Investigating early language acquisition with R

A reproducible workflow with formR, Shiny and Stan

Universitat Pompeu Fabra Barcelona

Gonzalo Garcia-Castro
Daniela S. Avila-Varela
Nuria Sebastian-Galles

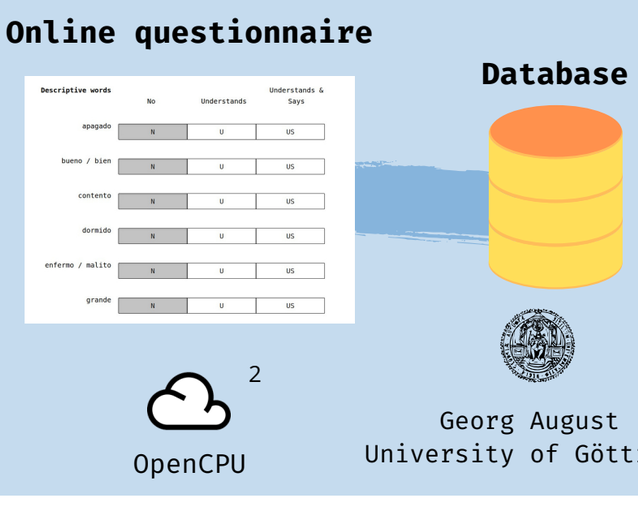
@gongcastro
/gongcastro
gongarciacastro

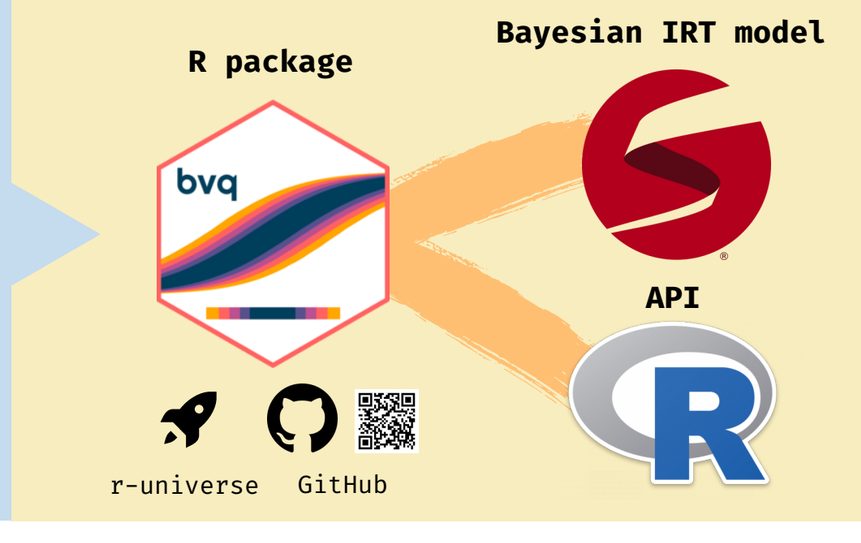
Online questionnaire to collect parental estimates **word acquisition** in 10-32 month-old bilinguals in Barcelona.

We used **open-source tools** to collect, model, visualize, and share the data, maximizing **computational reproducibility**.

Easily **extended** and **implemented** for other survey-based workflows.

Data collection: **formR** [1] | Data processing/modelling: **bvq** and **Stan** [3] | Outreach: **ShinyApp**





