Particle Filtering with Neural Language Models: Modelling the Effects of Memory on Incremental Sentence Processing

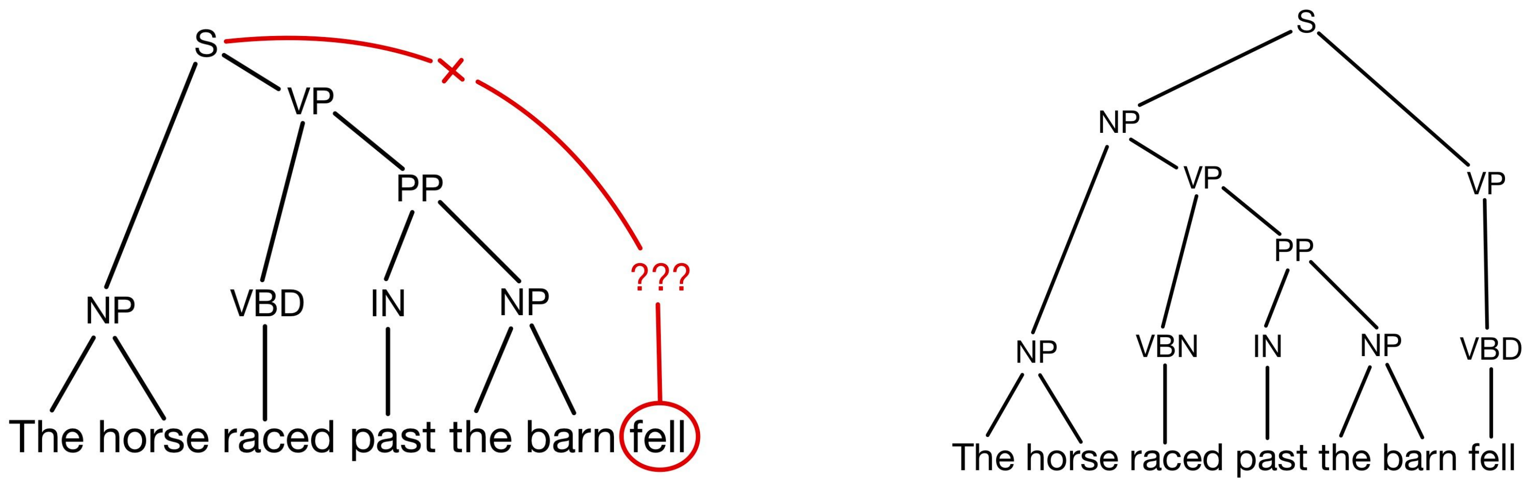
Sarah Payne[1], Peng Qian[2], Ethan Wilcox[2,3], and Roger Levy[2]

paynesa@sas.upenn.edu, pqian@mit.edu, wilcoxeg@g.harvard.edu, rplevy@mit.edu

1. Departments of Linguistics and Computer and Information Science, University of Pennsylvania
2. Department of Brain and Cognitive Science and Center for Brains, Minds, and Machines, MIT
3. Department of Linguistics, Harvard University

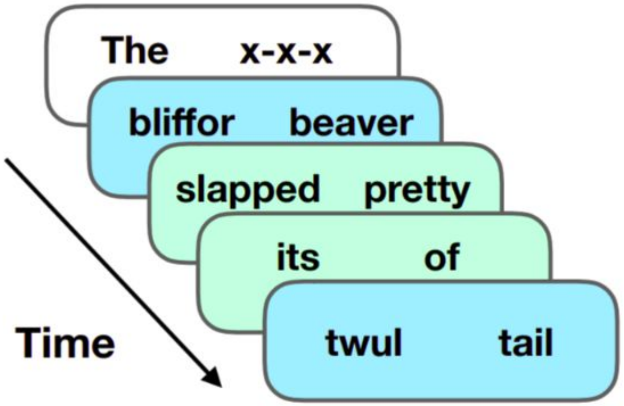
Background

- Sentence processing in humans is **online, incremental, and constrained by memory**
- **Language is ambiguous:** in "garden path" sentences, a locally likely structural hypothesis becomes globally implausible in the presence of disambiguating evidence



