Evaluating the Effect of Model Inductive Bias and Training Data in Predicting Human Reading Times

Ethan Wilcox♦, Jon Gauthier♠, Peng Qian♠, Jennifer Hu♠, and Roger Levy♠

♦ Linguistics, Harvard University, ♠ Brain and Cognitive Sciences, MIT

Language Models of Sentence Processing

Language Modeling: $P(x_i \mid x_1 \ldots x_{i-1})$

Background: What leads to good predictive power of human reading times?

Perplexity

- How good is the language model at predicting the upcoming token? (lower = better)
- Models with **lower perplexity** have a better predictive power (Fossum & Levy, 2012)
- **Linear relationship** between perplexity and Predictive Power (Goodkind and Bicknell, 2018)

Inductive Bias

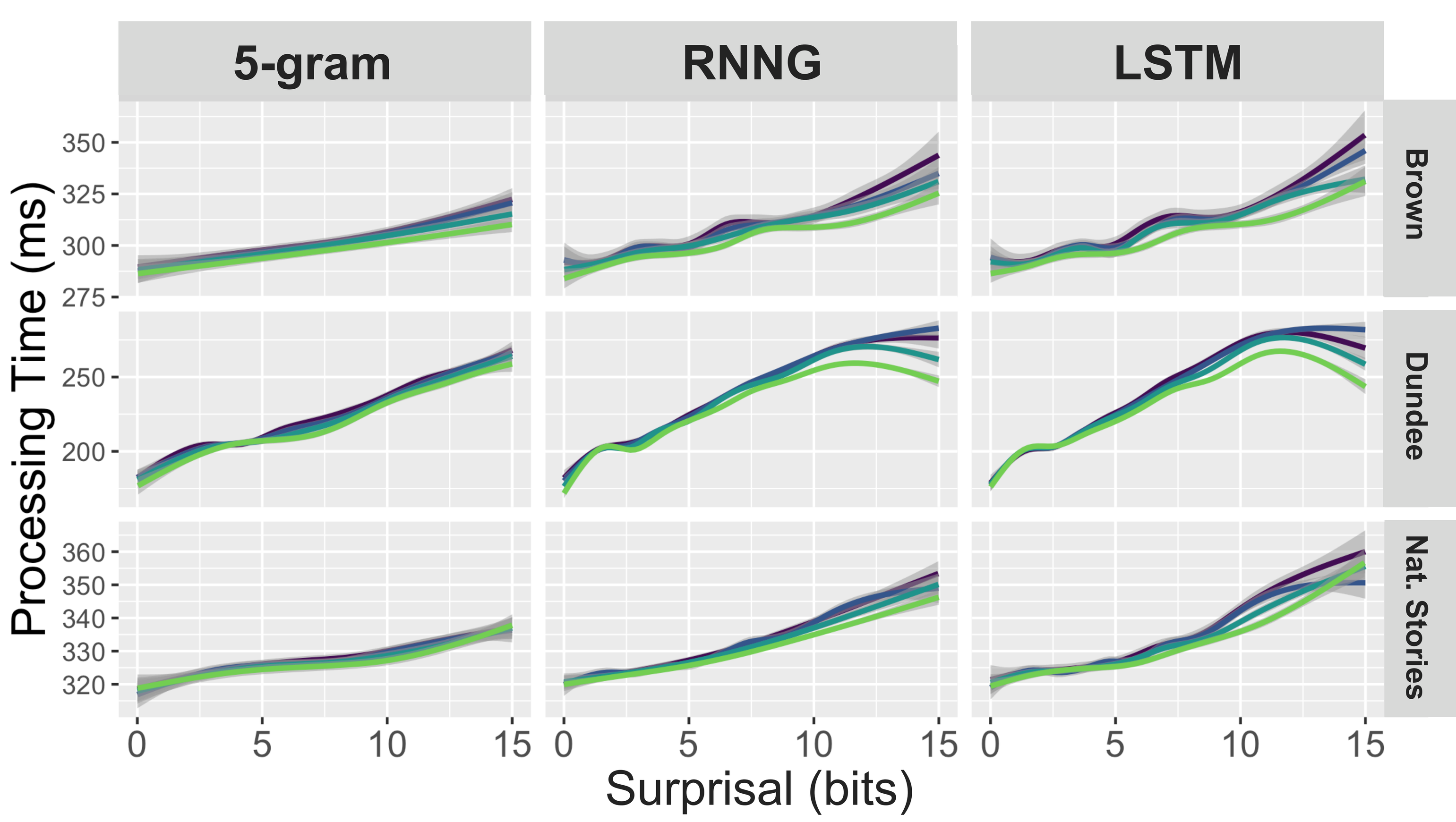
- Is the model trained with explicit structural supervision?
- Models **without hierarchical bias** have better predictive power (Frank & Bod, 2011)

Previous Work: Focuses on *n*-gram models. Do results hold for state-of-the-art neural network models?

