BEETLE

A Bilingual Model Suite for Modelling Second-Language Processing

Suchir Salhan, Catherine Arnett, James A. Michaelov, Paula Buttery

University of Cambridge · EleutherAI · Massachusetts Institute of Technology

Motivation

Multilinguality & Interpretability. How does cross-lingual transfer occur? How much does it vary by language pair? To what extent does training curriculum matter?

Human Language Processing. Most people are multilingual, but computational models of language are overwhelmingly monolingual.

Method

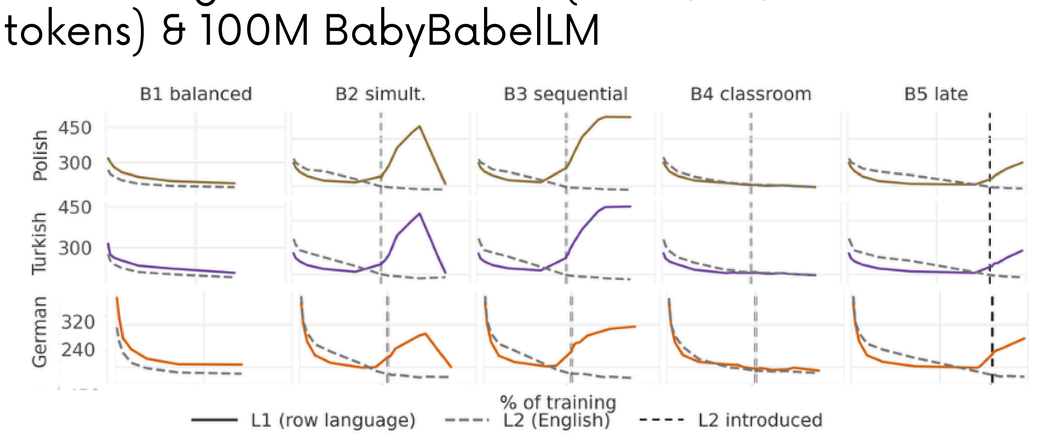
- **Data.** First-pass gaze duration from the Multilingual Eye Tracking Corpus (MECO): reading time data from second-language English speakers
- **Models.** Beetle models and 5 massively multilingual baseline models (Apertus, Gemma, Llama, Qwen, XGLM)
- **Modeling.** We calculate word surprisal from each Beetle model and use it to predict human times, controlling for word length, frequency, and position in the text using linear mixed-effects regression models
- **Comparison.** We compare the fit of the regressions with each language model's surprisal as a predictor. We consider a higher log-likelihood to indicate better modeling of the data

Training Framework

- A modular language model training & analysis framework
- Systematically manipulate training data curriculum, architecture, tokenizer
- Rich checkpointing for studying learning dynamics and catastrophic forgetting

Model Suite

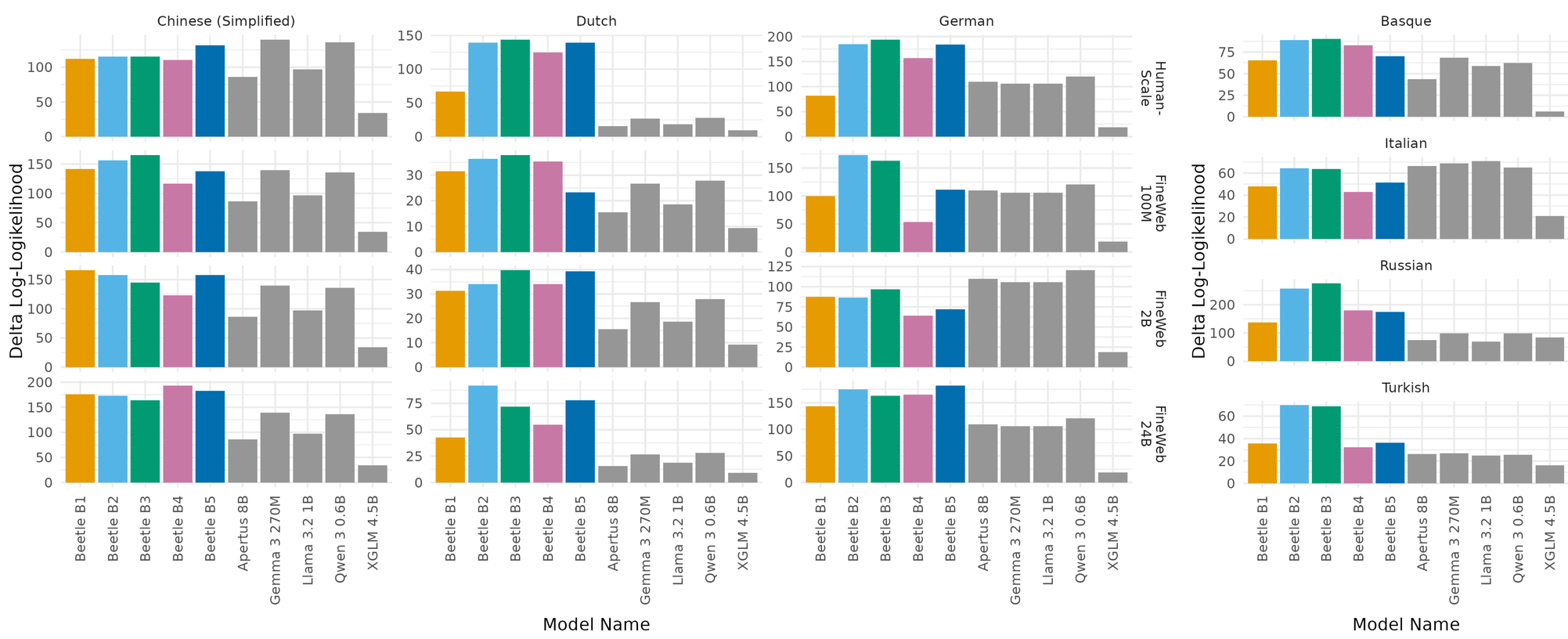
- 285 bilingual language models (21 unique L1s, L2 English)
- 5 acquisition-inspired data curricula:
  - Balanced (B1): 50/50 mix of L1 and L2
  - Simultaneous (B2): L1 only for first half, then 50/50 mix
  - Sequential (B3): L1 only for first half, L2 only for second half
  - Classroom (B4): episodic exposure, 80/20 mix
  - Late (B5): L2 onsets 80% through training
- 3 data regimes: FineWeb-2 (100M, 2B, 24B tokens) & 100M BabyBabelLM



Results

(A) Comparison of Beetle training data size and curriculum

(B) Extended analysis on FineWeb 100M



Takeaways

- The Beetle models generally show a closer fit to L2 reading times than the multilingual baselines
- Exposure curriculum matters: substantial variation across training schedules & joint pretraining leads to a worse fit than scheduled L2 introduction.
- No universally best curriculum for modeling L2 language processing - may require considering individual language experience of participants

Get in touch: sas245@cam.ac.uk

References

Jumelet et al. (2026). BabyBabelLm: A multilingual benchmark of developmentally plausible training data. EACL. **Penedo et al. (2025)**. FineWeb2: One pipeline to scale them all: Adapting pre-training data processing to every language. COLM. **Siegelman et al. (2022)**. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). Behavior Research Methods.

The Beetle model suite enables research on multilingual language acquisition and processing:

https://beetlelm.github.io