Reading Times and Garden Path Effects

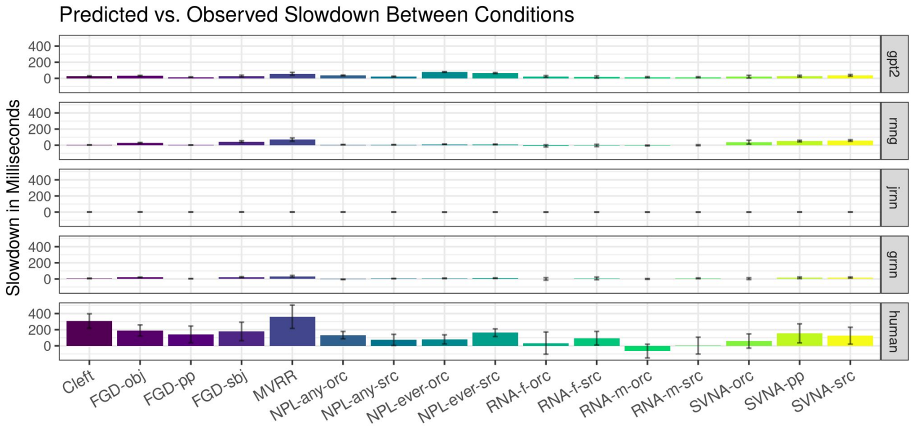
- Insight into sentence processing and garden path effects in humans can be gained via **eye tracking** and **maze tasks**
  - Longer fixation times = greater trouble incorporating word into hypothesized structure



Interpolated Maze (I-Maze)

Courtesy of Wilcox et al 2021

Surprisal and Fixation Time



Courtesy of Wilcox et al 2021

- **Surprisal:** log(1/P(w|C))
  - Smaller probability --> higher surprisal

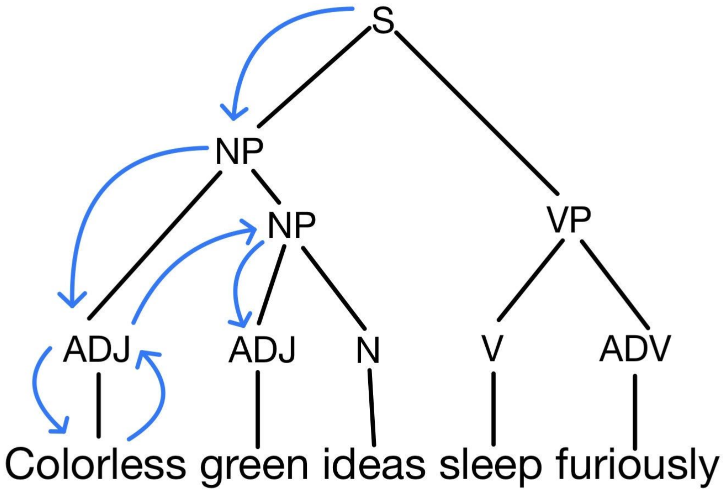
- **Surprisal** has been used with language models to model processing difficulty, but underpredicts the magnitude of garden path effects measured in humans
  - If surprisal + neural language model is an accurate model of garden path processing, we expect a linear relationship between surprisal and fixation/reading time

Surprisal and Memory Limitations

- **Hypothesis:** approximating the probability distribution $P(w|C)$ with limited parallel hypotheses via beam search or particle filtering will increase surprisal effects
    - Beam search will inflate surprisal effects at disambiguating words
    - In the presence of structural ambiguity, surprisal will be inflated for all words under particle filtering (Jensen's Inequality).

Our Model: Recurrent Neural Network Grammar (RNNG)

- Probabilistic model that **generates syntactic trees corresponding to structural hypotheses via depth-first search / top-down parsing** (Dyer et al 2016).
  - Explicit representation of structure is important for garden path effects, which result from structural ambiguity
- Three types actions are probabilistically generated by the model and are used to create the trees via a stack-based algorithm:
  - **NT:** open a non-terminal (e.g. NP)
  - **SHIFT:** add the next terminal (i.e. word)
  - **REDUCE:** close the current non-terminal

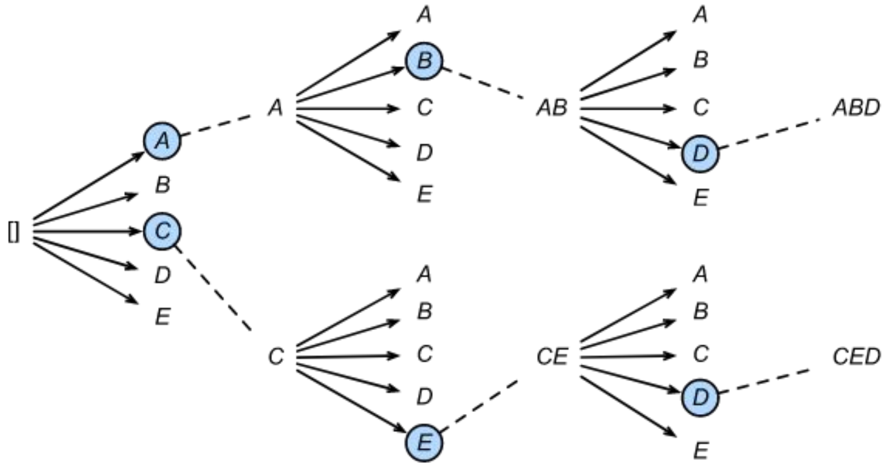


Our Model: Working Memory Limitations

- We use an RNNG trained on the BLLIP corpus (1.75 million sentences)
- We try three models of working memory limitations, all of which **keep a weighted set of *k* hypotheses** in parallel:
  - Word-synchronous beam search
  - Particle filtering
  - Particle filtering with resampling