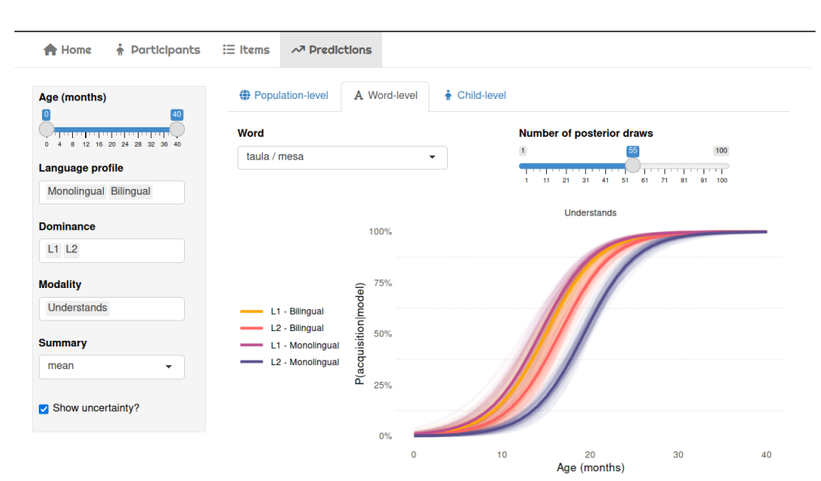
Online questionnaire

- Demographic information
- Language profile
- **1,600 Catalan/Spanish words**
- Complex item randomization design
- *N* = 586

Bayesian model

- Implemented in **Stan/brms**
  `4 MCMC chains, 6,000 samples`
- Incorporates **prior** knowledge
- **Multilevel structure** that estimated *participant*- and *item*-level parameters (IRT) [4]

Preprint



R package/API

Questionnaire responses

`bvq_responses(...)`

Participant and item-level information

`bvq_logs(...)`

Get vocabulary norms

`bvq_vocabulary(...), bvq_norms(...)`

Launch app

`bvq_app(...)`

Response (k) **to word** i **by participant** j

$$\text{Response}_{ij} \sim \text{Cumulative logit}(\theta_{k_{ij}})$$

where $k \in \{\text{No} \rightarrow \text{Understands}, \text{Understands} \rightarrow \text{Understands and Says}\}$

Distribution parameters

$$\theta_{k_{ij}} = (\beta_{0_k} + u_{0_{i_k}} + w_{0_{j_k}}) + (\beta_1 + u_{1_i} + w_{1_j}) \cdot \text{Age}_i +$$
$$(\beta_2 + u_{2_i} + w_{2_j}) \cdot \text{Length}_{ij} + (\beta_3 + u_{3_i} + w_{3_j}) \cdot \text{Exposure}_{ij} +$$
$$(\beta_4 + u_{4_i}) \cdot \text{Cognateness}_{ij} + (\beta_5 + u_{5_i} + w_{3_j}) \cdot (\text{Age}_i \times \text{Exposure}_{ij}) +$$
$$(\beta_6 + u_{6_i}) \cdot (\text{Age}_i \times \text{Cognateness}_{ij}) +$$
$$(\beta_7 + u_{7_i}) \cdot (\text{Exposure}_{ij} \times \text{Cognateness}_{ij})$$
$$(\beta_8 + u_{8_i}) \cdot (\text{Age}_i \times \text{Exposure}_{ij} \times \text{Cognateness}_{ij})$$

where:

β_{1-8}: fixed effects

u_{1-8_i}: participant-level random effects

w_{1-3_j}: TE-level random effects

Prior

$$\beta_{0_k} \sim \mathcal{N}(-0.25, 0.5);\ \beta_{1-5} \sim \mathcal{N}(0, 1)$$
$$\sigma_{u_{0-8}, w_{0-3}} \sim \mathcal{N}_+(1, 0.25);\ \rho_{u_{0-8}, w_{0-3}} \sim \text{LKJcorr}(2)$$

where:

$\sigma_{u_{0-8}, w_{0-3}}$: are the standard deviations of u and w

$\rho_{u_{0-8}, w_{0-3}}$ are the correlations between random effects in u and w

ShinyApp

- Data **visualization**
- Model estimates: **posterior distribution**, **posterior predictions**
- Model diagnostics (*R-hat*, *N-eff*)

Future steps

- Migrate OpenCPU/formR instance to own server
- Continous integration using GitHub Actions

References

1. Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. Behavior Research Methods, 52, 376-387.
2. Ooms, J. (2014). The OpenCPU system: Towards a universal interface for scientific computing through separation of concerns. arXiv preprint arXiv:1406.4806.
3. Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76.
4. Bürkner, P. C. (2019). Bayesian item response modeling in R with brms and Stan. arXiv preprint arXiv:1905.09501.

R QUE R II CONGRÉS DE R XIII JORNADES D'USUARIS DE R
BARCELONA 15-17 NOVEMBRE 2023 / CAMPUS POBLENOU UPF-BSC

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN