u_{part[n],1} + w_{item[n],1} + \alpha_{bad}$
- $\eta_{2,n} = u_{part[n],1} + w_{item[n],1} + \text{ORDER}_n \times (\beta_{\text{ORDER}} + u_{part[n],2} + w_{item[n],2}) + \text{CASE}_n \times (\beta_{\text{CASE}} + u_{part[n],3}) + \text{CASE}_n \times \text{ORDER}_n \times (\beta_{\text{CASE:ORDER}} + u_{part[n],4})$

Including Filler Item Data

- The information that not everything is rejected is contained in the filler items.
- Only data from fillers in narrow sense, no medium acceptability fillers
- Data for filler items treatment-coded (1 if (un)acceptable filler, 0 for the other and test items). Other effects 0 for fillers.
- We can then assume the following as η_2, the linear predictor for the 'normal' participants::

$$u_{part[n],1} + w_{item[n],1} + \text{ORDER}_n \times (\beta_{\text{ORDER}} + u_{part[n],2} + w_{item[n],2}) + \text{CASE}_n \times (\beta_{\text{CASE}} + u_{part[n],3}) \quad (3)$$
$$+\text{CASE}_n \times \text{ORDER}_n \times (\beta_{\text{CASE:ORDER}} + u_{part[n],4}) + \text{ACCEPT}_n \times \beta_{\text{ACCEPT}} + \text{REJECT}_n \times \beta_{\text{REJECT}}$$

- Priors: $\beta_{\text{ACCEPT}} \sim N(3, 1.5)$, $\beta_{\text{REJECT}} \sim N(-3, 1.5)$
- η_1 (for the participants rejecting everything) could then be:

$$u_{part[n],1} + w_{item[n],1} + \beta_{\text{REJECT}} + \text{ACCEPT}_n \times (\beta_{\text{ACCEPT}} - \beta_{\text{REJECT}}) \quad (4)$$

(If test items are really ungrammatical for these participants, they should be as bad as unacceptable fillers, so β_{REJECT} will be relevant for them. For acceptable fillers this effect has to be subtracted again.)

Results M_{diff}

I will call this model M_{diff}. Fit with PyMC (Martin & Kumar & Lao 2021) (4 chains, 30 000 draws per chain, using default Metropolis-within-Gibbs sampler for *b*, the default No-U-Turn sampler for the other parameters)

- Estimates for the thresholds are a bit smaller than in $M_{standard}$ overall: $\hat{\tau}_1 = -1.82\ [-2.33, -1.32]$, $\hat{\tau}_2 = 0.32\ [-0.16, 0.8]$, $\hat{\tau}_3 = 1.06\ [0.58, 1.56]$, $\hat{\tau}_4 = 2.95\ [2.43, 3.47]$
- $\hat{\beta}_{\text{ORDER}} = 0.8\ [0.58, 1.04]$, $BF_{10} = 60.77$
- $\hat{\beta}_{\text{CASE}} = -0.06\ [-0.39, 0.26]$, $BF_{10} = 0.04$
- $\hat{\beta}_{\text{CASE:ORDER}} = 0.26\ [0.1, 0.43]$, $BF_{10} = 1.77$
- $\hat{\beta}_{\text{REJECT}} = -1.9\ [-2.46, -1.35]$, $BF_{10} = 82.54$, $\hat{\beta}_{\text{ACCEPT}} = 4.02\ [3.3, 4.76]$, $BF_{10} = 79.55$
- $\hat{\pi} = 0.2\ [0.06, 0.36]$ (the posterior has a bit positive skew, but not much)
- $\hat{\sigma}_{u_1} = 0.9\ [0.65, 1.15]$

Learnings Binding

- For some (ca. 20 %) of the participants, the factors tested are irrelevant as they reject all test items
- Even if this is accounted for, the test items in the OS condition are not rated as fully natural
- But they are considerably more natural than the test items in the SO condition and unacceptable fillers, and we can be very sure about that.

The fact that the test items are not fully natural even when the behaviour of the participants who reject everything is accounted for needs to be explained.

