Syntactic Node Count as Index of Predictability

Kohei Kajikawa[1] and Shinnosuke Isono[2]

[1]Georgetown University

[2]National Institute for Japanese Language and Linguistics (NINJAL)

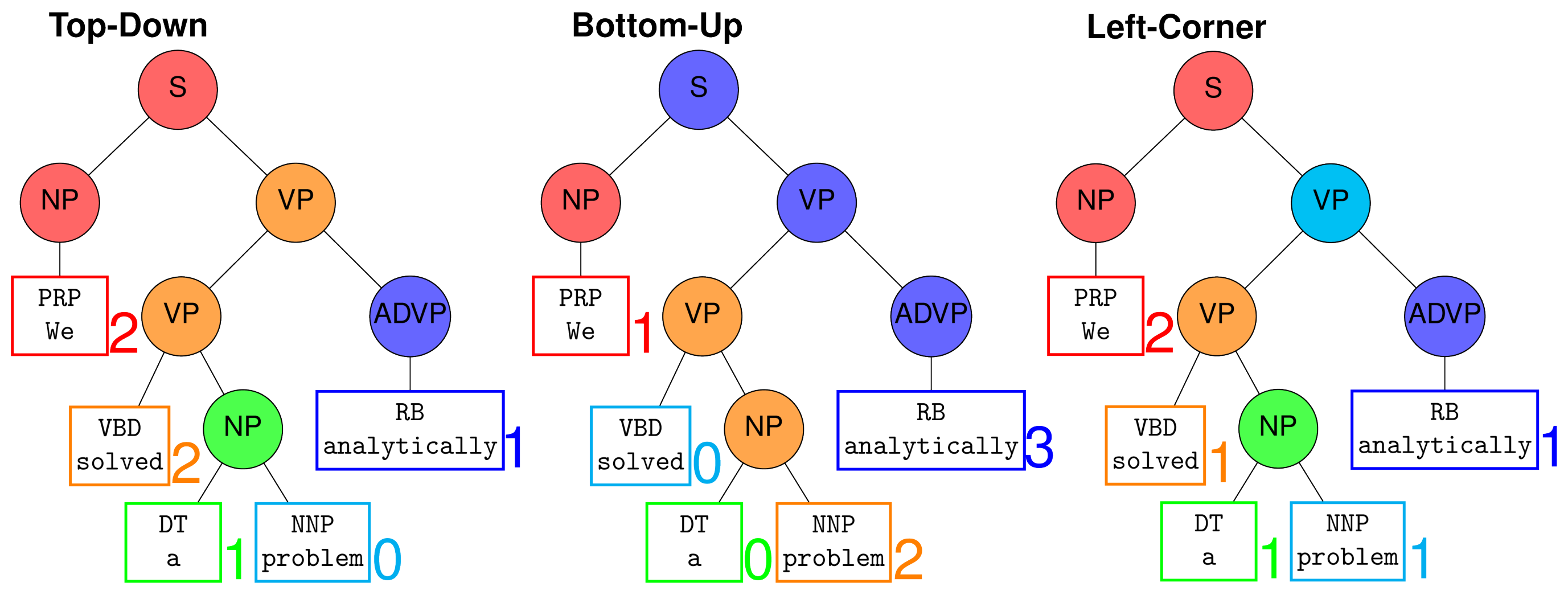
kk1571@georgetown.edu

- Neurolinguistic research has adopted syntactic **Node Count (NC)** as a metric of **complexity**.
- We find ***facilitatory*** effects of NC in early reading times, suggesting that NC captures **context richness**.
  - These effects are independent from GPT-2 surprisal, probably reflecting human-like prediction.
- ***Inhibitory*** effects of NC, predicted by the complexity hypothesis, are found only in a later region.
- These results call for a careful interpretation of NC in neurolinguistics.

