Narrative Cloze Task: Experiment 3

Erin Bennett
erindb@stanford.edu

Background

Scripts are a kind of knowledge that people have that helps them understand narrative texts, including filling in their representation of the story with events that were not mentioned but that probably happened.

Recent NLP work has developed initial script induction models and evaluated their performance at filling in the gaps in narrative text. Some of these models have access only to very pared-down linguistic information (essentially only the main verb of a sentence), which allows them to combine multiple sentences together but potentially looses a lot of important information, too.

Question

How do people fill in gaps in narrative text?

- To what extent can people accurately fill in gaps in natural narrative text?
- Whether or not people are able to fill in the original text, do they tend to agree with one another?
- How detailed does the contextual linguistic information need to be for people to sensibly fill in the gaps in narrative texts?

And how does this relate to the performance of script induction models?

Experiment

Design

51 Stimuli:

- 17 documents randomly sampled from "Dinners from Hell" blog (within Ss)
- 3 event chain / cloze test pairs for each document (between Ss)

3 linguistic conditions (between Ss):

- caveman
- event only
- original text

Pilot results

In the pilot, we lost data from some participants, and unfortunately, this was length/time based and so the most conscientious participants were excluded.

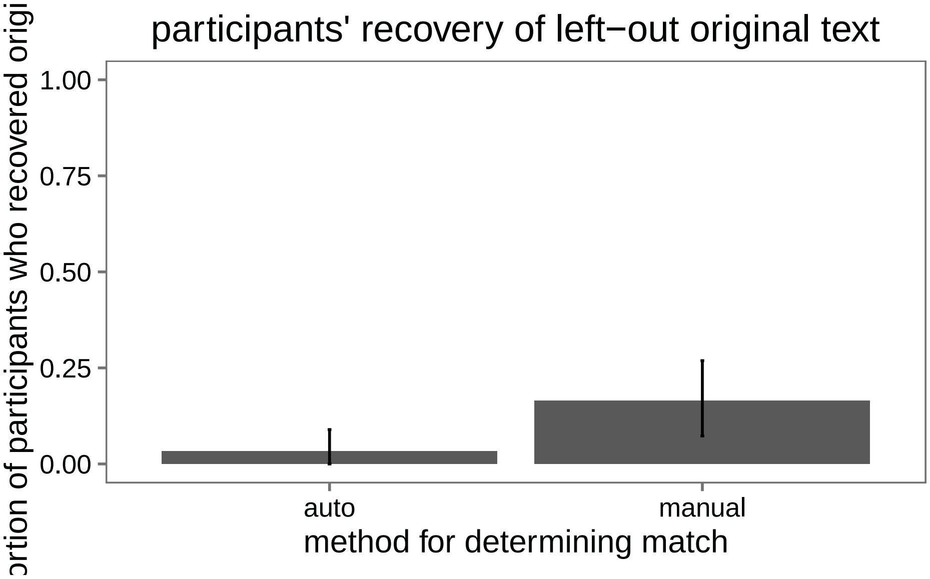
There were 51 cloze tasks in total, and only 17 were seen by each participant. We collected data from 24 participants, so not all of the cloze tasks had enough responses to make meaningful inferences about (e.g.) response overlap across participants.

Question 1: To what extent can people accurately fill in gaps in natural narrative text?

First, we wanted to know whether the main verb of the original sentence was mentioned in the participants' responses. We searched wordnet for synonyms of each of the verbs that participants used checked for any overlap with the synonyms of the original main verb.

Of the 41 stimuli presented with the original passage text as context, only 2 of them had main verbs that were shown to be recovered (i.e. shared a synonym with one a response verb) by at least one participant.

However, hand-coding whether the gist of the response was similar to the gist of the original event, we can see that people are actually doing a lot better on this task. Still, only 0.25% of the stimuli were recovered by any of the participants, and on those conditions, only 0.658% of participants chose that response.

