Explanations meeting (6/13/16)

- 3 kinds of projects
  - Idea 1: Computational psycholinguistics: Can we give a meaning to "a because b" that works.
    - Comprehension and Productions / Endorsements
    - What kinds of explanations are we dealing with? [perhaps not worth thinking about yet]
    - Obstacles
      - Lots of potential phenomena
      - Way forward: Design an experimental paradigm that gets interesting explanations
        - Focus on production / endorsement
          - These may not be totally separate because endorsement may depend upon alternatives (which basically makes it production)
        - Methodological point: May not be able to separate intuitive theory from explanations
          - Then, our approach may be to draw coherence between explanations (why questions?), fact questions, predictions, other dependent measures
        - Model
          - Inspired by Lewis, semantics of "a because b" as a counterfactual: (If not b then not a)
            - Something about a and b with presupposition
          - Open questions about the pragmatics model
            - What are the alternatives?
            - Projection / presupposition
            - Slack variables? [in the generative process of counterfactual worlds]
              - Stickiness variable: How happy are you to be in this counterfactual world
              - If A and B themselves are events, then not-A could be the logical negation, or some area in event space
            - QUD?
            - Perhaps the thing you assert the counterfactual are things robust to changes.
        - What's the paradigm?
          - Block world
          - Artifacts, Agents, Kinds, Events, Games
          - Seiver, Gopnik and Goodman (2013). Covariation data of individuals playing with certain toys (why? because it's a cool toy. vs. because she likes it)
          - Short movie clips and get people's explanations? [3-4 events; 15-30s]
            - nonverbal*; Coffee and cigarettes ?*
            - Write down by hand the events in the movie clip.
            - Ask for explanations of some events.
      - Noah talked to L. A. Paul
        - What's the difference between "cause" and "because"
        - Her take: "cause" has to be about events, but "because" can be about more abstract things (e.g. "it's black because it's a tire") <-- MH feels that that's a weird explanation
  - Idea 2: Learning from explanations (both psychology and machine learning)
    - If we had a bunch of "a because b", can we use that to learn a model of the world?
      - And is that better than just "a and b"
      - There are some choices to be made about "what kind of model are you trying to learn"
    - Similar to Leon style learning with RSA (here, + counterfactuals)
      - Counterfactuals could be with respect to either a structured or unstructured model (e.g., generative, multi-layered neural net)
  - Idea 3: ProbProg systems that explain themselves
    - Usual thing: Uncertainty, do inference, query for variable

  - Missing: An explanation of why that variable has that value
  - Why might it?
    - Because something about the observations
    - Because something about the latent variables
    - Because a law-like counterfactual
- Implemented models
  - Church, with easy syntax
  - WebPPL with maybe more complex syntax
- Next steps
  - Videos (criteria outlined above)
  - Write a few simple WebPPL programs: what kinds of explanations could you give for these models?
    - Baking: sometimes order matters
    - Knitting: if you make errors, you have to go back
    - Route finding
    - Video games
    - Real ML models