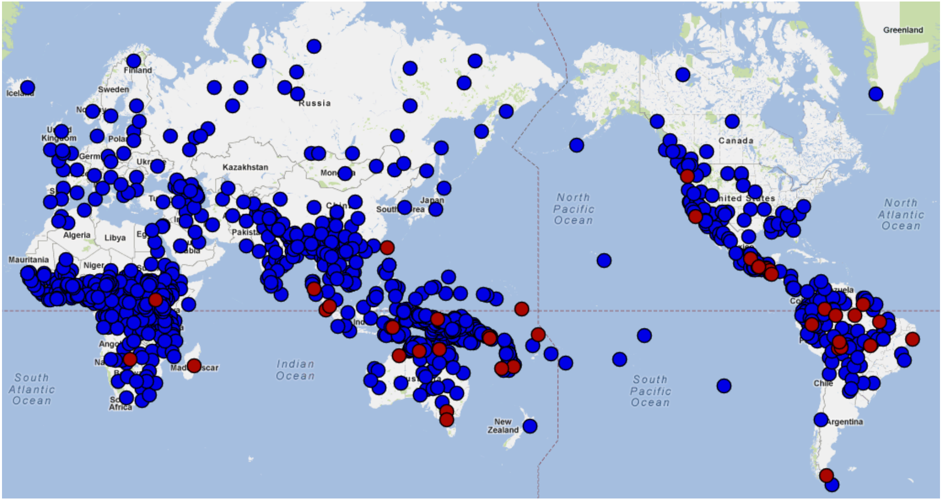
Greenbergian Universals and Bayesian inference

Jenny Culbertson

Simulating Language, 11 March, 2015

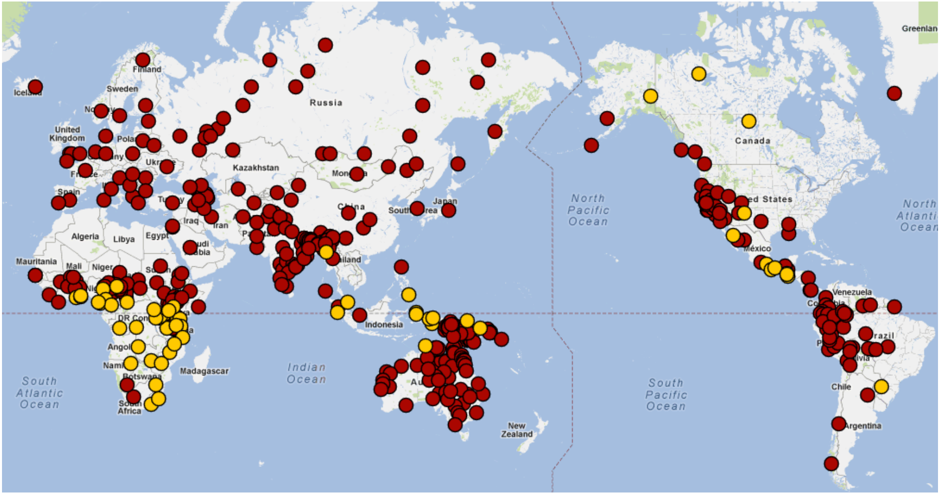
Greenberg's Universal 1

1. SOV, SVO, VSO (not VOS, OSV, OVS)



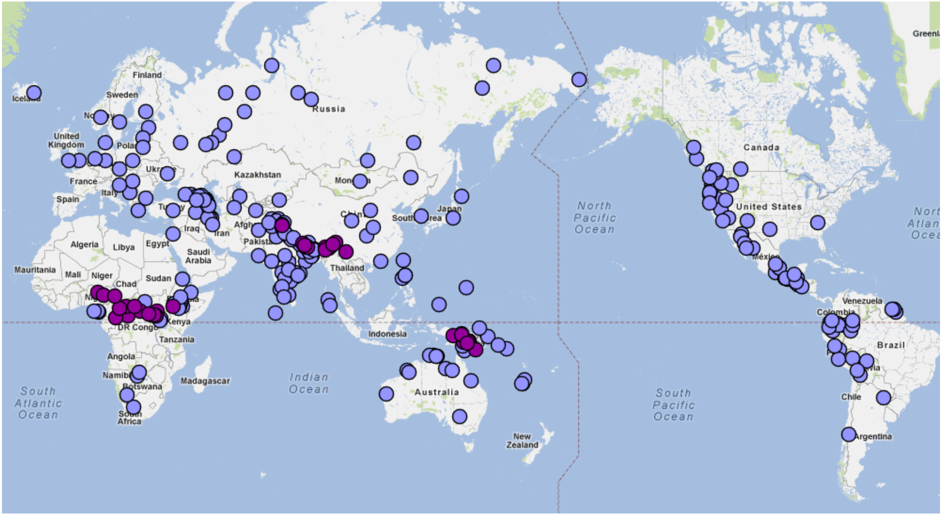
Greenberg's Universal 26

26. Suffixes (not prefixes)



Greenberg's Universal 18

18. If Adjective-Noun → Numeral-Noun



Diversity constrained by cognitive biases?

- Lots of variation across languages
- Lots of confounding factors (e.g.,...?)

Diversity constrained by cognitive biases?

- Lots of variation across languages
- Lots of confounding factors (e.g.,...?)
- But could indicate cognitive biases

Diversity constrained by cognitive biases?

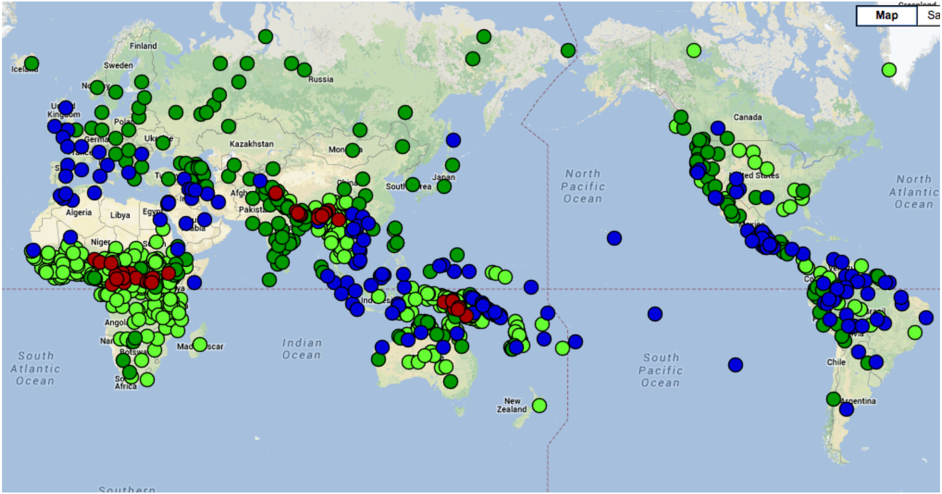
- Lots of variation across languages
- Lots of confounding factors (e.g.,...?)
- But could indicate cognitive biases
    - Cognitive bias = prior bias
    - Non-uniform preference among patterns
    - (Could be innate or learned)
    - (Could be general or specialized for language)

Diversity constrained by cognitive biases?

- Lots of variation across languages
- Lots of confounding factors (e.g.,...?)
- But could indicate cognitive biases
  - Cognitive bias = prior bias
  - Non-uniform preference among patterns
  - (Could be innate or learned)
  - (Could be general or specialized for language)
- How to investigate? Preferences in a single generation??

Universal 18

18. If Adjective-Noun $\rightarrow$ Numeral-Noun



Universal 18

- Actually, there is more than one asymmetry here...

	N-Adj	Adj-N
Num-N	17%	27%
N-Num	52%	4%

Universal 18

- Actually, there is more than one asymmetry here...