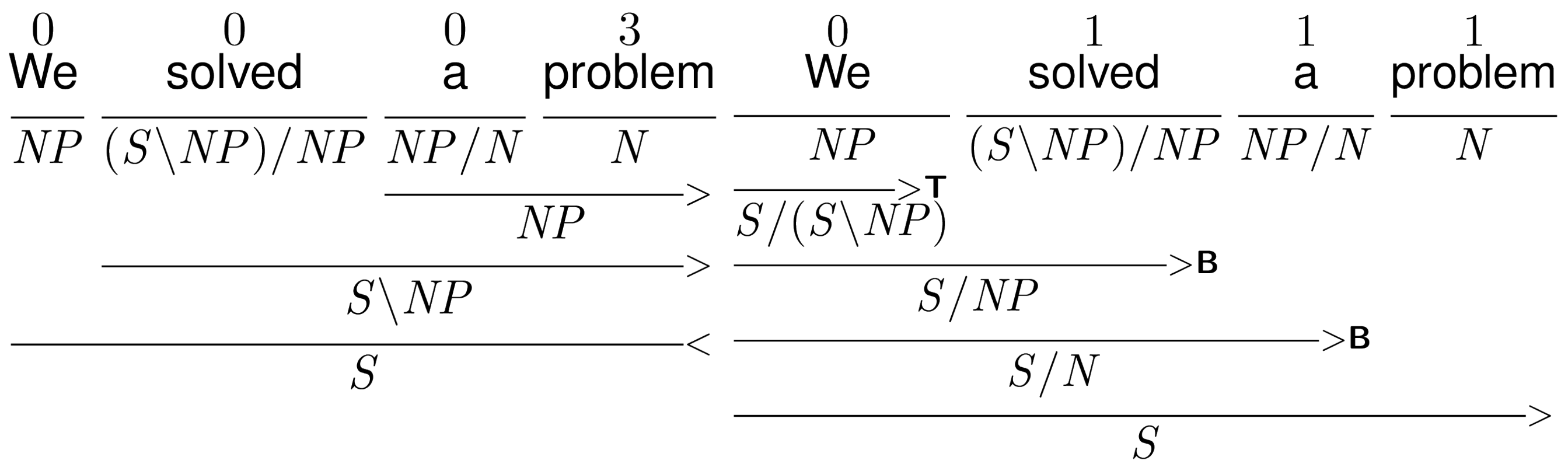
Background

NC is the number of parsing steps at each word → used as a complexity metric in neurolinguistics [1,2]

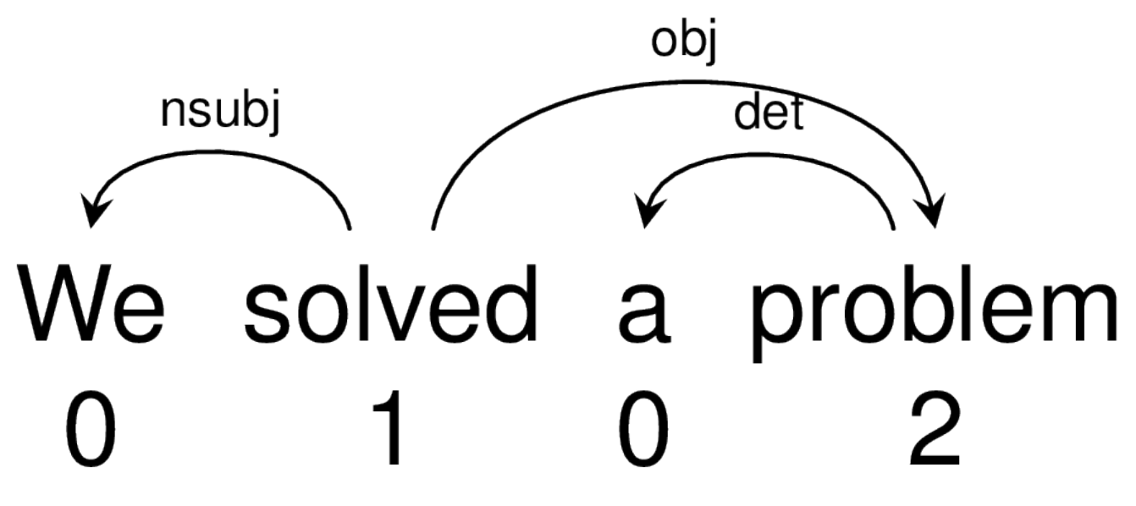
Phrase Structure Grammar (PSG)



Combinatory Categorial Grammar (CCG)