Arguably account of acceptability of N P REFL in diverse settings needed, but this goes beyond this talk.

Additional Example

Crossover Violations English Relative Clauses

Longstanding conflicting judgments about the existence of weak crossover (WCO) effects in English restricted relative clauses in examples like (8a) (Howitt & Scontras & Polinsky 2025).

(8) (adapted from Howitt & Scontras & Polinsky 2025: 5)

a. The plane$_i$ which$_i$ its$_i$ pilot t_i flew around the country avoided turbulence.

b. The plane$_i$ which$_i$ the pilot flew t_i around the country avoided turbulence.

The analysis of the data by (Howitt & Scontras & Polinsky 2025) here strongly builds on a re-analysis of that data by Tibor Kiss.

Additional Example

Crossover Violations English Relative Clauses

Howitt & Scontras & Polinsky (2025) conducted a slider-scale acceptability judgment study on the phenomenon:

- 24 test items (2 of which excluded), 6 fillers only for attention checks
- 3 factors: WCO (yes or no), ANIMACY (of relative clause head), DETERMINER (R-expression or quantificational), but ANIMACY and DETERMINER turned out not to have an effect
- data from 78 participants considered in analysis

Authors checked for differences between participants (not their main interest) by looking for bimodality in a histogram. For them there seems to be none (but hard to tell in my opinion).

Crossover Violations: Lectal Variants

Some reason to assume that there could be different groups of participants: wco relevant for some but not others.

No (non-attention-check) filler data that could be used.

I ran linear models as well as (for this kind of data arguably more appropriate) Zero-one-inflated Beta (ZOIB, see i.a. Ospina & Ferrari 2010; Liu & Eugenio 2018) models on the data with latent discrete variables governing the choice of linear predictor as before. Linear predictors for the linear model (ZOIB model analogous, same formula for mean, φ, α, γ):

$$\eta_1 = \alpha + u_{part[n],1} + w_{item[n],1} + \text{WCO}_n \times (\beta_{\text{WCO}} + u_{part[n],2} + w_{item[n],2}) \quad (5)$$

$$\eta_2 = \alpha + u_{part[n],1} + w_{item[n],1} \quad (6)$$

Crossover Violations: Lectal Variants

- $\pi \sim Beta(1, 1)$
- $b \sim Bernoulli(\pi)$
- $\eta_n = \begin{cases} \eta_{1,n} & \text{if } b_{part[n]} = 1 \\ \eta_{2,n} & \text{if } b_{part[n]} = 0 \end{cases}$

Results:

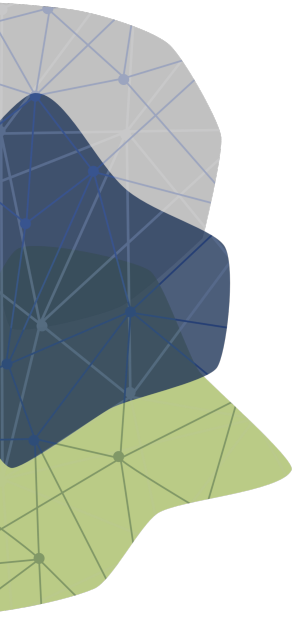
- Parameters for population-level effects do not differ substantially from standard models, though they are a bit more pronounced for the ZOIB model
- participants' group-level effect standard deviations, too
- $\hat{\pi} = 0.71$ [0.51, 0.92] (linear model), $\hat{\pi} = 0.67$ [0.51, 0.84] (ZOIB), indicating that there is no WCO effect in this setting for ca. ⅓ of the participants

Conclusion

I discussed a way to handle assumed differences between participants in statistical models via different linear predictors and latent discrete variables governing group membership.

Such models would probably not be the first or only ones one uses, but arguably they can help us getting a better understanding of the data and identifying linguistically relevant patterns:

- In the binding study, they showed that the test items in the arguably grammatical condition are not fully acceptable even when it is taken into account that some participants reject them for presumably independent reasons.
- For WCO in English relative clauses the models based on data by Howitt & Scontras & Polinsky (2025) suggest that a significant share of speakers does not get the effect, which is in line with the patterns in the theoretical literature and may itself need an explanation.



Lectal Variants as Linear Predictors: A Case Study

Thank you for your attention

Simon Masloch
simon.masloch@rub.de

LDSL
Linguistic Data Science Lab

RUHR
UNIVERSITÄT
BOCHUM

RUB

References

- Bürkner, Paul-Christian. 2017. Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). https://doi.org/10.18637/jss.v080.i01.
- Diewald, Nils et al. 2016. KorAP architecture: Diving in the Deep Sea of Corpus Data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 3586–3591. Portorož/Paris: European Language Resources Association (ELRA).
- Howitt, Katherine & Scontras, Gregory & Polinsky, Maria. 2025. English restrictive relative clauses are subject to crossover violations. *Linguistic Inquiry*. https://doi.org/10.1162/LING.a.541.
- Kupietz, Marc et al. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*, 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

References

- Liu, Fang & Eugenio, Evercita C. 2018. A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research* 27(4). 1024–1044. https://doi.org/10.1177/0962280216650699.
- Martin, Osvaldo A. & Kumar, Ravin & Lao, Junpeng. 2021. *Bayesian Modeling and Computation in Python*. Boca Raton: CRC Press. https://doi.org/10.1201/9781003019169.
- Masloch, Simon & Poppek, Johanna M. & Kiss, Tibor. 2025. On the (im-)possibility of reflexive binding into the subject of German experiencer-object verbs. In Bîlbîie, Gabriela & Schaden, Gerhard (eds.), *Empirical issues in syntax and semantics. Selected papers from CSSP 2023*, 197–222. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.15450442.

LDSL
Linguistic Data Science Lab

References

- Ospina, Raydonal & Ferrari, Silvia L. P. 2010. Inflated beta distributions. *Statistical Papers* 51(1). 111–126. https://doi.org/10.1007/s00362-008-0125-4.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
- Rychlý, Pavel. 2008. A lexicographer-friendly association score. In Sojka, Petr & Horák, Aleš (eds.), *Proceedings of recent advances in Slavonic natural language processing, RASLAN 2008*, 6–9. Brno: Masaryk University.
- Stan Development Team. 2024. *Stan User's Guide*. Version 2.35. https://mc-stan.org/docs/2_35/stan-users-guide/.
- Vehtari, Aki & Gelman, Andrew & Gabry, Jonah. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

LDSL
Linguistic Data Science Lab

Materials MPK

Test items:

- Verb-final clause of interest embedded in matrix clause
- Inanimate NP subject containing embedded PP containing *sich*
- NPs chosen based on analysis of frequent N-P combinations
- Object as only possible antecedent

Measures N P REFL

In MPK, we took great care to use only N P REFL sequences sounding natural:

- Nouns frequently having a preposition as its right neighbour were extracted from DeReKo (Kupietz et al. 2010) using KorAP (Diewald et al. 2016).
- 327 nouns manually chosen. For each of them collocation scores with 81 prepositions (as direct right neighbours) and possessive pronouns (maximally three words to the left of the noun) computed.
- Noun-preposition combinations then chosen from pairs with a high logDice (Rychlý (2008)); manually checked whether use of an embedded reflexive is overshadowed by a possessive construction (as judged by us).

Results $M_{standard}$

In MPK, we took the model to support the main hypothesis:

- Binding into the subject of German of EO verbs is licit only if it is *not* backward
  ⇒ No need to postulate a peculiar syntactic structure for German EO verbs to account for their (reflexive) binding behaviour.
- No evidence for reconstruction. Binding patterns follow surface order.
  ⇒ No need to assume fixed base order to explain patterns.
  (Caveat: If fixed base-order is SO and Principle A anywhere condition, data compatible. Data also compatible under standard assumption that scrambling reconstructs for reflexive binding)

Results M_{choice}

M_{choice} is a model without filler item data as on slide 15

Fit with PyMC (Martin & Kumar & Lao 2021) (4 chains, 40 000 draws per chain, using default Metropolis-within-Gibbs sampler for *b*, the default No-U-Turn sampler for the other parameters)

Overall results similar to $M_{standard}$

- $\hat{\beta}_{\text{ORDER}}$ = 0.9 [0.53, 1.29]. Bit larger
- All τs ca. 0.2 smaller ⇒ Overall acceptability in normal group a bit higher
- $\hat{\pi}$ = 0.15 [0, 0.31], but skewed distribution with density peak at ca. 0.1
- $\hat{\alpha}_{bad}$ = −1.93 [-4.17, 0.11]
- $\hat{\sigma}_{u_1}$ = 1.43 [1.02, 1.86]

LDSL
Linguistic Data Science Lab

Results M_{choice}: Individual Participants

- Which participants does the model consider to belong to the group of participants who reject everything?
- The *b*[*p*] (part_group[p]) for participant *p* is either 0 or 1 for each draw. Mean value indicates share of draws in which the participant was considered to belong to the group of participants who consider everything bad.
- Surprisingly, there are only 6 participants with values of ca. 50 % upwards. For all of them the 89 % credible interval includes both values.

	mean	sd	hdi_5.5%	hdi_94.5%
part_group[39]	0.874	0.332	0.0	1.0
part_group[1]	0.734	0.442	0.0	1.0
part_group[20]	0.594	0.491	0.0	1.0
part_group[6]	0.528	0.499	0.0	1.0
part_group[15]	0.498	0.500	0.0	1.0
part_group[42]	0.497	0.500	0.0	1.0
part_group[28]	0.271	0.444	0.0	1.0
part_group[24]	0.267	0.442	0.0	1.0
part_group[8]	0.246	0.431	0.0	1.0
part_group[7]	0.228	0.419	0.0	1.0
part_group[47]	0.224	0.417	0.0	1.0
part_group[37]	0.159	0.366	0.0	1.0

Model comparison

$M_{\text{standard+fillers}}$ and M_{diff} have very similar predictive performance as measured by $elpd_{loo}$ (Vehtari & Gelman & Gabry 2017): $elpd_{loo}^{M_{\text{diff}}} - elpd_{loo}^{M_{\text{standard+fillers}}} = 4.07, SE = 6.6$

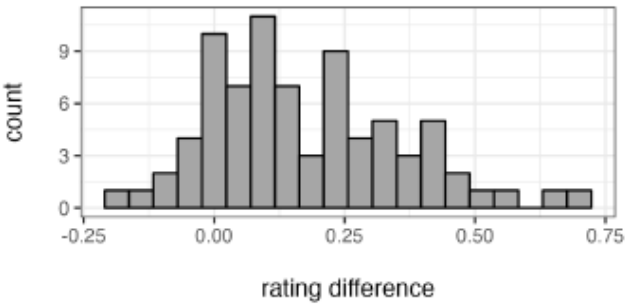
Bayes Factors would be more appropriate because we are not primarily interested in predictive performance, but ultimately in the posterior odds.

To perform a model comparison that takes the fact that it is theoretically unpleasant to assume too much variation between participants into account properly, one would need real, informative priors for the participants' group-level effects.

Hard to come up with, risk of building the intended results into the priors

Participant Variability Checks (Howitt & Scontras & Polinsky 2025)

- Authors checked for differences between participants (not their main interest) by looking for bimodality in a histogram with differences between non-WCO-violation condition and WCO-violation condition per participant.
- Did not see any (ibid., p. 8), but hard to tell in my opinion.



Reproduced from (Howitt & Scontras & Polinsky 2025: figure 3)

Hypothesis Testing

- Is there a need to assume that some participants have a different linear predictor than the others?
- If we have informative priors for π, we can test for this hypothesis because it would correspond to the point 1 (or 0) hypothesis for π.
- We can simply compute the BF_{10} against this point Null then. (In the case of the linear model including participant differences for Howitt & Scontras & Polinsky's (2025) data, we get a BF of ca. 4 against the Null that $\pi = 1$, but one would really need a serious prior and should perform sensitivity tests to do this properly.)