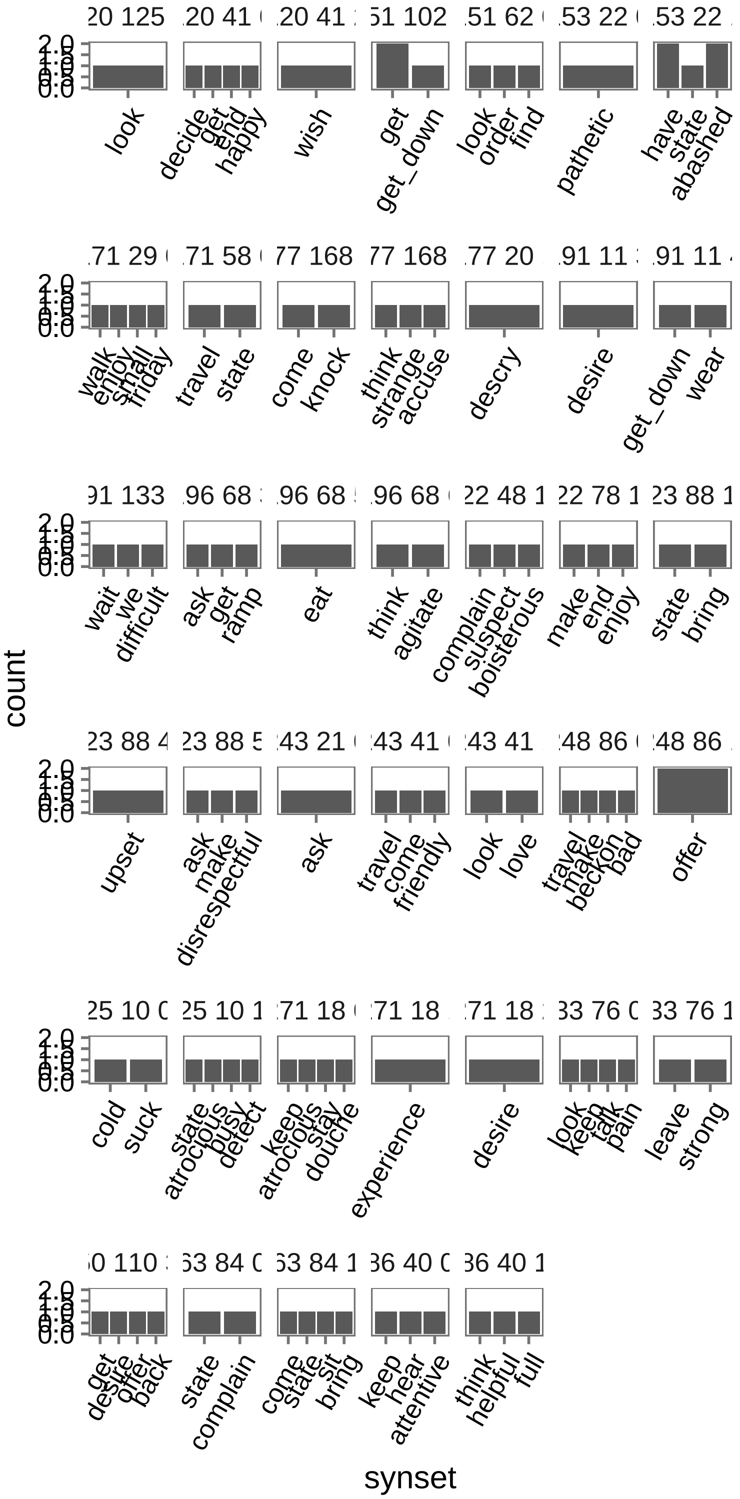


Overall, people don't seem to perform especially well on the task of recovering left-out sentences.

Question 2: Whether or not people are able to fill in the original text, do they tend to agree with one another?