	N-Adj	Adj-N
Num-N	17%	27%
N-Num	52%	4%

- Related to another bias you've read about??

Setting up the experiment

- The conditions

	N-Adj	Adj-N
Num-N	3	1
N-Num	2	4

- Easy or hard to learn...?

Setting up the experiment

- The conditions

	N-Adj	Adj-N
Num-N	3	1
N-Num	2	4

- Easy or hard to learn...?
- Adding in regularization...

Setting up the experiment

- The conditions

	N-Adj	Adj-N
Num-N	3	1
N-Num	2	4

- Easy or hard to learn...?
- Adding in regularization...
  - 70% dominant pattern, 30% minority pattern
  - What would regularization look like in this case?

Formulating hypotheses

- Training = listing to Adj-N, N-Adj, Num-N, N-Num phrases
- Testing = producing phrases

Formulating hypotheses

- Training = listing to Adj-N, N-Adj, Num-N, N-Num phrases
- Testing = producing phrases
- Three reasonable hypotheses...

Formulating hypotheses

- Training = listing to Adj-N, N-Adj, Num-N, N-Num phrases
- Testing = producing phrases
- Three reasonable hypotheses...

H1. Learning involves tracking input statistics

Formulating hypotheses

- Training = listing to Adj-N, N-Adj, Num-N, N-Num phrases
- Testing = producing phrases
- Three reasonable hypotheses...

H1. Learning involves tracking input statistics

H2. Learners regularization variation

Formulating hypotheses

- Training = listing to Adj-N, N-Adj, Num-N, N-Num phrases
- Testing = producing phrases
- Three reasonable hypotheses...

H1. Learning involves tracking input statistics

H2. Learners regularization variation

H3. Learners regularize but only orders that are easy to learn

Formulating hypotheses

- In terms of Bayesian inference...

Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

  H2. Input likelihood $\times$ regularization prior

Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

  H2. Input likelihood $\times$ regularization prior

  H3. Input likelihood $\times$ regularization prior $\times$order prior

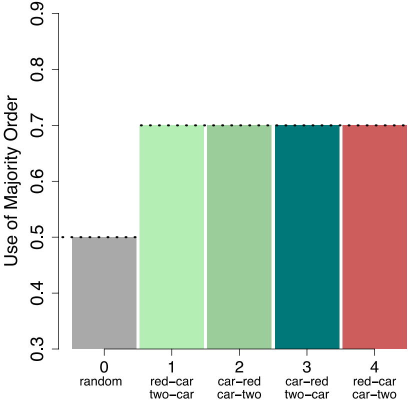
Making predictions

- Three predicted outcomes...

Making predictions

- Three predicted outcomes...

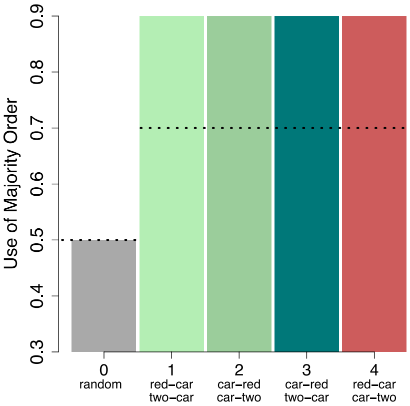
1. Probability matching



Making predictions

- Three reasonable outcomes...

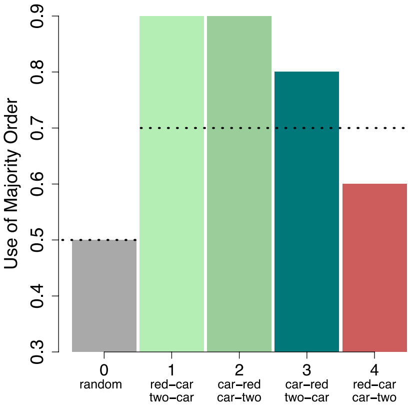
1. Probability matching
2. Across the board regularization



Making predictions

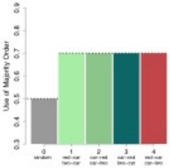
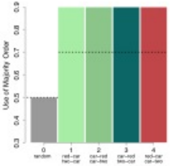
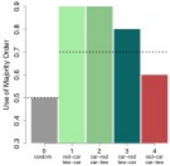
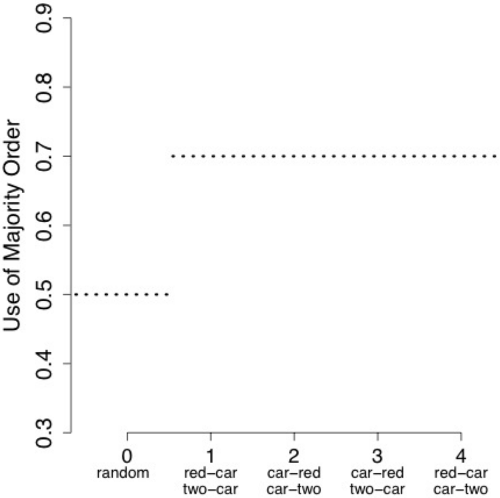
- Three reasonable outcomes...

1. Probability matching
2. Across the board regularization
3. Regularization modulated by order



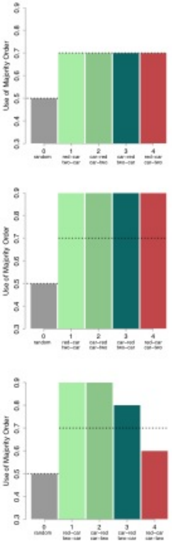
Results

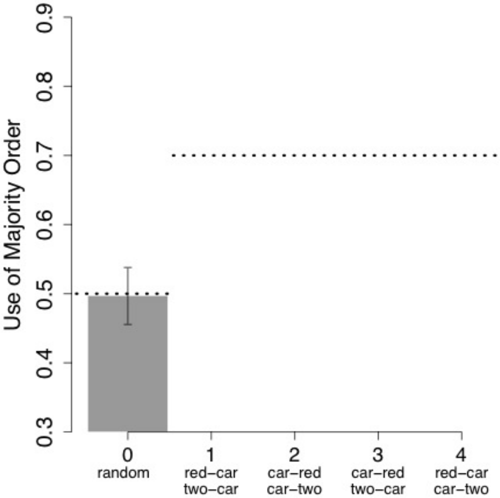
- Participants: 65 native-English-speaking undergrads

Results

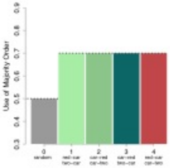
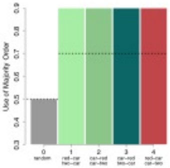
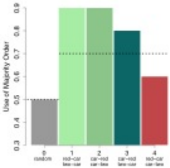
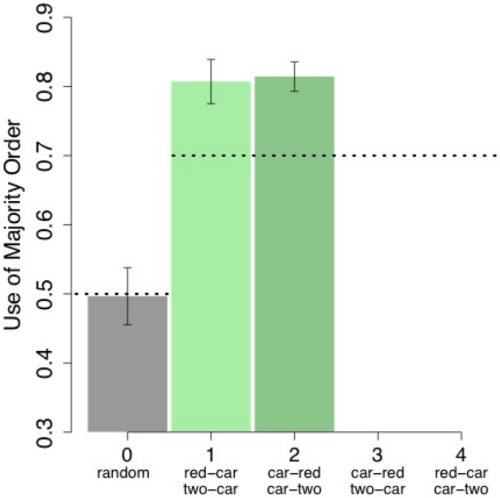
- Participants: 65 native-English-speaking undergrads





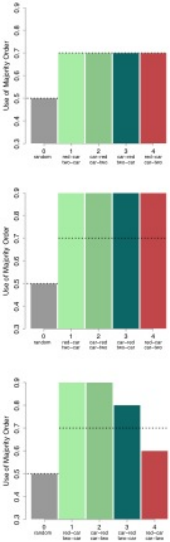
Results

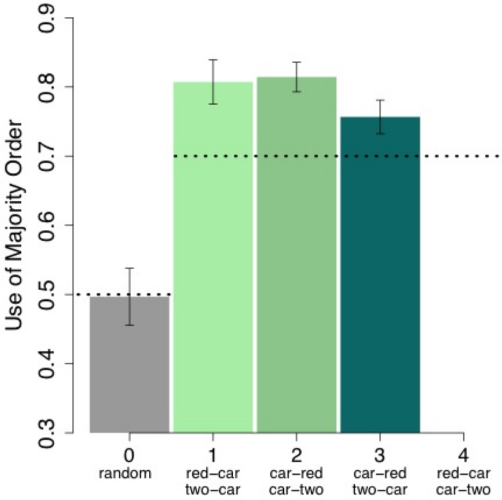
- Participants: 65 native-English-speaking undergrads

Results

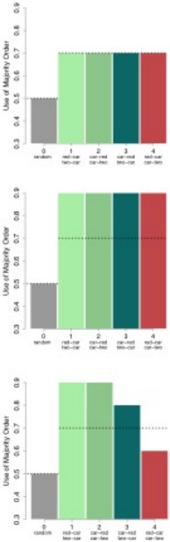
- Participants: 65 native-English-speaking undergrads

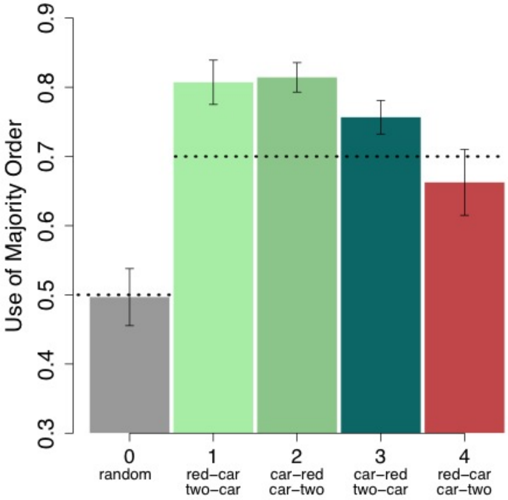




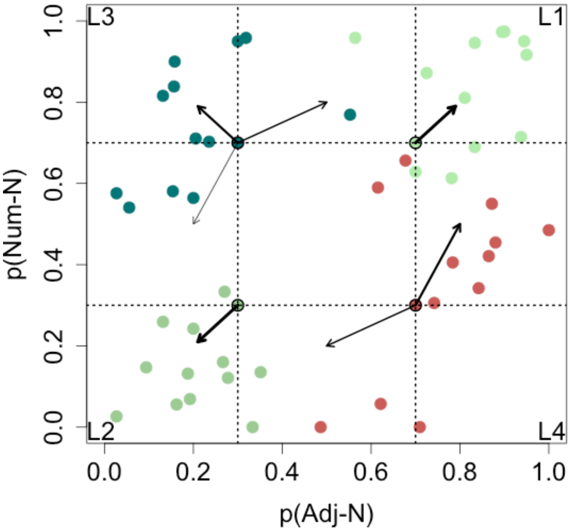
Results

- Participants: 65 native-English-speaking undergrads





Individual learner outcomes



Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

  H2. Input likelihood $\times$ regularization prior

  H3. Input likelihood $\times$ regularization prior $\times$ order prior

Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

  H2. Input likelihood $\times$ regularization prior

  H3. Input likelihood $\times$ regularization prior $\times$ order prior

- Likelihood
- Regularization prior
- Ordering prior

Likelihood

- Coin toss example
  - How many heads out of total tosses?
  - Fair coin?
  - Biased coin?

Likelihood

- Coin toss example
  - How many heads out of total tosses?
  - Fair coin?
  - Biased coin?
- Likelihood

binomial(5 heads $|p =$0.5, 10 tosses$) = 0.2$

binomial(5 heads $|p =$0.9, 10 tosses$) = 0.001$

Likelihood

- Adj, N ordering
  - How many Adj-N out of total Adj utterances?
  - Does the grammar tend to use Adj-N?

Likelihood

- Adj, N ordering
  - How many Adj-N out of total Adj utterances?
  - Does the grammar tend to use Adj-N?
- Likelihood

  binomial(28 Adj-N $|p =$0.5, 40 Adj) $=$ 0.005

Likelihood

- Adj, N ordering
  - How many Adj-N out of total Adj utterances?
  - Does the grammar tend to use Adj-N?
- Likelihood

binomial(28 Adj-N $|p =$0.5, 40 Adj) $=$ 0.005

binomial(28 Adj-N $|p =$0.7, 40 Adj) $=$ 0.14

Likelihood

- Adj, N ordering
  - How many Adj-N out of total Adj utterances?
  - Does the grammar tend to use Adj-N?
- Likelihood

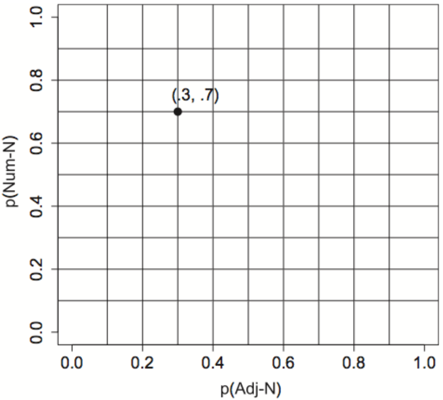
binomial(28 Adj-N $|p=$0.5, 40 Adj) $=$ 0.005

binomial(28 Adj-N $|p=$0.7, 40 Adj) $=$ 0.14

binomial(28 Adj-N $|p=$0.3, 40 Adj) $=$ 0.0000018

Likelihood

- Adj *and* Num ordering
  - Grid of possible probability combos
  - Each assigns likelihood to a set of counts
  - (Total likelihood just multiplies Adj and Num likelihoods)



Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

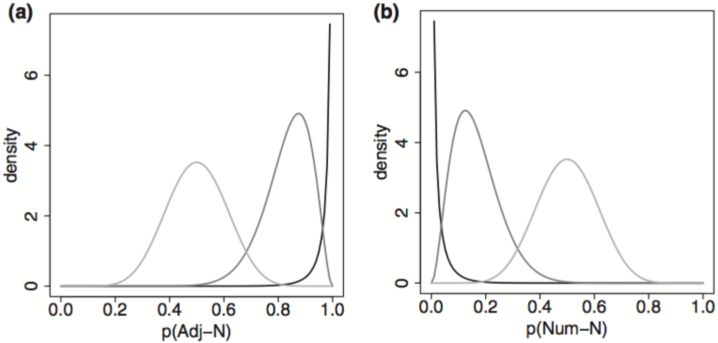
  H2. Input likelihood $\times$ regularization prior $\times$ flat order prior

  H3. Input likelihood $\times$ regularization prior $\times$ biased order prior

- Likelihood
- Regularization prior
- Ordering prior

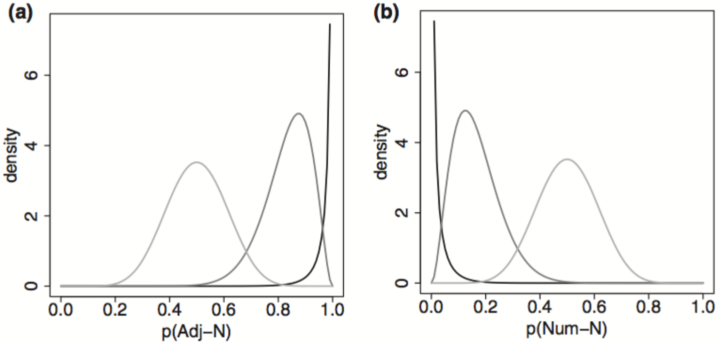
Regularization prior

- Which points in the grid are more likely a priori?



Regularization prior

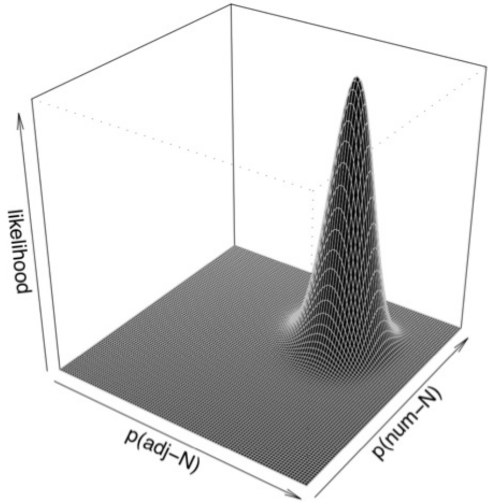
- Which points in the grid are more likely a priori?



- *Asymmetrical* beta distributions: skewed parameters $\rightarrow$ one-way regularization

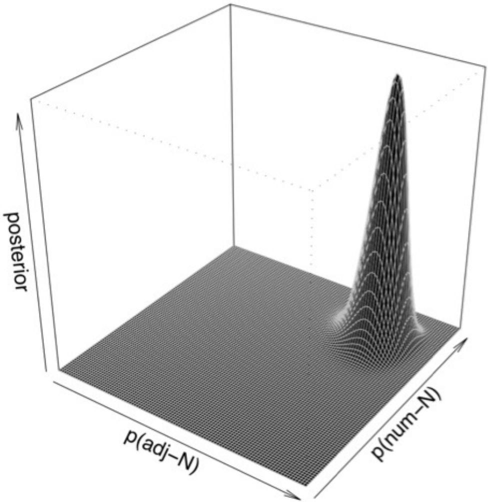
Effect of prior on posterior

- Likelihood alone vs. likelihood $\times$ regularization prior



Effect of prior on posterior

- Likelihood alone vs. likelihood $\times$ regularization prior



Regularization prior

- Which points in the grid are more likely a priori?
- Parameters of the beta: α, β
- Same as the regularization prior from Reali & Griffiths, but asymmetrical

Regularization prior

- Which points in the grid are more likely a priori?
- Parameters of the beta: α, β
- Same as the regularization prior from Reali & Griffiths, but asymmetrical
- Conceptually: prior *counts*, e.g. of Adj-N utterances

Formulating hypotheses

- In terms of Bayesian inference...

  H1. Input likelihood $\times$ flat/uninformative prior

  H2. Input likelihood $\times$ regularization prior $\times$ flat order prior

  H3. Input likelihood $\times$ regularization prior $\times$ biased order prior

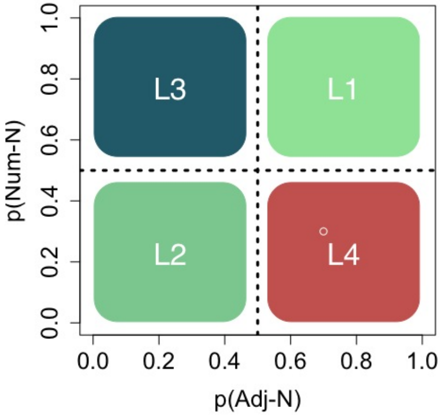
- Likelihood
- Regularization prior
- Ordering prior

Ordering prior

- Which patterns are more likely a priori?
- Combination of two beta distributions gives *pattern type*

Ordering prior

- Which patterns are more likely a priori?
- Combination of two beta distributions gives *pattern type*



Ordering prior

- Which pattern is more likely a priori?
- Combination of two beta distributions gives *pattern type*
- Ordering prior is probability of each type, e.g.

