

Large-scale eye-tracking while reading benchmark shows surprisal captures early fixations, but not regressions

William Timkey¹, Suhas Arehalli², Kuan-Jung Huang³, Grusha Prasad⁴, Tal Linzen¹, Brian Dillon³

¹New York University, ²Macalester College, ³University of Massachusetts Amherst, ⁴Colgate University

Introduction

- **Garden path sentences** are temporarily ambiguous between multiple syntactic structures
- Readers incur a processing cost when the sentence is disambiguated in an unexpected way:

The experienced soldiers warned about the dangers **conducted** the midnight raid.

Surprisal-based accounts:

- Readers maintain many parallel interpretations
- Garden path effects are driven by the disambiguating region's low predictability, measured by surprisal (Hale 2001, Levy 2008)
- Surprisal-difficulty relationship is linear, and constant (not specific to garden paths)

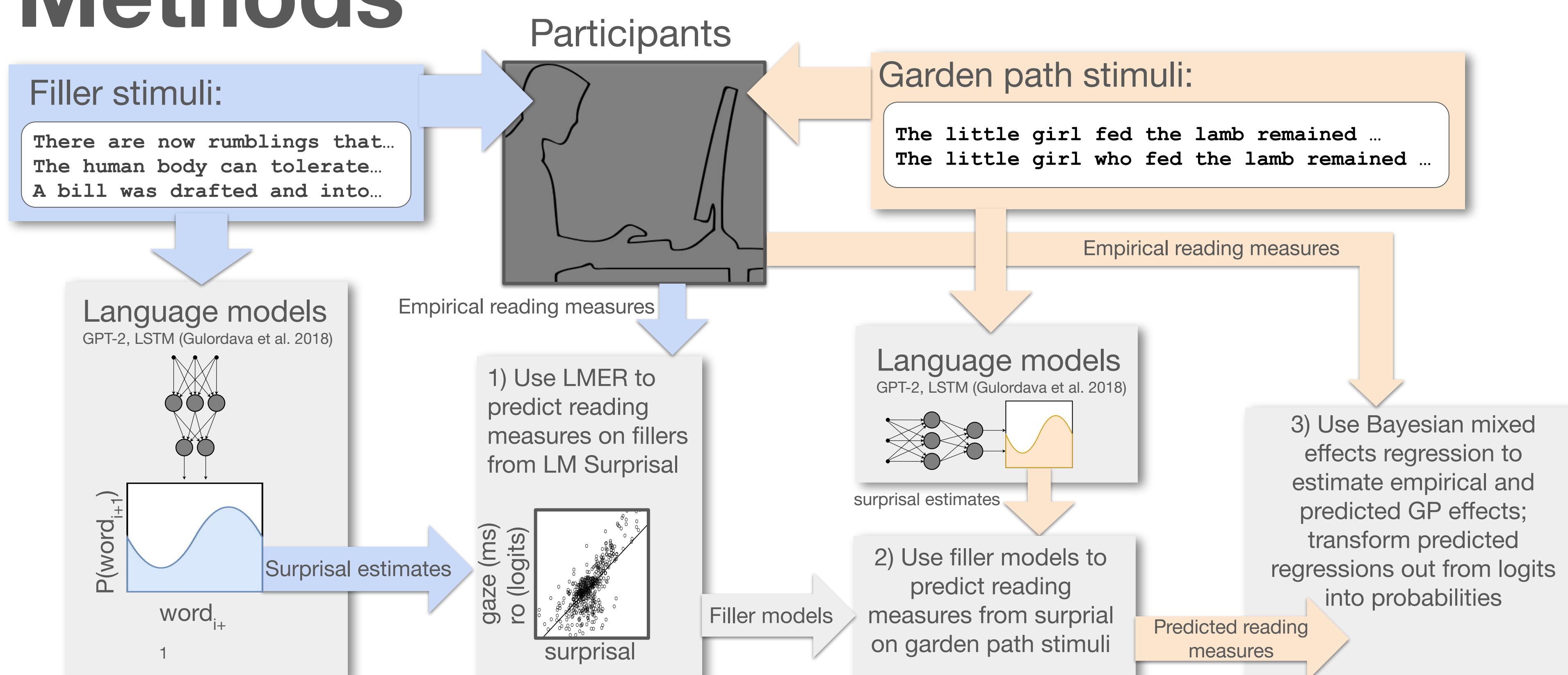
Two-stage accounts:

- Readers maintain a single interpretation at a time
- Parser crashes when input cannot be integrated into existing structure, triggering a costly structural reanalysis stage (Frazier & Rayner 1982)

- Surprisal estimates from neural language models drastically underpredict the magnitude of garden path effects in self-paced reading (Huang et al. 2024)
- But, SPR conflates distinct sources of difficulty (word recognition, syntactic integration+reanalysis) into a single RT measure
- Surprisal might explain some sources of difficulty, but not others (e.g. Hale et al. 2018)

- In a large-scale (n=384) eye-tracking while reading study, we investigate whether surprisal can explain two distinct measures of processing difficulty in garden path sentences:
 - **First pass/gaze duration:** Total fixation time on a word before exiting to the left or right (associated with word recognition) (Reichle et al. 2009)
 - **First pass regressions out:** Probability of regressing to a previous word after the first pass (associated with integration difficulty) (Reichle et al. 2009)

Methods



Materials

1) Main verb / Reduced relative (MVRR)

- 1a) The little girl fed the lamb **remained relatively** calm despite having asked for beef.
1b) The little girl who was fed the lamb **remained relatively** calm despite having asked for beef.