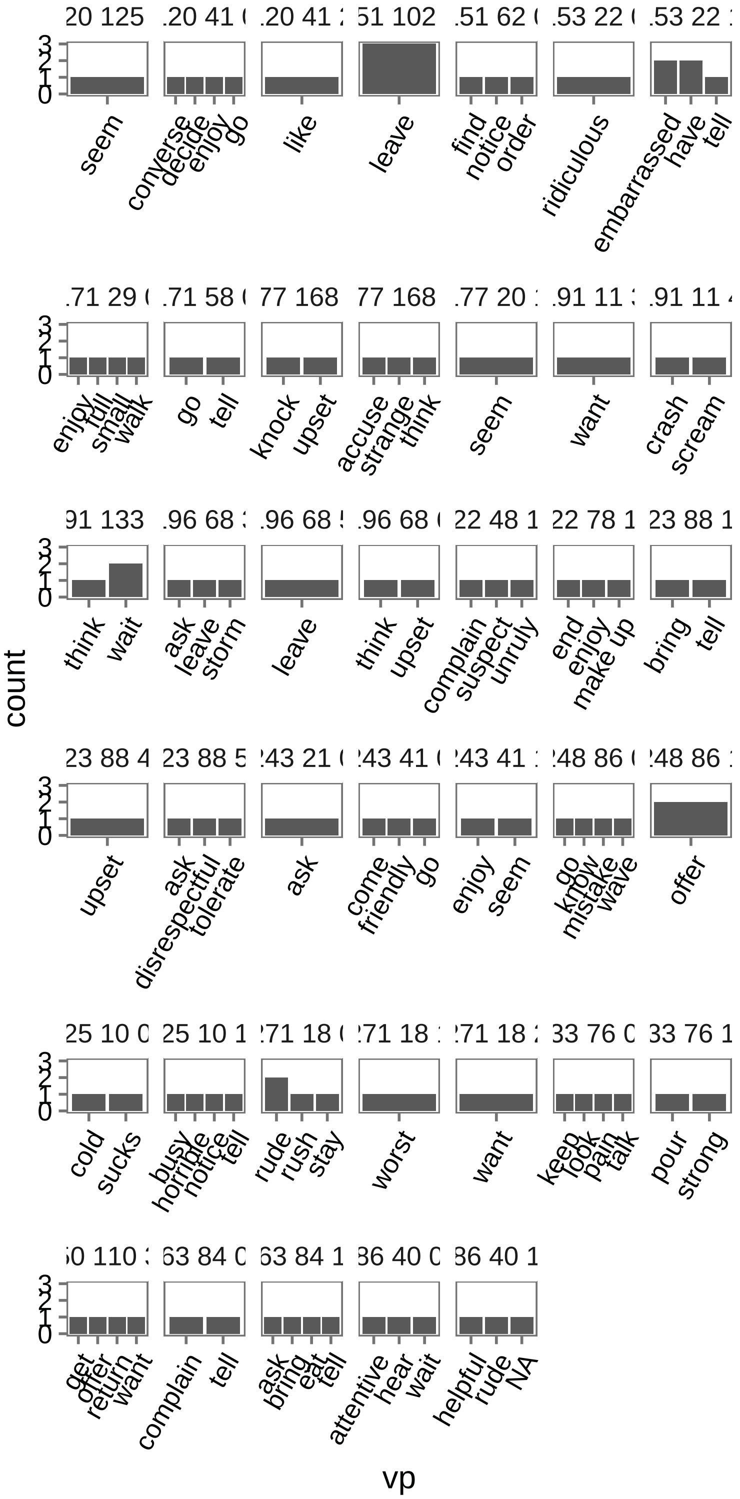
We also wanted to know to what extent people agree on the main verb for the sentences they produce for a given cloze task.

We used Stanford CoreNLP dependency parser to extract the root verb (or occasionally adjectival predicate: e.g. 'red' in 'The apple is red') from the sentences that people produced. For each predicate, we looked for the first synonym set in wordnet. If there was a synonymset for that word, we replaced it with the name of that synset (e.g. we replaced "travel" with its synset name "go").