Word-Synchronous Beam Search

- Variant of beam search where at each word, **the model recursively enumerates and applies all possible next actions until enough of the high-scoring states reach the next lexical action** (Hale et al 2018)
    - The beam is composed of the top *k* of the actions that reach the next lexical state
    - Ensures that all hypotheses at each timestamp end in the same lexical action corresponding to the generation of the next word

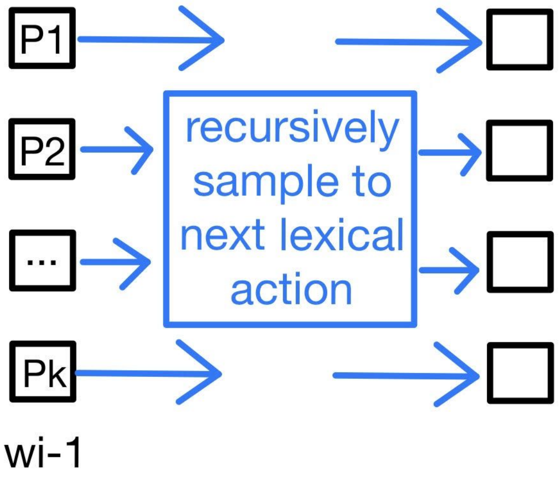


Particle Filtering

- **Sequential Monte Carlo method** to approximate the probability of the i^{th} word w_i given the set of previous actions/structure $y_{1 \dots i}$

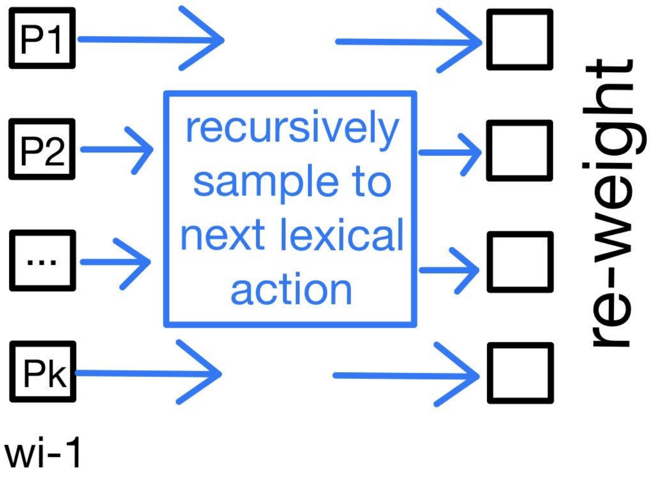
Particle Filtering

- **Sequential Monte Carlo method** to approximate the probability of the i^{th} word w_i given the set of previous actions/structure $y_{1...i}$
  - For each particle, we recursively sample and apply actions until we get to a lexical action



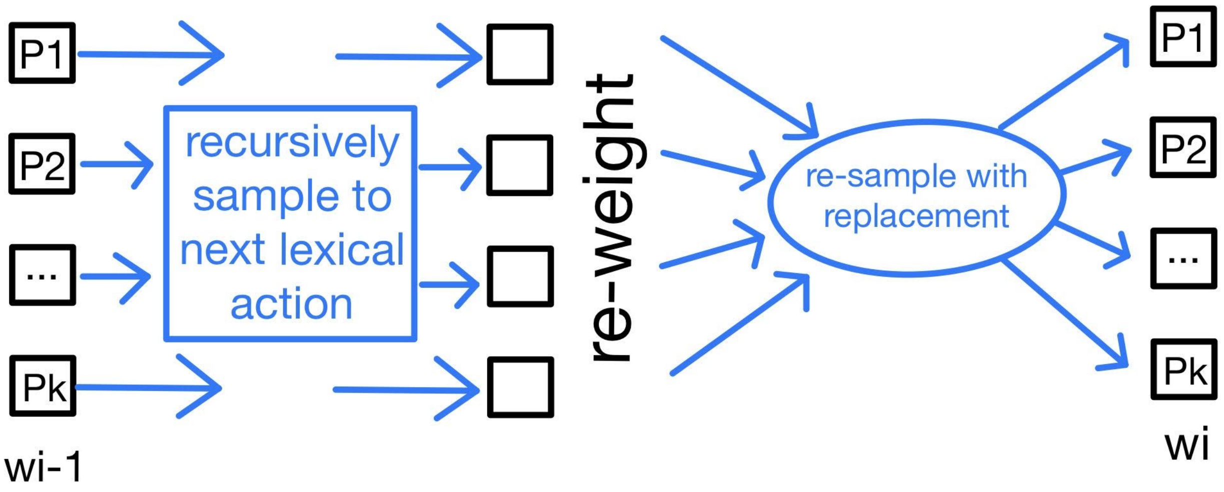
Particle Filtering

- **Sequential Monte Carlo method** to approximate the probability of the i^{th} word w_i given the set of previous actions/structure $y_{1...i}$
    - For each particle, we recursively sample and apply actions until we get to a lexical action
    - We re-weight each particle by the probability of w_i occurring in that structure: $P(w_i \mid y_{i...1})$



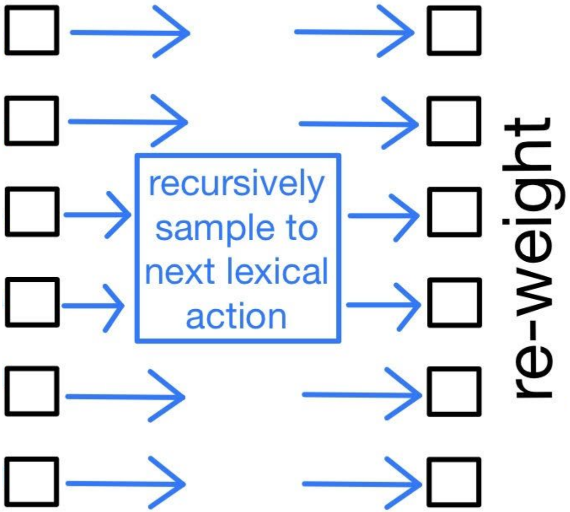
Particle Filtering

- **Sequential Monte Carlo method** to approximate the probability of the i^{th} word w_i given the set of previous actions/structure $y_{1...i}$
  - For each particle, we recursively sample and apply actions until we get to a lexical action
  - We re-weight each particle by the probability of w_i occurring in that structure: $P(w_i \mid y_{i...1})$
  - Finally, we re-sample with replacement to get a set of k particles for w_i.



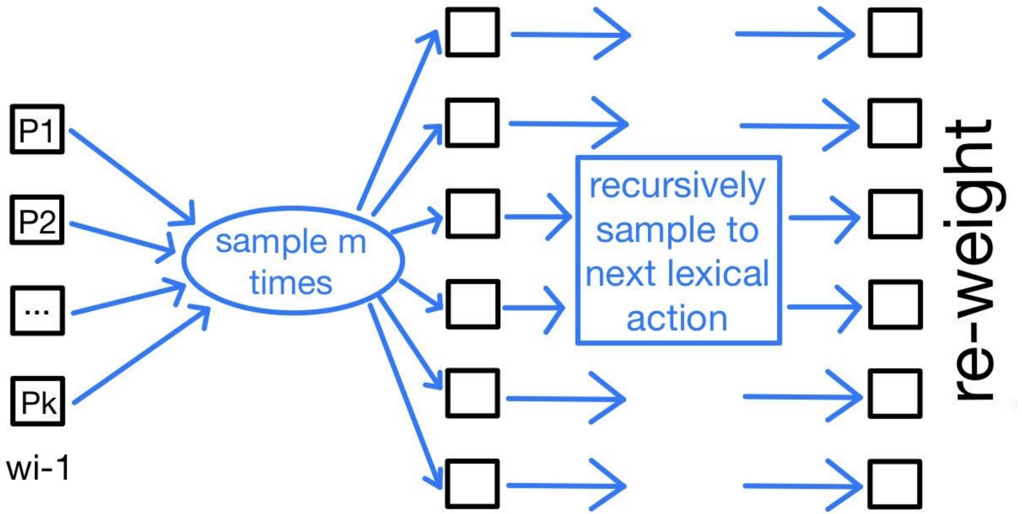
Particle Filtering with Resampling

- A modified version of particle filtering where we sample *m*, *m* > *k*, values from our *k* particles and recursively resample with each until we reach a lexical action in order to **better approximate the action distribution.**



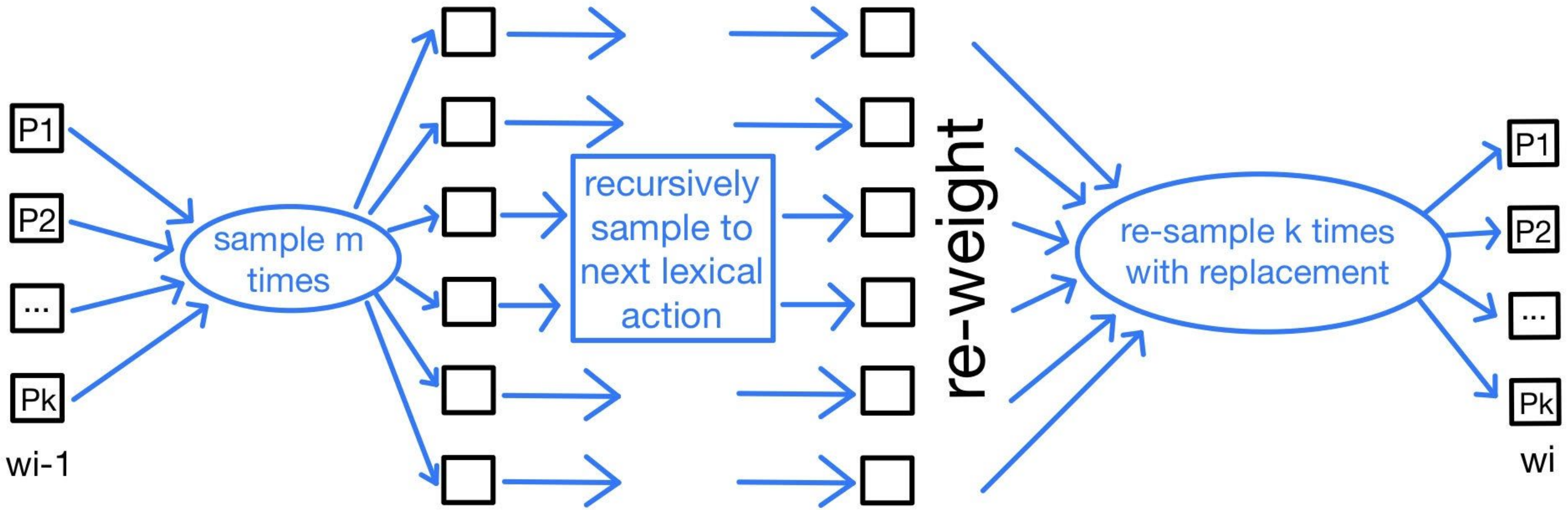
Particle Filtering with Resampling

- A modified version of particle filtering where we sample *m*, *m* > *k*, values from our *k* particles and recursively resample with each until we reach a lexical action in order to **better approximate the action distribution.**



Particle Filtering with Resampling

- A modified version of particle filtering where we sample *m, m > k,* values from our *k* particles and recursively resample with each until we reach a lexical action in order to **better approximate the action distribution.**



Model Testing and Comparison

- We compare our model's predictions against human i-maze times for main verb/reduced relative (MVRR) ambiguity
- We consider the effects of k (the simulated number of hypotheses in working memory) on Noun-Phraze/Zero (NPZ) ambiguity

Garden Path Sentences: Main Verb/Reduced Relative (MV/RR) Ambiguity

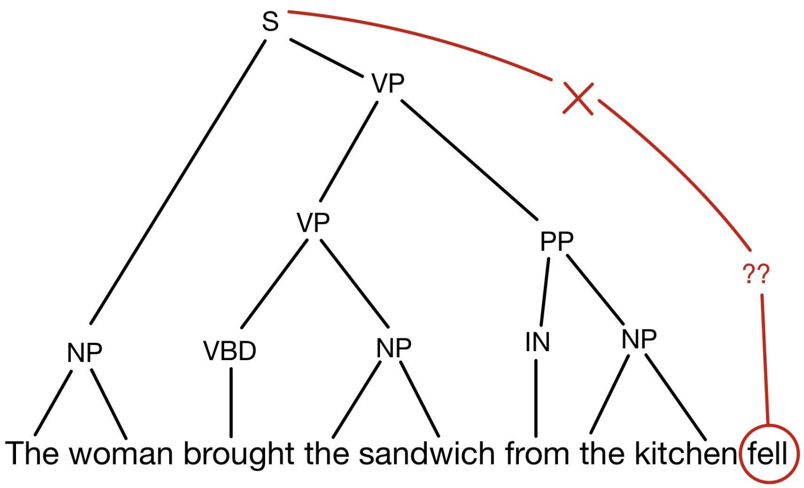
These sentences cause garden paths by **leading the reader to interpret the start of a relative clause as a main verb.** We use 2x2 conditions:

1. Ambiguity of the verb:

Garden Path: "The woman <u>brought</u> the sandwich from the kitchen fell"
Unambiguous: "The woman <u>given</u> the sandwich from the kitchen fell"

2. Reduction of the relative clause:

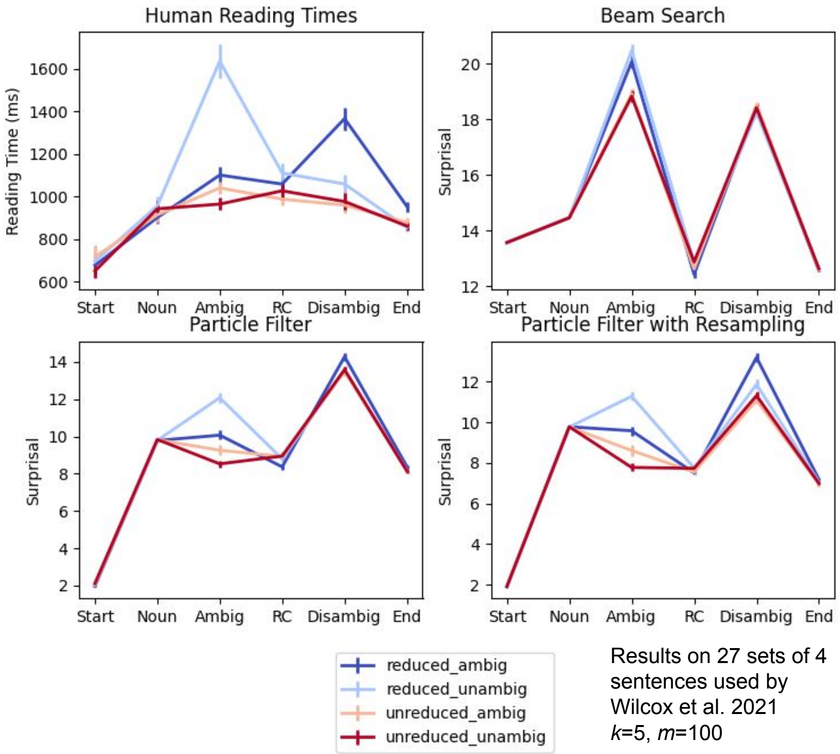
Garden Path: "The woman <u>brought</u> the sandwich from the kitchen fell"
Unambiguous: "The woman <u>who was brought</u> the sandwich from the kitchen fell"



Garden Path Sentences: MV/RR Ambiguity

- We test the three models on 27 sets of 4 sentences used by Wilcox et al. 2021 and compare to the human results from this study.

- We see correct relative surprisals at the ambiguous verb and disambiguator
  - However, the relative magnitudes of surprisal at the ambiguous verb and disambiguator do not match humans
  - We also see a spike across all conditions for the disambiguator not present in humans



Results on 27 sets of 4 sentences used by Wilcox et al. 2021
k=5, *m*=100

Garden Path Sentences: Noun Phrase Zero (NP/Z) Ambiguity

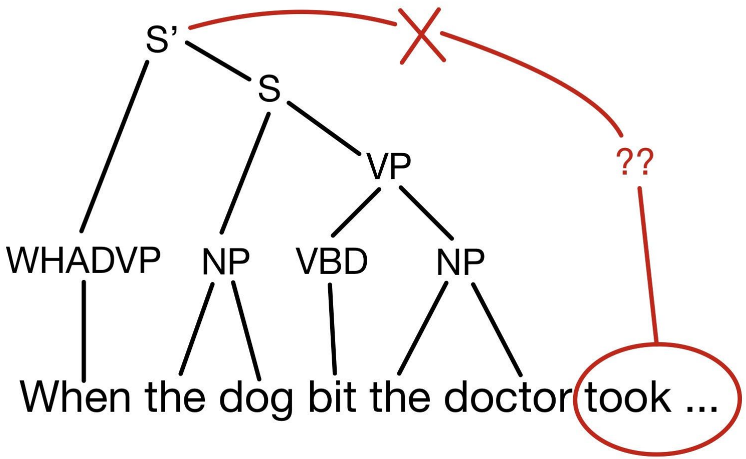
These sentences cause garden paths by **leading the reader to interpret the subject of the second clause as the object of the first clause**. We use 2x2 conditions:

1. Transitivity of the verb:

Garden Path: "When the dog <u>bit</u> the doctor took off the restraint"
Unambiguous: "When the dog <u>struggled</u> the doctor took off the restraint"

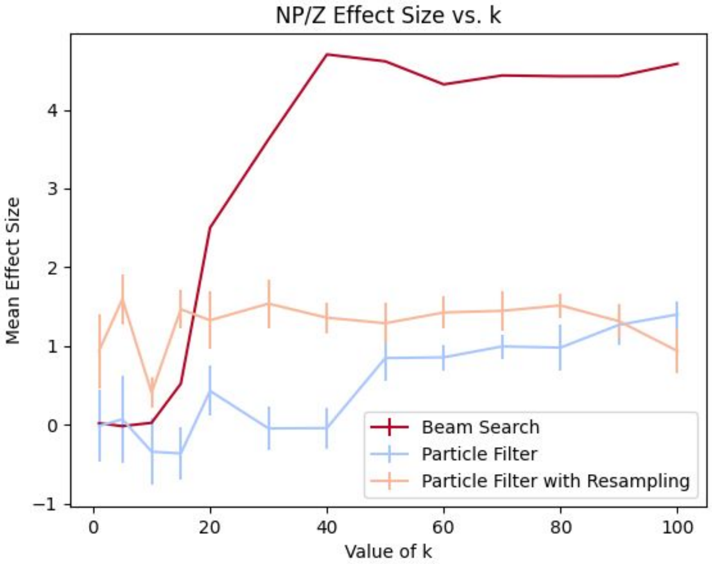
2. Comma between clauses:

Garden Path: "When the dog <u>bit</u> the doctor took off the restraint"
Unambiguous: "When the dog <u>bit,</u> the doctor took off the restraint"



Garden Path Sentences: NP/Z Ambiguity

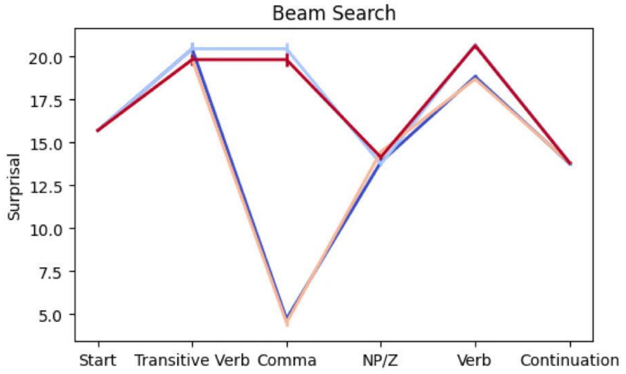
- From reading times, we have that the difference in disambiguator surprisal between the comma and no-comma cases should be greater for the transitive than intransitive verb.
- We measure effect size as transitive difference - intransitive difference



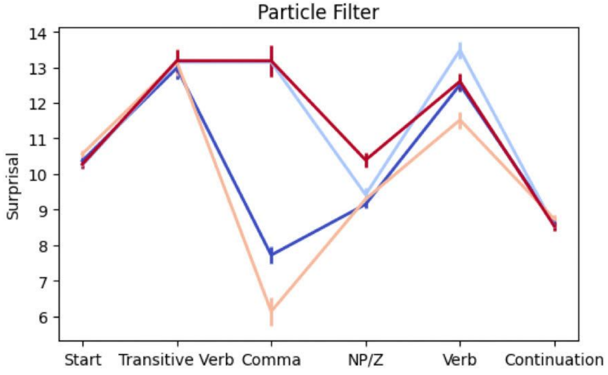
For particle filtering and beam search, the effect size is larger for larger values of *k*, but for particle filter with re-sampling, it is similar for all *k* and slightly larger for smaller values of *k*.

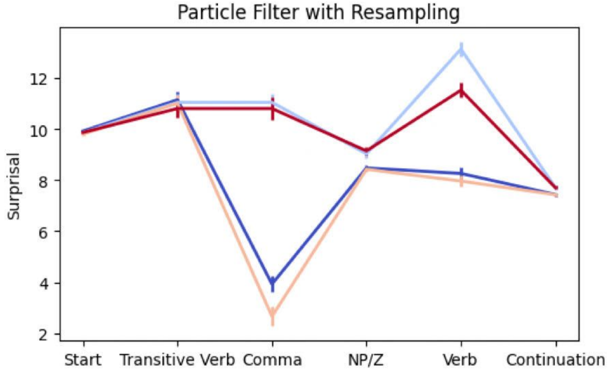
Results on 24 sets of 4 sentences from Hu et al. 2020, *m*=100

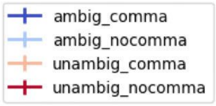
Garden Path Sentences: NP/Z Ambiguity



- At comma/lack thereof, no-comma conditions show far higher surprisal
- At disambiguating verb, spike in surprisal
  - Only particle filtering with resampling differentiates between the spikes for the comma and no-comma conditions
  - Only particle filtering with resampling shows the interaction we expect to see



