One way to think about the connection of computation with acquisition

- Computation = information processing done by human minds during language acquisition
  - Theoretical research: what is it that’s being computed
    - Ex: knowledge of phonological/syntactic/semantic structure, where words are in fluent speech
  - Experimental research: when it’s being computed & constraints on how it’s computed
    - Ex: known/achieved by a certain age, with cognitive limitations on memory and processing
One way to think about the connection of computation with acquisition

- Computation = information processing done by human minds during language acquisition
  - Theoretical research: what is it that’s being computed
  - Experimental research: when it’s being computed & constraints on how it’s computed
  - Corpus research: what it’s being computed from
    - Ex: which data appear in the input and with what frequency

Poverty of the stimulus claim depends directly on what, when, how, and what from

Poverty of the stimulus is one motivation for Universal Grammar: what children need to accomplish these computations
- domain-specific or domain-general
- innate/maturing or derived from prior experience
- (NSF) “Testing the Universal Grammar Hypothesis” with Jon Sprouse: syntactic islands
- Pearl & Lidz (2009): English anaphoric one

Another way to think about the connection of computation with acquisition

- Computation = information processing done by computers to help understand the information processing done by human minds during language acquisition

Modeling learnability vs. modeling acquirability

- Modeling learnability
  - “Can it be learned at all by a simulated learner?”
  - “ideal”, “rational”, or “computational-level” learners
  - what is possible to learn

- Modeling acquirability (Johnson 2004)
  - “Can it be learned by a simulated learner that is constrained in the ways humans are constrained?”
  - more “realistic” or “cognitively inspired” learners
  - what is possible to learn if you’re human
Adapting a learnability model to be an acquirability model:
Word segmentation (Pearl, Goldwater, & Steyvers forthcoming, submitted)
- Do ideal learner solutions transfer to constrained learners?
- Surprise finding: constrained learners can do as well or better

Surprise finding: constrained learners can do as well or better

Considering acquirability and learnability:
Metrical phonology (Pearl 2008, 2009, submitted)
- Framework for testing theories of knowledge representation: using an argument from acquisition
- Benefits: informing theory and informing acquisition

Input (specific linguistic observations)
Abstract internal representation/generalization
Output (specific linguistic productions)

Language acquisition computation as induction

Typically an ideal observer approach asks what the optimal solution to the induction problem is, given particular assumptions about knowledge representation and available information.

Constrained learners implement ideal learners in more cognitively plausible ways.
- How might limitations on memory and processing affect learning?
Word segmentation

- A big deal: basis for more complex linguistic knowledge

- Cognitive modeling: Given a corpus of fluent speech or text (no utterance-internal word boundaries), we want to identify the words.

- One of the first problems infants must solve when learning language.

- Infants make use of many different cues.
  - Phonotactics, allophonic variation, metrical (stress) patterns, effects of co-articulation, and statistical regularities in syllable sequences.

- Statistics may provide initial bootstrapping.
  - Used very early (Thiessen & Saffran, 2003)
  - Language-independent, so doesn't require children to know some words already

Bayesian inference

- Useful tool for linguistic research: a more sophisticated form of statistical learning that does not require us to trivialize the complexity of linguistic knowledge

- Allows us to combine probabilistic methods with structured linguistic representations and predict the likelihood of things we rarely or never see (allowing generalizations from a data subset)
Bayesian inference: model goals

- The Bayesian learner seeks to identify an explanatory linguistic hypothesis that accounts for the observed data.
- Conforms to prior expectations.

\[ P(h|d) \propto P(d|h)P(h) \]

- Ideal learner: Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.
- Constrained learner: Use same probabilistic model, but algorithm reflects how humans might implement the computation.

Bayesian segmentation

- In the domain of segmentation, we have:
  - Data: unsegmented corpus (transcriptions)
  - Hypotheses: sequences of word tokens

\[ P(h|d) \propto P(d|h)P(h) \]

- 1 if concatenating words forms corpus,
  0 otherwise.

Corpus: "look at the doggie"
- P(d) = 1
- P(d) = 0

 enforce assumptions or biases in the learner.

Optimal solution is the segmentation with highest probability.