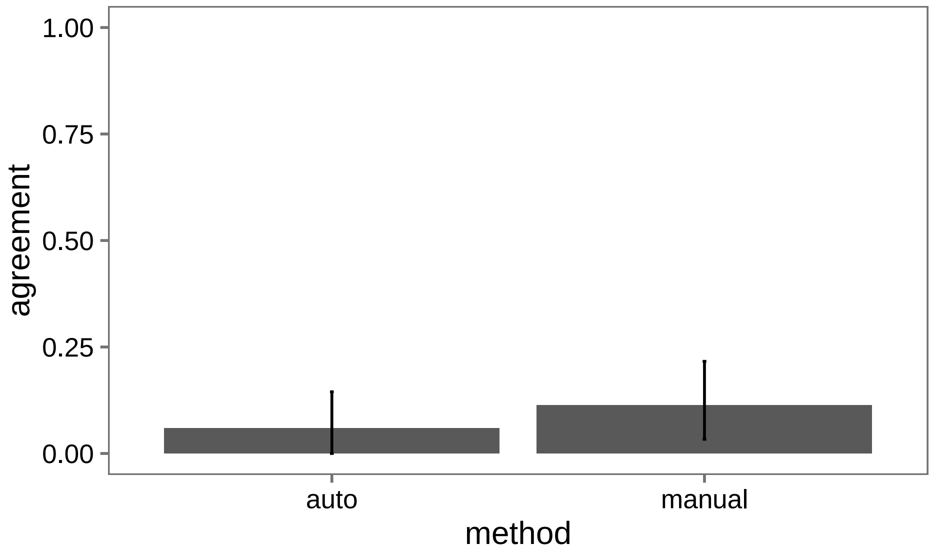
For most of the cloze tasks, there was not much overlap in responses.

In addition to this automated method of finding main event predicates, we also hand-annotated the pilot data with the experimenter's gloss of the main verb.



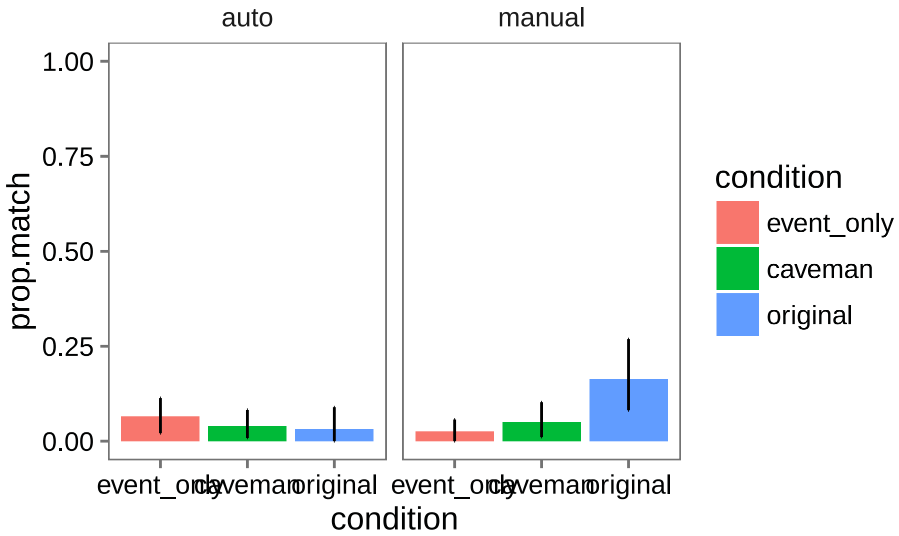
This method shows a tiny bit more overlap in participants' responses.

For all cloze tasks in the pilot where we have data from more than one participant, we calculated the entropy of the distribution of responses and divided by the maximum possible entropy given the number of participants. We subtracted this number from one to get an "agreement" measure. So a value of 0 corresponds to no agreement, and larger values correspond to more agreement. (Entropy by itself will not be an informative measure of agreement, since we have different numbers of participants in each condition.)