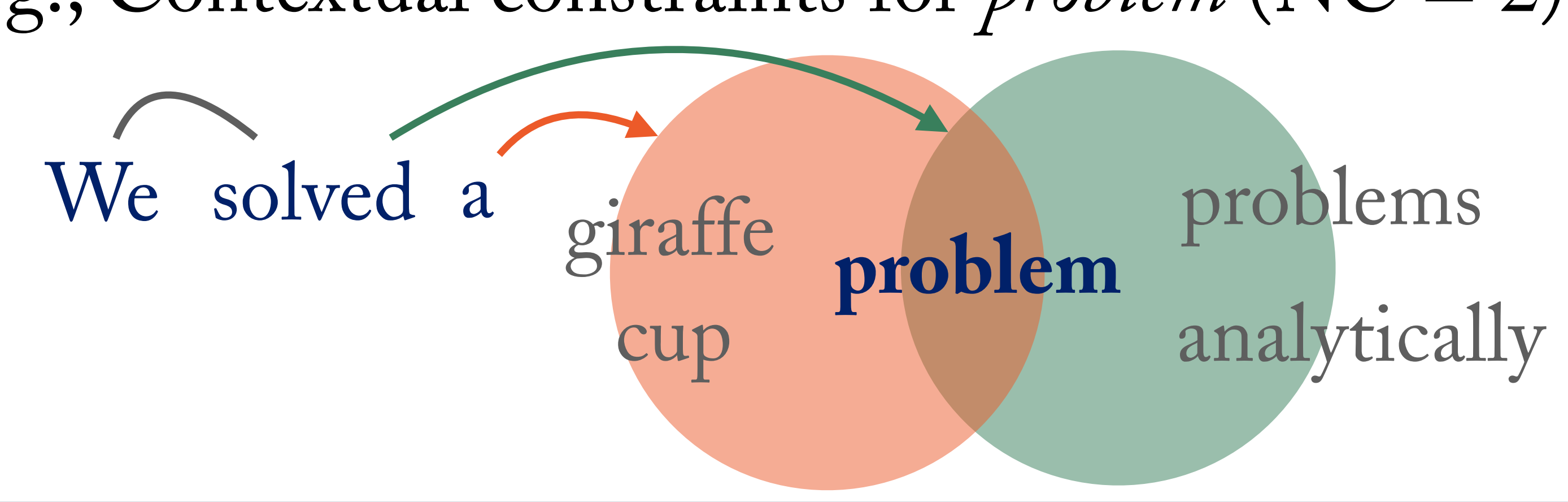


Dependency Grammar



- Some studies report negative (facilitatory) effect of NC on reading times [3,4]
- Possible reason for this: NC is correlated with the amount of **lexical/syntactic/semantic** constraints.
  →Current study: Examine NC's effect on RTs in detail

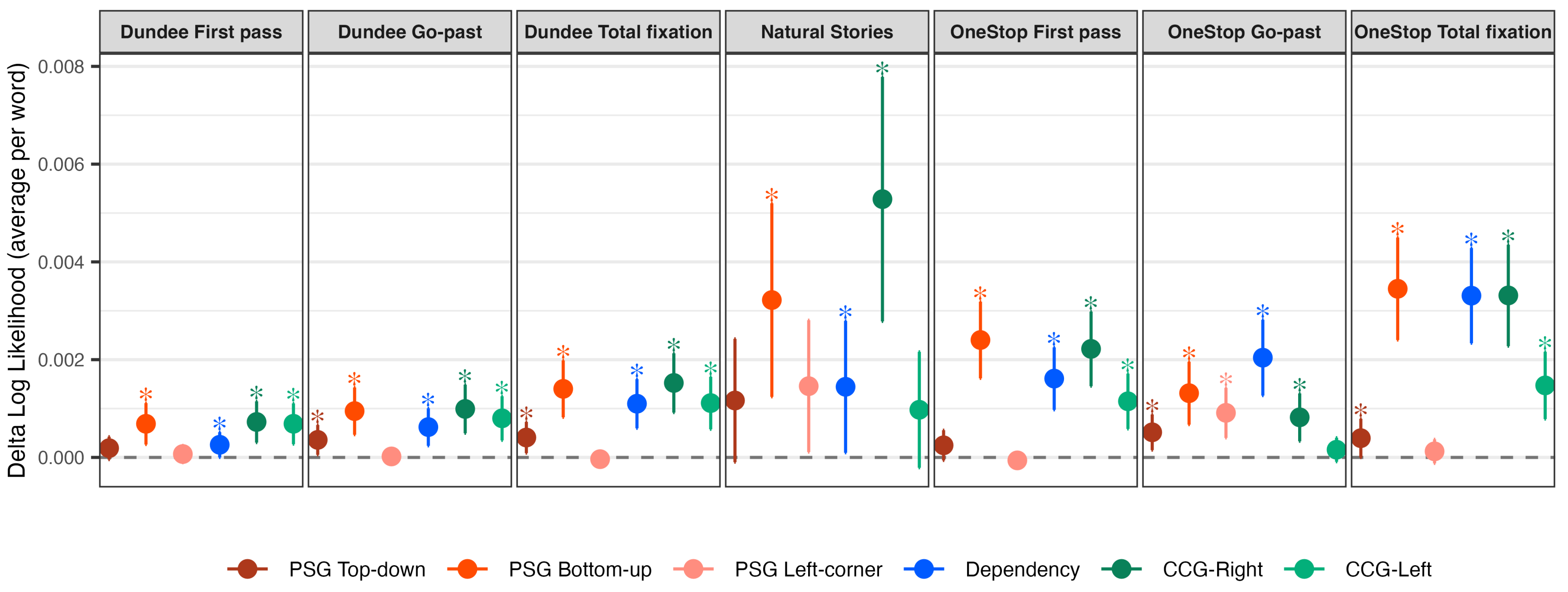
e.g., Contextual constraints for *problem* (NC = 2)