Result 1: Surprisal as a linking function

Reading time is *linearly correlated* with **surprisal** (Smith & Levy, 2013)

$$\text{Surprisal} = -log(\,P(x_i \mid x_1 \ldots x_{i-1}))$$



GAM regression shows linear relationship between surprisal & reading time for all models tested.

Methods

12 Architecture X Training Data Pairs

Tag	Model Name	Architecture Type	Inductive Bias
●	N-Gram	Statistical	Local Window Only
■	LSTM	Neural Network	Linear locality+unstructured memory store
▲	RNNG	Neural Network	Supervised with Penn-Treebank style parses

Training Size	Millions of Tokens
Extra Small	1
Small	5
Medium	14
Large	42

Test Corpus	Data Type	Text Type
Dundee	Eye-Tracking	News Text
Natural Stories	Self-Paced Reading	Stories with rare syntactic structures
Brown	Self-Paced Reading	Stories

Measuring Model Predictive Power

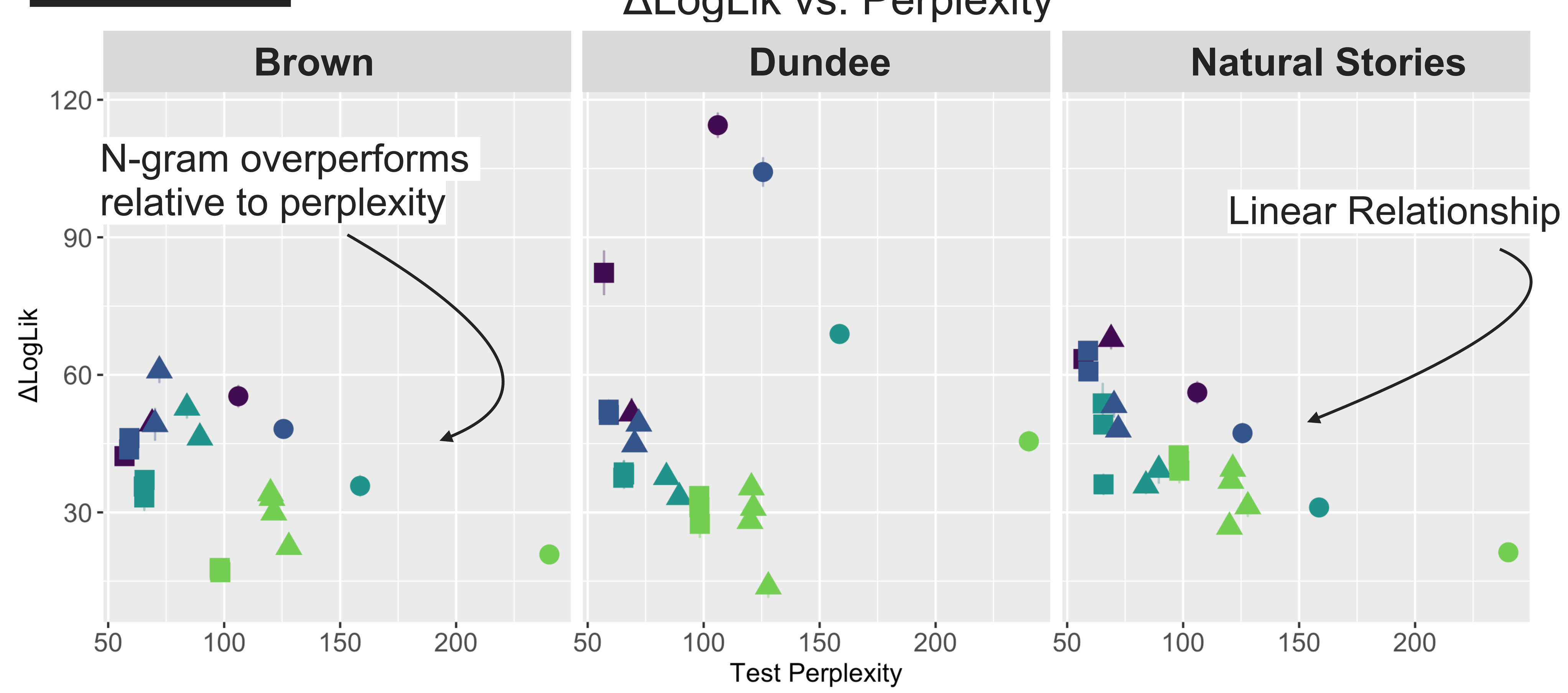
ΔLogLiklihood: difference between baseline and LM-Derived linear regression models following Frank & Bod (2011); Goodkind & Bicknell (2018)

```
lm(read time ~  word length + word frequency)
lm(read time ~ surprisal + word length + word frequency)
```