Goldwater, Griffiths, and Johnson (2007, 2009)
Investigating learner assumptions

- If a learner assumes that words are independent units, what is learned from realistic data? [unigram model]
- What if the learner assumes that words are units that help predict other units? [bigram model]

Approach of Goldwater, Griffiths, & Johnson (2007, 2009): use a Bayesian ideal observer to examine the consequences of making these different assumptions.

Corpus: child-directed speech samples

- Bernstein-Ratner corpus:
  - 9790 utterances of phonemically transcribed child-directed speech (19-23 months), 33399 tokens and 1321 unique types.
  - Average utterance length: 3.4 words
  - Average word length: 2.9 phonemes

- Example input:
  
Results: Ideal learner (Standard MCMC)

<table>
<thead>
<tr>
<th></th>
<th>Word Tokens</th>
<th>Boundaries</th>
<th>Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>Prec</td>
</tr>
<tr>
<td>Ideal (unigram)</td>
<td>61.7</td>
<td>47.1</td>
<td>92.7</td>
</tr>
<tr>
<td>Ideal (bigram)</td>
<td>74.6</td>
<td>68.4</td>
<td>90.4</td>
</tr>
</tbody>
</table>

Correct segmentation: “look at the doggie. look at the kitty.”

Best guess of learner: “lookat the doggie. lookat the kitty.”

- Word Token Prec = 2/5 (0.4), Word Token Rec = 2/8 (0.25)
- Boundary Prec = 3/3 (1.0), Boundary Rec = 3/6 (0.5)
- Lexicon Prec = 2/4 (0.5), Lexicon Rec = 2/5 (0.4)

Results: Ideal learner (Standard MCMC)

<table>
<thead>
<tr>
<th></th>
<th>Word Tokens</th>
<th>Boundaries</th>
<th>Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>Prec</td>
</tr>
<tr>
<td>Ideal (unigram)</td>
<td>61.7</td>
<td>47.1</td>
<td>92.7</td>
</tr>
<tr>
<td>Ideal (bigram)</td>
<td>74.6</td>
<td>68.4</td>
<td>90.4</td>
</tr>
</tbody>
</table>

- The assumption that words predict other words is good: bigram model generally has superior performance
- Note: Training set was used as test set
- Both models tend to undersegment, though the bigram model does so less (boundary precision > boundary recall)
How about constrained learners?

- The constrained learners use the same probabilistic model, but process the data incrementally (one utterance at a time), rather than all at once.

  - Dynamic Programming with Maximization (DPM)
  - Dynamic Programming with Sampling (DPS)
  - Decayed Markov Chain Monte Carlo (DMCMC)

Pearl, Goldwater, & Steyvers forthcoming, submitted

Dynamic Programming: Maximization

For each utterance:
- Use dynamic programming to compute probabilities of all segmentations, given the current lexicon.
- Choose the best segmentation.
- Add counts of segmented words to lexicon.

Algorithm used by Brent (1999), with different model.

Results: Ideal learner sample segmentations

Unigram model

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to see the book</td>
<td>0.33</td>
</tr>
<tr>
<td>Look there's a boy with his hat and a doggie</td>
<td>0.21</td>
</tr>
<tr>
<td>You want to look at this</td>
<td>0.15</td>
</tr>
</tbody>
</table>
| Have a drink | ...

Bigram model

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to see the book</td>
<td>0.33</td>
</tr>
<tr>
<td>Look there's a boy with his hat and a doggie</td>
<td>0.21</td>
</tr>
<tr>
<td>You want to look at this</td>
<td>0.15</td>
</tr>
</tbody>
</table>
| Have a drink | ...

Considering human limitations

What if the only limitation is that the learner must process utterances one at a time?

Pearl, Goldwater, & Steyvers forthcoming, submitted
Considering human limitations

What if humans don’t always choose the most probable hypothesis, but instead sample among the different hypotheses available?

Dynamic Programming: Sampling

For each utterance:
- Use dynamic programming to compute probabilities of all segmentations, given the current lexicon.
- Sample a segmentation.
- Add counts of segmented words to lexicon.

Decayed Markov Chain Monte Carlo

For each utterance:
- Probabilistically sample \( s \) boundaries from all utterances encountered so far.
- \( \text{Prob(sample } b \text{)} = b^a \exp(-d) \) where \( b \) is the number of potential boundary locations between \( b \) and the end of the current utterance and \( d \) is the decay rate (Marthi et al. 2002).
- Update lexicon after the \( s \) samples are completed.
Decayed Markov Chain Monte Carlo

For each utterance:
- Probabilistically sample $s$ boundaries from all utterances encountered so far.
- Prob(sample $b$) = $b_a^{-d}$ where $b_a$ is the number of potential boundary locations between $b$ and the end of the current utterance and $d$ is the decay rate (Marthi et al. 2002).
- Update lexicon after the $s$ samples are completed.

Results: unigrams vs. bigrams

![Graph comparing results](chart.png)

Results averaged over 5 randomly generated test sets (~900 utterances) that were separate from the training set (~8800 utterances), all generated from the Bernstein/Ratner corpus.

DMGCMC Unigram: d=1, s=20000
DMGCMC Bigram: d=0.25, s=20000

Note: s=20000 means DMGCMC learner samples 89% less often than the Ideal learner.

Like the Ideal learner, the DPM & DMGCMC bigram learners perform better than the unigram learner, though improvement is not as great as in the Ideal learner. The bigram assumption is helpful.
Results: unigrams vs. bigrams

However, the DPS bigram learner performs worse than the unigram learner. The bigram assumption is not helpful.

\[ F = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}} \]

Precision: \#correct / \#found
Recall: \#found / \#true

Unigram comparison: DPM, DMCMC > Ideal, DPS performance
Interesting: Constrained learners outperforming unconstrained learner when words are believed to be independent units.

Results: unigrams vs. bigrams

Lexicon = a seed pool of words for children to use to figure out language-dependent word segmentation strategies.
Results: unigrams vs. bigrams for the lexicon

Like the Ideal learner, the DPM bigram learner yields a more reliable lexicon than the unigram learner.

However, the DPS and DMCMC bigram learners yield less reliable lexicons than the unigram learners.

Unigram comparison: DMCMC > Ideal > DPM > DPS performance
Interesting: Constrained learner outperforming unconstrained learner when words are believed to be independent units.

Bigram comparison: Ideal > DPM > DMCMC > DPS performance
More expected: Unconstrained learner outperforming constrained learners when words are believed to be predictive units (though not by a lot).
Results: under vs. oversegmentation

Undersegmentation: boundary precision > boundary recall
Oversegmentation: boundary precision < boundary recall

The DMCMC unigram learner, like the ideal learner, tends to undersegment.

All other learners, however, tend to oversegment.

Results: main points

- A better set of cognitively inspired statistical learners
  - While no constrained learners outperform the best ideal learner on all measures, all perform better on realistic child-directed speech data than a transitional probability learner (Gambell & Yang 2006, over syllables: word token F-score = 29.6; Brent 1999, over phonemes: word token precision and recall scores =~ 40, lexicon precision scores =~ 15).

- Ideal learner behavior doesn’t always transfer
  - While assuming words are predictive units (bigram model) significantly helped the ideal learner, this assumption may not be as useful to a constrained learner (depending on how cognitive limitations are implemented).
  - Undersegmentation doesn’t always occur (though it may match children’s behavior better (Peters 1983)).
 Constraints on processing are not always harmful
- Decayed MCMC learner can perform well even with more than 99.9% less processing than the unconstrained ideal learner (ask for details!)
- Constrained unigram learners can sometimes outperform the unconstrained unigram learner (“Less is More” hypothesis: Newport 1990).

More sophisticated statistical learning can be a way to solve the initial chicken-and-egg problem for word segmentation
- Constrained statistical learning, as a language-independent strategy, may provide a lexicon reliable enough for children to learn language-dependent strategies from.