Analysis

Predictive power of Node Count

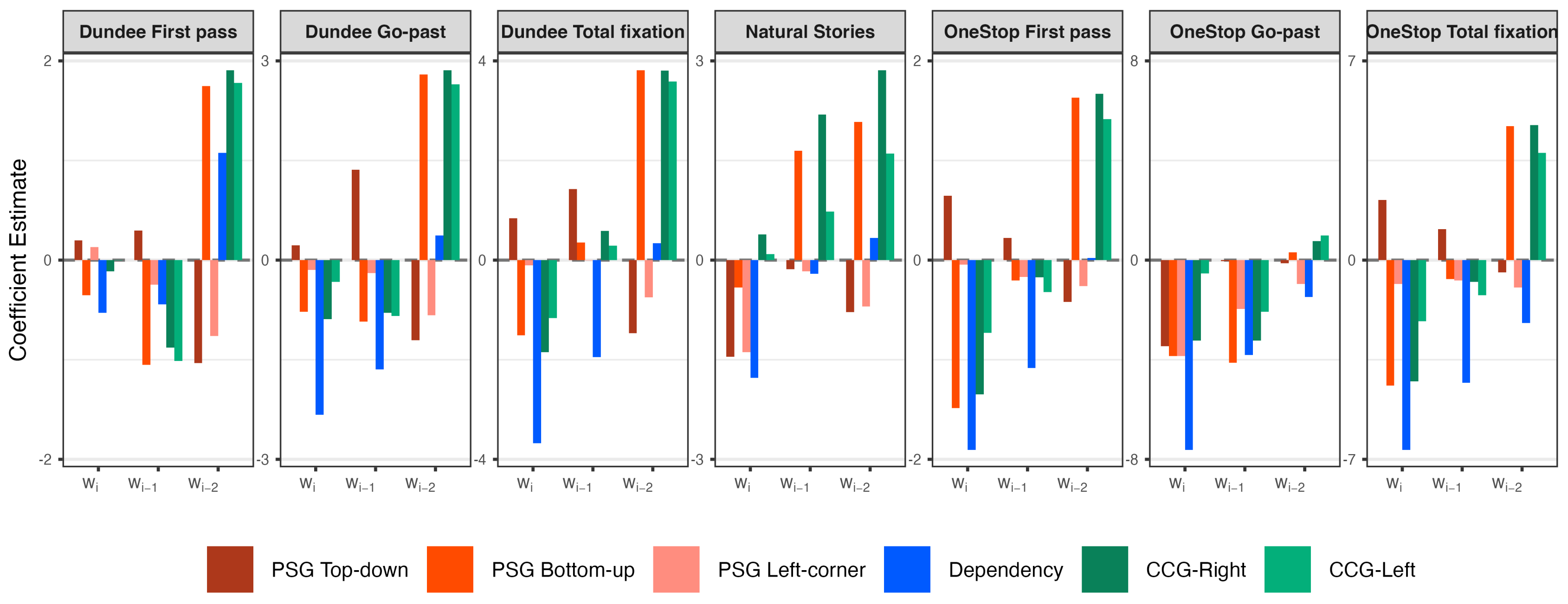
- Predictive power of NC evaluated using 10-fold CV:
  - ΔLoglik = Loglik(Baseline+NC$_{i:i-2}$) − Loglik(Baseline)
  - Baseline: RT$_i$ ~ position$_i$ + len$_i$*freq$_i$ + len$_{i-1}$*freq$_{i-1}$ + len$_{i-2}$*freq$_{i-2}$



- The predictive power of NC is clearest for variants that are more directly tied to the amount of contextual information

Ambivalent nature of Node Count

- Regression coefficients β for NC:



- NC shows **negative** effects in early regions, but **positive** effects in a later region, suggesting that NC reflects both ***predictability effects*** and ***the cost of late integration***