Current + Previous word (eye-tracking) / Current + 2 Previous Words (SPR)

Result 2

ΔLogLik vs. Perplexity



- Model perplexity is strongly correlated with predictive power
- n-gram model *over performs* based on its perplexity

Predictive Power vs. Syntactic Generalizations

Syntactic Generalization Score

- Performance on targeted syntactic evaluation tests (Marvin & Linzen, 2018)
- Average accuracy across ~700 hand-crafted test items over 6 syntactic categories

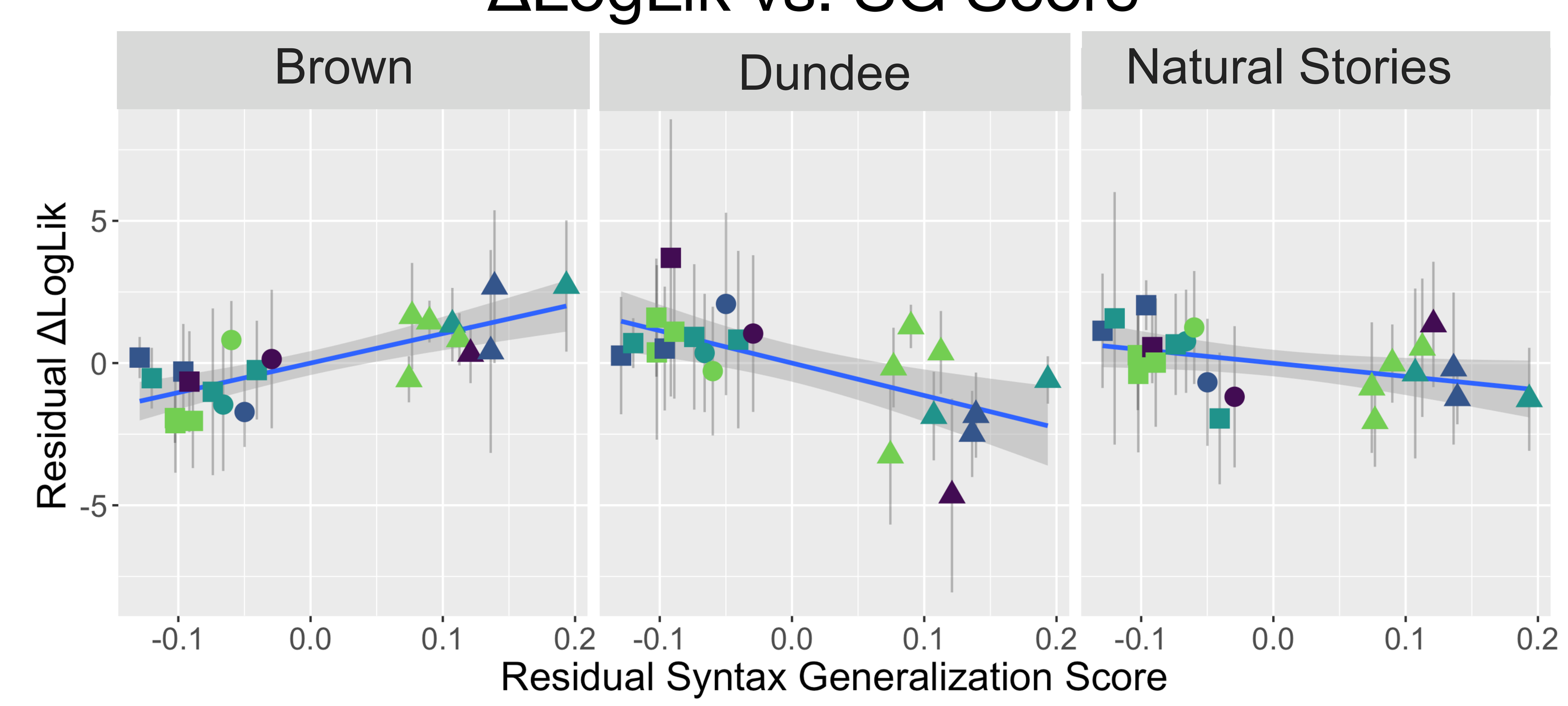
Example: Negative Polarity Items. Is P(`ever` | `context`) less in (a) than in (b) ?

`(a) ✗The senator who we liked ever supported…`

`(b) ✓No senator who we liked ever supported…`

Result 3

ΔLogLik vs. SG Score



- Supports hypothesis that eye movements are more sensitive to n-gram probabilities (McDonald & Shillcock 2003)

Discussion

- Linear relationship between reading time and surprisal in new neural network models
- Model **perplexity** is correlated with predictive power
- No strong relationship between **Inductive Bias** / Syntactic Generalization and predictive power.

References: Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. • Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. Psychological science, 22(6), 829-834. • Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. • Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. • Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319.