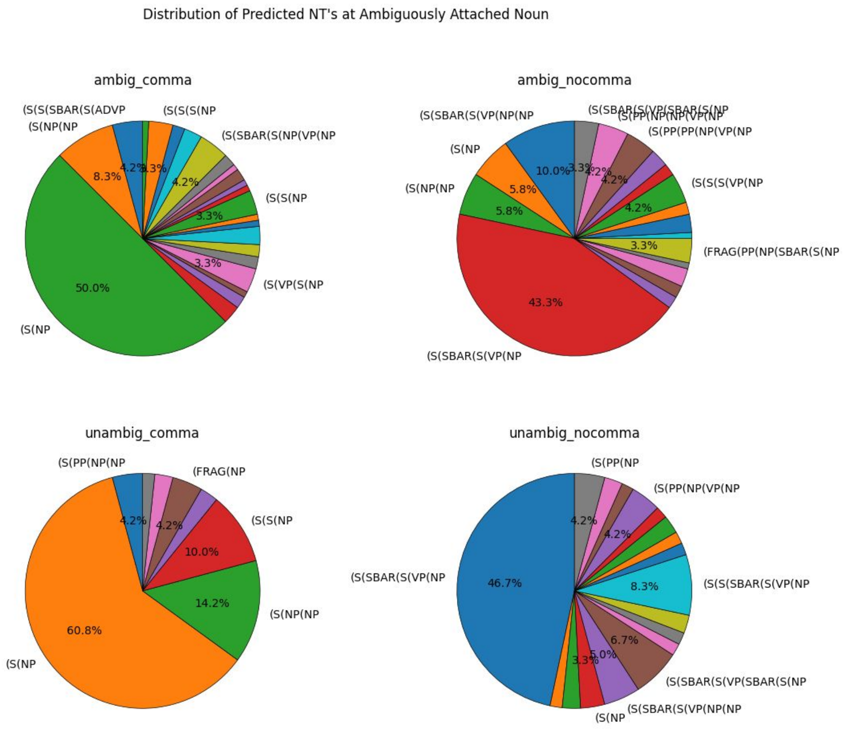




Results on 24 sets of 4 sentences from Hu et al. 2020, *k=5, m*=100

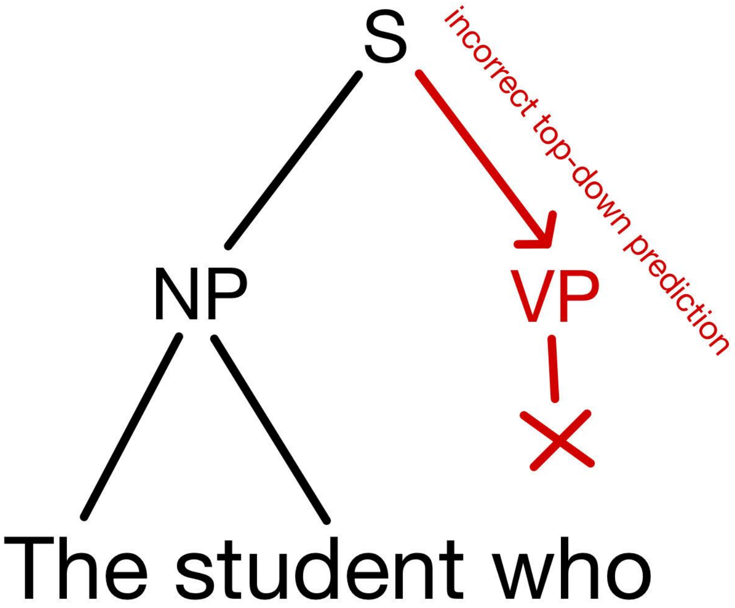
Garden Path Sentences: NP/Z Ambiguity

- Does the model know transitivity?
  - More likely to predict the garden path parse in the unambiguous case than the ambiguous case



Parsing Order and Distribution Approximation

- If the model makes an incorrect top-down prediction, **it cannot recover when it encounters the next word**.
- Resampling ensures that each hypothesis may be expanded more than once vs. regular particle filtering
- **Future work: explore other parsing orders, such as left corner**



Discussion & Conclusion

- **For smaller values of *k*, a better approximation of the action distribution yields larger garden path effects.**
  - Particle filtering with resampling combines small *k* and a more accurate approximation
- Even under these conditions, however, the model still fails to fully predict human garden path effects.
- **Future:**
  - More accurate approximation of the distribution -- resampling and transitivity issues
  - More accurate parsing algorithms

Thank you for an amazing summer!!!

- My collaborators and mentors: Peng Qian, Ethan Wilcox, Roger Levy, Yevgeni Berzak, and everyone in the Computational Psycholinguistics Lab
- Mandana Sassanfar and everyone in MSRP-Bio