Results: main points

Where to go from here: exploring acquirability

- Explore robustness of constrained learner performance across different corpora and different languages
  - Is it just for this data set of English that we see these effects?
    - English to children aged 9 months or younger (portion of Brent corpus (Brent & Siskind 2001) containing ~28K utterances)
    - (Pearl et al., in prep) results show same performance trends: constrained learners performing equivalently or better than the unconstrained ideal learner
  - Is it just for this language that we see these effects?
    - In progress: Spanish to children a year or younger (portion of JacksonThal corpus (Jackson-Thal 1994) containing ~3600 utterances)

- Simple intuitions about human cognition (e.g., memory and processing limitations) can be translated in multiple ways
  - Here: processing utterances incrementally, keeping a single lexicon hypothesis in memory, implementing recency effects

- Investigate other implementations of constrained learners
  - Imperfect memory: Assume lexicon precision decays over time, assume calculation of probabilities is noisy
  - Knowledge representation: assume syllables are a relevant unit of representation (Jusczyk et al. 1999), assume stressed and unstressed syllables are tracked separately (Curtin et al. 2005, Pelucchi et al. 2009)

Today’s Plan: Both are useful

- Adapting a learnability model to be an acquirability model:
  - Word segmentation (Pearl, Goldwater, & Steyvers forthcoming, submitted)
    - Do ideal learner solutions transfer to constrained learners?
    - Surprise finding: constrained learners can do as well or better

- Considering acquirability and learnability:
  - Metrical phonology (Pearl 2008, 2009, submitted)
    - Framework for testing theories of knowledge representation: using an argument from acquisition
    - Benefits: informing theory and informing acquisition
Knowledge Representation Motivations

- One traditional motivation for proposals of knowledge representation (such as parameters or constraints): The knowledge representation helps explain the constrained variation observed in adult linguistic knowledge across the languages of the world

  **Argument from constrained cross-linguistic variation**

- Another (sometimes implicit) motivation for proposals of knowledge representation: Having this knowledge representation pre-specified allows children to acquire the right generalizations from the data as quickly as they seem to do

  **Argument from acquisition**

- Using computational methods and available empirical data, we can quantify this argument and explicitly test different proposals for knowledge representation

- At the same time, we can explore how acquisition could proceed if children were using these different knowledge representations

  Pearl 2008, 2009, submitted

A generative system of metrical phonology

- Observable data: stress contour

  - Observational data: OCtopus
  
  - Underlying representation/analysis:

  - Using computational methods and available empirical data, we can quantify this argument and explicitly test different proposals for knowledge representation

  - At the same time, we can explore how acquisition could proceed if children were using these different knowledge representations

  Pearl 2008, 2009, submitted
Two Knowledge Representations

- Tractable explorations
  - Parametric system: 5 parameters & 4 sub-parameters (Halle & Vergnaud 1987, Dresher & Kaye 1990, Dresher 1999)
  - Hypothesis space: 156 legal grammars
  - Hypothesis space: 10! grammars (3,628,800)

Comparing Knowledge Representations

- Quantity Sensitivity
- Extrametricality
- Parse, Non-Final
- Align-Left, Align-Right
- Feet Headedness
- Boundedness
- Feet Directionality
- Correct grammar produces compatible contour
- OCtopus
- Best candidate for the correct grammar has a compatible contour

Non-trivial case study: English

- Non-trivial because there are many data that are ambiguous for which parameter value or constraint ranking they implicate
- Non-trivial because there are many irregularities
  - Analysis of child-directed speech (8-15 months) from Brent corpus (Brent & Siskind 2001) from CHILDES (MacWhinney 2000): 904,084 tokens, 7390 types
  - For words with 2 or more syllables:
    - 174 unique syllable-rime type combinations (ex: closed-closed (VC VC))
    - 85 of these 174 have more than one stress contour associated with them (unresolvable): no one grammar can cover all the data
    - Ex for VC VC type: her SELF
    - SOME WHERE

Cognitively inspired learners using parameters

- Learner’s hypothesis space: Set of 156 legal grammars
- Target state = grammar for English (Halle & Vergnaud 1987, Dresher & Kaye 1990, Dresher 1999) derived from cross-linguistic variation and adult linguistic knowledge: quantity sensitive, VC syllables are heavy, rightmost syllable is extrametrical, feet are constructed from the right, feet are 2 syllables, feet are headed on the left

Premise: This is the grammar that best describes the systematic data of English, even if there are exceptions.
Cognitively inspired learners using parameters

**Empirical grounding**
- Learner’s input based on the number of words likely to be heard on average in a 6 month period: 1,666,667. (Akhtar et al. (2004), citing Hart & Risley (1995)).
- Input distributions derived from child-directed speech distributions.
  - Brent corpus (Brent & Siskind 2001): 8 - 15 months
  - Child’s syllabification of words: MRC Psycholinguistics Database (Wilson 1988)
  - Associated stress contour: CALLHOME American English Lexicon (Canavan et al. 1997)

**Cognitively inspired learners using parameters**

**Learner’s algorithm:**
- Incremental update: words are processed one at a time, as they are encountered. (Assumes word segmentation is operational. Jusczyk, Houston, & Newsome (1999) suggests that 7-month-olds can segment some words successfully.)
- Words are divided into syllables, with syllable rime identified as closed (VC), short (V), long (VV), or superlong (VVC). Jusczyk, Goodman, & Baumann (1999) and Turk, Jusczyk, & Gerken (1995) suggest young infants are sensitive to syllables and properties of syllable structure.
- Sub-parameters are not set until the main parameter is set. This is based on the idea that children only consider information about a sub-parameter if they have to.

**Cognitively inspired learners using parameters**

**Learner’s algorithm:**
- Probabilistic generation and testing of parameter value combinations (grammars) (Yang 2002)
- For each parameter, the learner associates a probability with each of the competing parameter values. Initially all values are equiprobable.
  - Ex: Quantity Sensitivity
    - Value 1: Quantity Sensitive (0.5)
    - Value 2: Quantity Insensitive (0.5)
- For each data point, a grammar is probabilistically generated, based on the probabilities associated with each parameter’s values.

**Cognitively inspired learners using parameters**

**Learner’s algorithm:**
- The selected grammar is then used to generate a stress contour, based on the syllable structure of the word.
- If the generated contour matches the observed contour, all participating parameter values are rewarded. If it mismatches, all values are punished.
- Over time (as measured in data points encountered), the probability associated with a parameter value will approach either 1.0 or 0.0, based on rewards and/or punishments. Once the probability is close enough, the learner sets the appropriate parameter value.
Acquirability results: parameters

- Four different implementations of reward/punishment tried (two Naïve Parameter Learner variants that use Linear reward-penalty schemes (Yang 2002) and two incremental Bayesian variants)

- Only one variant (one of the linear reward-penalty ones) was ever successful at converging on the adult English grammar, and then only once every 3000 runs! This seems like very poor performance from these cognitively inspired learners.

Problem with constrained learners?

- Maybe the problem is with the constrained learning algorithms: Are they identifying sub-optimal grammars for the data they encounter?
  - If so, ideal learners should find the optimal grammars that are most compatible with the English child-directed speech data

Premise: The adult English grammar is the grammar that best describes the systematic data of English, even if there are exceptions.

Implication: The adult English grammar is the grammar that is best able to generate the stress contours for the English data (most compatible).

- English grammar compatibility with data:
  - Generates contours matching 73.0% observable data tokens, where every instance of a word is counted (62.1% types, where frequency is factored out and a word is counted only once no matter how often it occurs)
  - Note: not expected to be at 100% because of irregularities in English data

Problem for any parametric learner

- Average compatibility of grammars selected by constrained learners:
  - 73.6% by tokens (63.3% by types)

- The cognitively inspired learners are identifying the more optimal grammars for this data set - it's just that these grammars don't happen to be the adult English grammar!
  - Learnability Implication: The problem isn't because these learners are constrained. Unconstrained learners would have the same problem.
  - English grammar compared to other 155 grammars
    - Rank 52nd by tokens, 56th by types
    - English grammar is barely in the top third - unsurprising that probabilistic learners rarely select this grammar, given the child-directed speech data

Problem for any parametric learner

- Parametric child learner has a learnability problem: can't get to adult target state given the data available to children

But what about a child learner using the OT knowledge representation? Pearl in prep.
OT system test

- 10 constraints (Hammond 1999, Prince & Smolensky 1993, Tesar & Smolensky 2000)
  - Hypothesis space: 10! grammars (3,628,800)

  Weight-To-Stress Principle: VV, VC
  - Parse, Non-Final
  - Align-Left, Align-Right
  - FootBin: syllables, moras
  - Trochaic, iambic

- Adult English grammar (Hammond 1999, Pater 2000):
  - Combination of constraint orderings
    - FootBin, Trochaic, WSP(VV) > Non-Final > Align-Right > Parse > Align-Left
    - Trochaic > iambic
    - Non-Final > WSP(VC)
  - 720 grammars of 3,628,800 follow these orderings (720 ways to be English)

- Compatibility of English OT grammars with child-directed speech data
  - Compatible grammar’s best candidate has a stress contour that matches the observed stress contour for any given data point

<table>
<thead>
<tr>
<th>(OC) to pus</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>oc (TO pus)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(oc TO) pus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Maximum compatibility score for any English grammar:
  - 24.2% of data tokens (26.6% of types)
  - (32 grammars with this score)
  - Maybe we simply can’t find grammars that are much better, given these constraints?

- Maximum compatibility score for any non-English grammar:
  - 74.8% of data tokens (87.5% of types)
  - (1600 grammars with this score)

- The English OT grammars are clearly sub-optimal for this data set - but how do they compare overall to the other grammars in the hypothesis space?

Upshot: The OT system representation doesn’t look much better for learners trying to acquire an adult English grammar from child-directed speech.
### Parameters vs. OT comparison

<table>
<thead>
<tr>
<th></th>
<th>Parameters</th>
<th>OT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammars in hypothesis space</td>
<td>156</td>
<td>3,628,800</td>
</tr>
<tr>
<td>Best grammar compatibility</td>
<td>76.5% (tokens)</td>
<td>74.6% (tokens)</td>
</tr>
<tr>
<td></td>
<td>70.3% (types)</td>
<td>67.5% (types)</td>
</tr>
</tbody>
</table>

- Either knowledge representation contains grammars that are compatible with a reasonable majority of the English child-directed speech data.

### Problem for both learners

- Parametric child learner has a learnability problem: can’t get to adult target state given the data available to children.

- OT child learner has a learnability problem, too (possible an even greater one): can’t get to adult target state given the data available to children, and adult grammar accounts for a much smaller portion of the available data.

<table>
<thead>
<tr>
<th></th>
<th>Parameters</th>
<th>OT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammars in hypothesis space</td>
<td>156</td>
<td>3,628,800</td>
</tr>
<tr>
<td>Best grammar compatibility</td>
<td>76.5% (tokens)</td>
<td>74.6% (tokens)</td>
</tr>
<tr>
<td></td>
<td>70.3% (types)</td>
<td>67.5% (types)</td>
</tr>
</tbody>
</table>

- The ranking in the hypothesis space for the (best) English grammar for either knowledge representation is fairly similar (around the top third of the hypothesis space).

### Parameters vs. OT comparison

<table>
<thead>
<tr>
<th></th>
<th>Parameters</th>
<th>OT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammars in hypothesis space</td>
<td>156</td>
<td>3,628,800</td>
</tr>
<tr>
<td>Best grammar compatibility</td>
<td>76.5% (tokens)</td>
<td>74.6% (tokens)</td>
</tr>
<tr>
<td></td>
<td>70.3% (types)</td>
<td>67.5% (types)</td>
</tr>
<tr>
<td>% of hypothesis space (best) English grammar scores lower than</td>
<td>28.3% (tokens)</td>
<td>31.9% (tokens)</td>
</tr>
<tr>
<td></td>
<td>31.1% (types)</td>
<td>34.8% (types)</td>
</tr>
<tr>
<td>(Best) English grammar compatibility</td>
<td>73.0% (tokens)</td>
<td>24.2% (tokens)</td>
</tr>
<tr>
<td></td>
<td>62.1% (types)</td>
<td>26.6% (types)</td>
</tr>
</tbody>
</table>

- However, the best English grammar compatibility is very low for OT, compared to the English grammar in the parametric system.
Getting out of the learnability problem: 2 options

Option 1: change the target state

- Initial knowledge state of learner
- Child-directed speech
- Adult knowledge (target state)
- Other data

Other target state

Prediction: Children initially select non-English grammars, given these data. If so, we should be able to use experimental methods to observe them using non-English grammars for an extended period of time.