2) Direct object / Sentential complement (NP/S)

- 2a) The little girl found the lamb **remained relatively** calm despite the absence of its mother.
2b) The little girl found that the lamb **remained relatively** calm despite the absence of its mother.

3) Noun phrase / Zero complement (NP/Z)

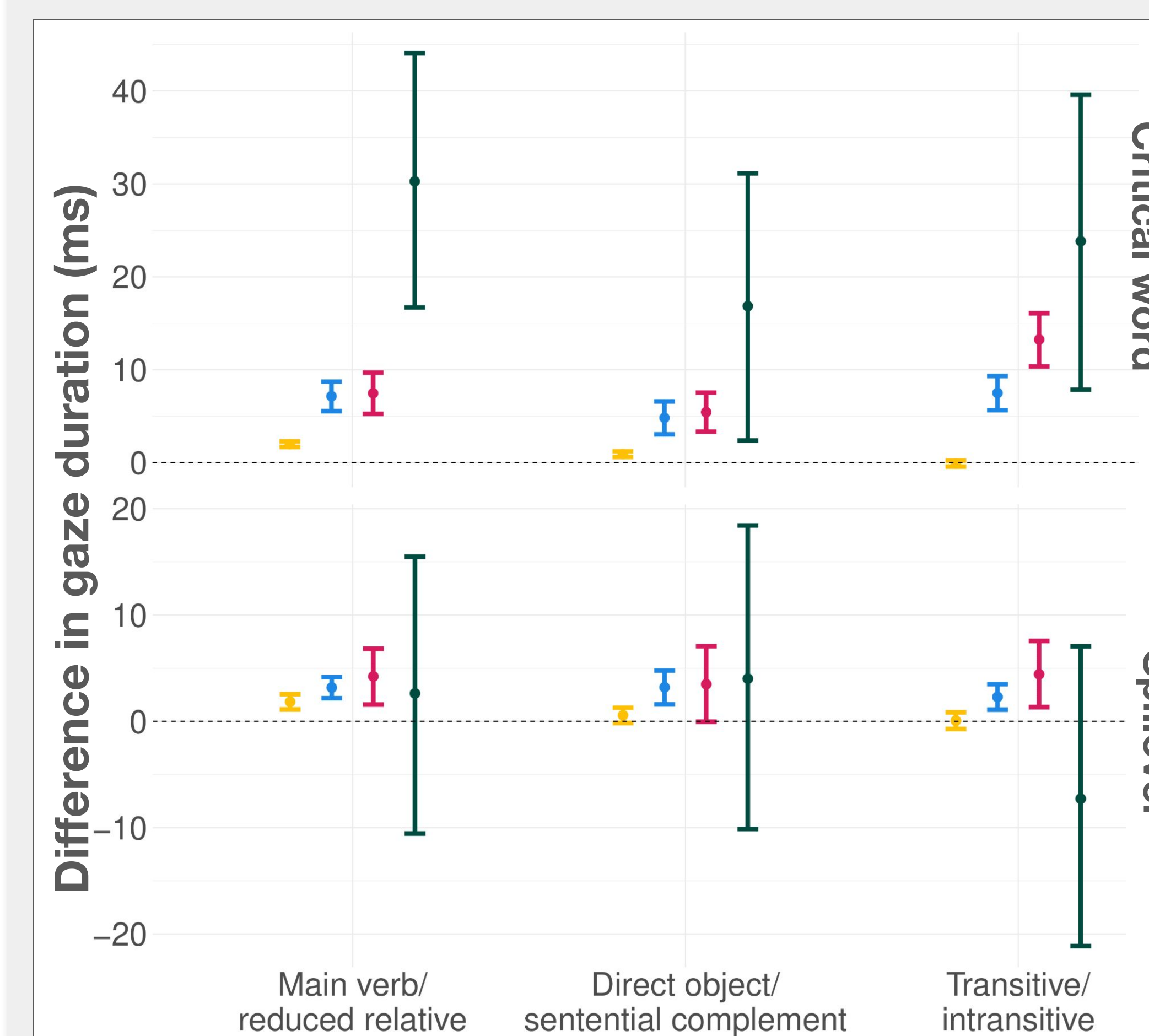
- 3a) When the little girl attacked the lamb **remained relatively** calm despite the sudden assault.
3b) When the little girl attacked, the lamb **remained relatively** calm despite the sudden assault.