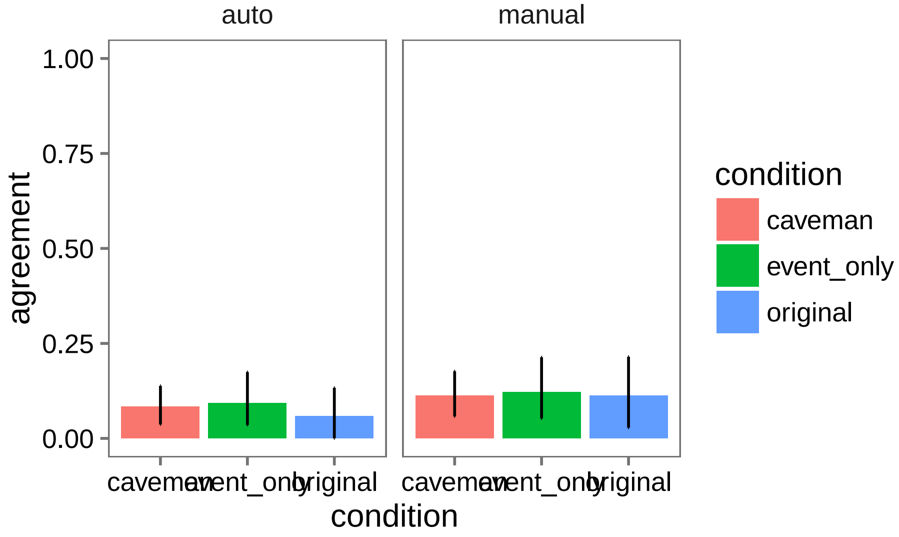


Overall, people didn't agree with one another very much on their answers to these cloze tasks.

Question 3: How detailed does the contextual linguistic information need to be for people to sensibly fill in the gaps in narrative texts?



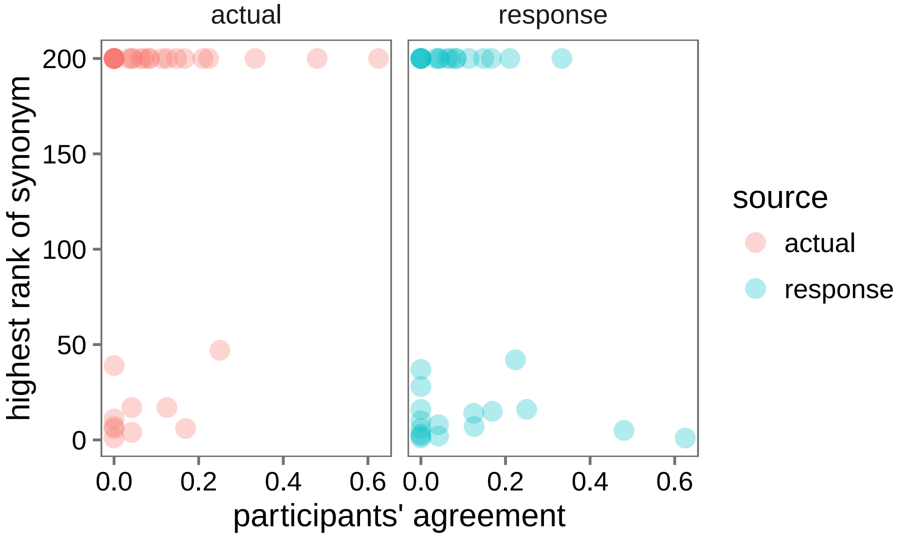
When we manually checked responses for having similar meaning to the original sentences, we find that including the full original text from the context increases the probability that people will recover something close to the original left-out sentence.



We found no difference in agreement across the three linguistic conditions.

Question 4: How does human performance on this task relate to the performance of a PMI-based script induction model?

We measured the PMI model performance on the cloze task by taking the highest ranked guess that was a synonym of (or exactly equal to) the actual main verb or the main verbs that people provided. There doesn't seem to be a relationship between how much people agreed and how highly the model could rank a matching verb. If no synonym matched, or if all rankings were below 200, we recorded the model's ranking as 200.



There does not appear to be a relationship between agreement among people across all conditions and the ranking that PMI models give to the correct event, or the main verbs that people provide.

Informativity in human responses

is there anything in the data that suggests people are being informative?

Informativity model of narrative generation

Future work

One important observation is that most things that are mentioned in a narrative are at least a little bit uncommon given the context, otherwise they would not be interesting/informative enough to mention. How does this affect what people and machines can learn from narrative text?

References