Effects of multiple formalisms

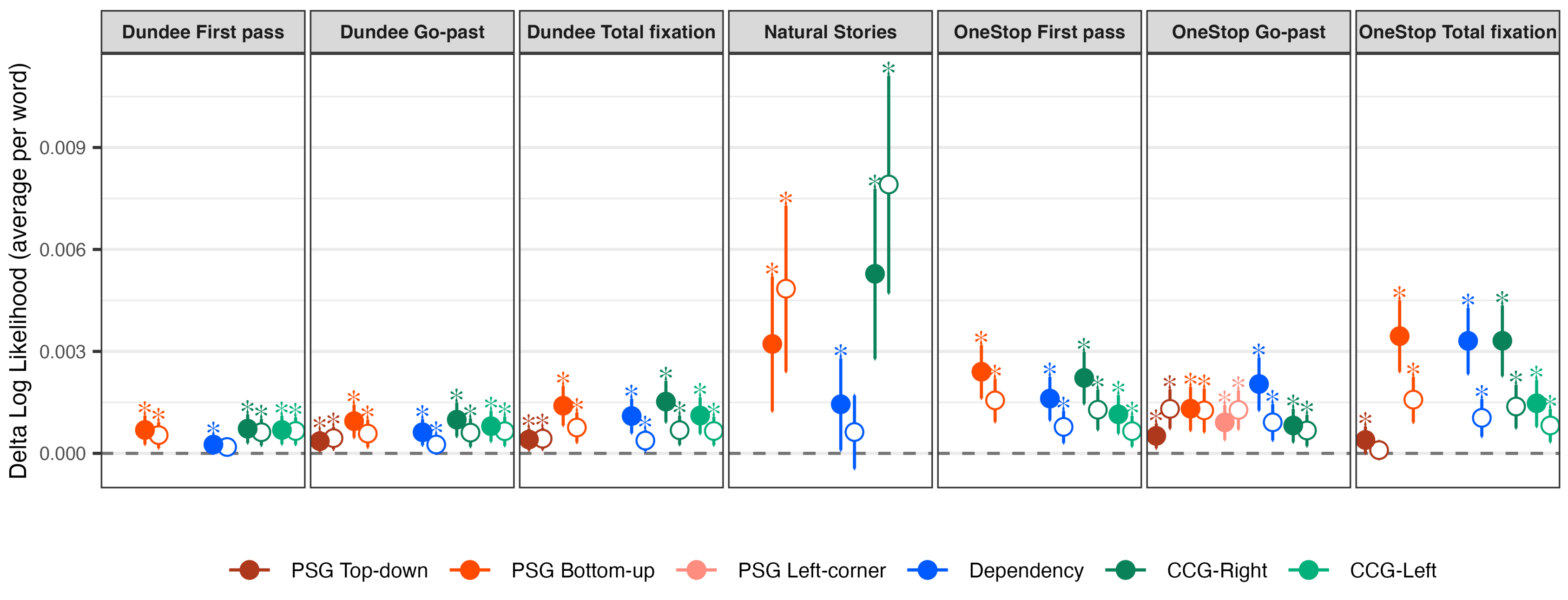
- Best models were selected by AIC; tested on held-out data

	PSG Top-down			PSG Bottom-up			PSG Left-corner			Dependency			CCG-Right			CCG-Left		
	w_i	w_{i-1}	w_{i-2}	w_i	w_{i-1}	w_{i-2}	w_i	w_{i-1}	w_{i-2}	w_i	w_{i-1}	w_{i-2}	w_i	w_{i-1}	w_{i-2}	w_i	w_{i-1}	w_{i-2}
Dundee First pass																	▽	△
Dundee Go-past					▽	△				▽								
Dundee Total fixation										▽	▽				△			
Natural Stories				▽						▽	▽		△	△	△			
OneStop First pass				▽		△				▽	▽							
OneStop Go-past	▽	▽		▽	▽					▽	▽							
OneStop Total fixation				▽	▽					▽	▽	▽			△	△		▽

△: Positive ▽: Negative

- Independent effects suggest multiple structural processing

Independence from GPT-2 surprisal



●: **Before** controlling for GPT-2 surprisal ○: **After** controlling for GPT-2 surprisal

- 28/31 patterns remains significant after controlling for GPT-2 surprisal—NC is not subsumed by co-occurrence statistics
- NC captures ***structure-mediated*** predictability/cost

References: [1] Brennan et al. 2012 *Brain and Language*. [2] Brennan et al. 2016 *Brain and Language*. [3] Asahara et al. 2019 *Gengo Kenkyu*. [4] Isono 2024 *Cognition*.