Ambiguous
Unambiguous
Critical word
Spillover

Results

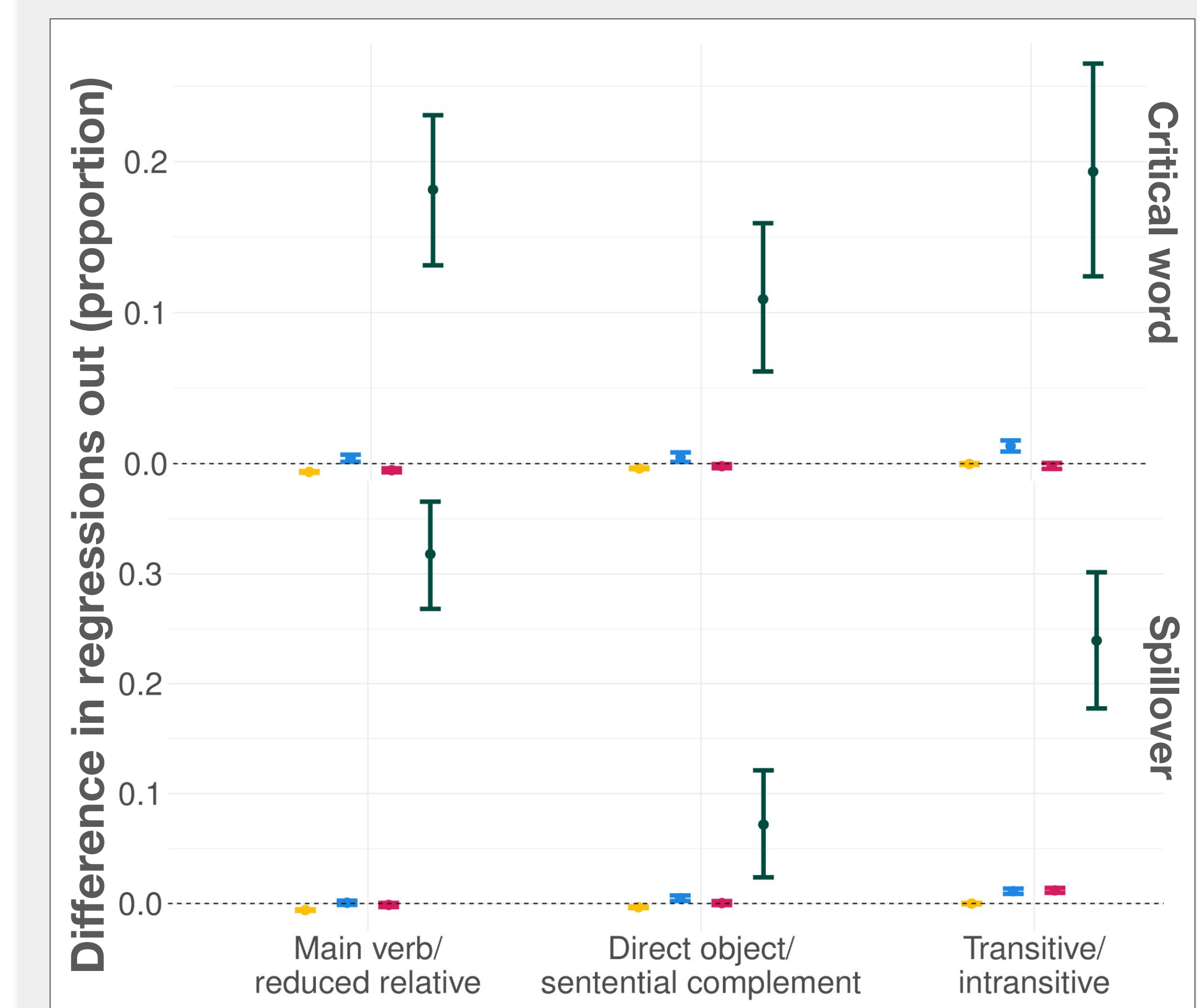
• No surprisal baseline • Wiki-LSTM • GPT-2 • Empirical

Gaze Duration



- Humans show garden path effects in critical region, but not spillover region in all three constructions
- **Surprisal predicts effects with magnitudes consistent with human data in NP/S and NP/Z constructions**
- Surprisal underpredicts effect in MV/RR by a factor of ~4.5x, compared to ~28x in SPR data

Regressions Out



- Humans show garden path effects in critical region and spillover region in all three constructions
- **Surprisal does not predict the magnitude of garden path effects in regressions out**
- Surprisal also fails to predict the direction of the effect in many cases

Discussion

- Surprisal can account for the early stages of garden pathing (i.e. difficulty recognizing the critical word), but fails to account for whether a readers initiate regressive eye movements
- Reanalysis “trigger” may be unrelated to predictability
- Suggests a two-stage account that incorporates surprisal in the first stage, and an explicit reanalysis mechanism in the second stage