  [0.25, 0.25, 0.25, 0.25]

Ordering prior

- Which pattern is more likely a priori?
- Combination of two beta distributions gives *pattern type*
- Ordering prior is probability of each type, e.g.

  [0.25, 0.25, 0.25, 0.25]

  [what would a biased one look like??]

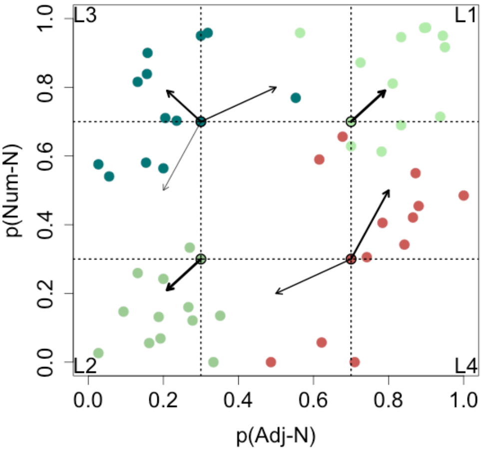
Complete prior

- Complete prior probability of a grammar p(Adj-N), p(Num-N) is a sum over four beta combinations of:
  - prior probability of p(Adj-N) given regularization bias $\times$
  - prior probability of p(Num-N) given regularization bias $\times$
  - prior probability of particular combination of betas
- e.g., prior for p(Adj-N)=0.8, p(Num-N)=0.2

$$
\begin{aligned}
&\mathit{beta}(0.8|\alpha = 10, \beta = 2) \times \mathit{beta}(0.2|\alpha = 10, \beta = 2) \times 0.25+\\
&\mathit{beta}(0.8|\alpha = 2, \beta = 10) \times \mathit{beta}(0.2|\alpha = 2, \beta = 10) \times 0.25+\\
&\mathit{beta}(0.8|\alpha = 2, \beta = 10) \times \mathit{beta}(0.2|\alpha = 10, \beta = 2) \times 0.25+\\
&\mathit{beta}(0.8|\alpha = 10, \beta = 2) \times \mathit{beta}(0.2|\alpha = 2, \beta = 10) \times 0.25+
\end{aligned}
$$

Looking for prior biases

- What parameters make the testing data most likely?

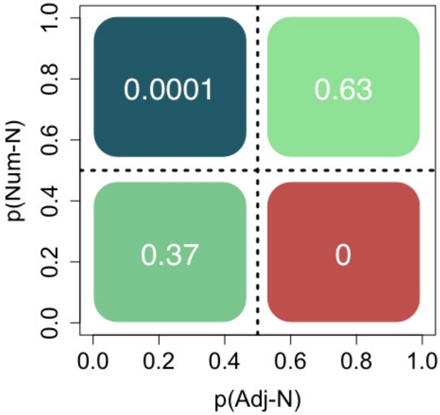


Looking for prior biases

- What parameters make the testing data most likely?
- Regularization parameters (α, β) very skewed (16.5, 0.001)

Looking for prior biases

- What parameters make the testing data most likely?
- Regularization parameters (α, β) very skewed (16.5, 0.001)
- Prior probability of pattern types:



Posterior (finally!)

- What kinds of p(Adj-N), p(Num-N) pairs are learners likely to acquire given set of prior parameters?

Posterior (finally!)

- What kinds of p(Adj-N), p(Num-N) pairs are learners likely to acquire given set of prior parameters?

- Prior probability of p(Adj-N)=high, p(Num-N)=high is high
- Prior probability of p(Adj-N)=low, p(Num-N)=low is high
- Prior probability of p(Adj-N)=low, p(Num-N)=high is pretty low
- Prior probability of p(Adj-N)=high, p(Num-N)=low is zero!

	N-Adj	Adj-N
Num-N	17%	27%
N-Num	52%	4%

For the lab...

- Calculate posterior distributions
- Recreate model predictions
- Investigate the effect of the prior parameters on predicted grammars
- Extra-credit: iterate it