- Kehoe 1998: elicitation task with English 34-month-olds used items that were compatible with the grammars modeled learners often chose here

Option 2: change the initial state

- Initial knowledge state of learner
- Child-directed speech
- Adult knowledge (target state)

A different (enriched) initial state

- Maybe young children have additional boosts
  - Pearl (2008) explores the effects of a bias to only learn from data perceived as unambiguous for a parametric learner, and finds that the learners with this knowledge are successful if parameters are set in certain orders.
  - Required knowledge at the initial state:
    - importance of unambiguous data (and a method for identifying these data for each parameter value)
    - parameter-setting order constraints (and potentially a method for deriving these constraints)
Testing proposals of knowledge representation

- Began by exploring cognitively plausible learners to test theories about knowledge representation (argument from acquisition)

- When they failed at the acquisition task, we asked what the cause of the failure was - due to learners being constrained or due to something about the language acquisition computation?

- Led us to examine learnability considerations, given the data
  - Highlighted learnability issues for probabilistic learners seeking optimal solutions given child-directed speech data

A useful framework: what comes next

- Change knowledge representation
  - Theoretical + computational investigations: perhaps different parameters or constraints make the adult English grammar more acquirable from child-directed speech
  - Different theoretical proposals can be motivated and tested via computational methods

A useful framework: what comes next

- Change premise about trajectory of children’s acquisition
  - Experimental investigations: exploring English children’s initial knowledge states before they have knowledge of morphology and adult lexicon items
  - This then informs future computational investigations and thus any arguments from acquisition for a given theoretical proposal of knowledge representation

A useful framework: what comes next

- Change learner’s initial knowledge state
  - Computational investigations: strategies learners can use to solve acquisition problem as currently defined
  - Describe the required initial knowledge state to make acquisition possible for learners using specific knowledge representations, thereby creating a way to explicitly compare different knowledge representations
  - Knowledge representations requiring a less enriched initial state may be more desirable
Computation in Acquisition: Revisited

Many places where the concept of computation connects with the information-processing task of acquisition
- Understanding the computation in human minds (what, when, how, what from)
- Using computational methods to understand that computation

The End & Thank You!

Special thanks to...
- Sharon Goldwater
- Mark Steyvers
- Ivano Caponigro
- Jon Sprouse
- Diogo Almeida
- Amy Weinberg
- Roger Levy
- the Boston University Conference on Language Development 2007, 2009
- the Psychocomputational Models of Human Language Acquisition Workshop 2009
- the Learning Meets Acquisition Workshop 2009
- the Generative Approaches to Language Acquisition North America 2008
- the Computational Models of Language Learning Seminar at UCI 2008
- UC San Diego Linguistics Department
- UC Irvine Machine Learning Group
- UCLA Linguistics Department
- USC Linguistics Department

This work was supported by NSF grant BCS-0843896 and CORCL grant MI 14B-2009-2010.

Results: Exploring different performance measures

- Some positions in the utterance are more easily segmented by infants, such as the first and last word of the utterance (Seidl & Johnson 2006).
  - The first and last word are less ambiguous (one boundary known) (first, last > whole utterance)
  - Memory effects & prosodic prominence make the last word easier (last > first, whole utterance)
  - The first/last word are more regular, due to syntactic properties (first, last > whole utterance)
Results: Exploring different performance measures

Unigrams vs. Bigrams, Token F-scores

whole utterance
first word
last word

Results averaged over 5 randomly generated test sets (~900 utterances) that were separate from the training sets (~8800 utterances), all generated from the Bernstein/Ratner corpus.

DMCMC Unigram: d=1, s=20000
DMCMC Bigram: d=0.25, s=20000

Unigram Ideal, Unigram DMCMC, Bigram DPS, both DPM learners: improvement on first and last words.

Unigram Ideal, Unigram DMCMC, Bigram DPS, both DPM learners: improvement on first and last words.

Interesting:
Constrained unigram learners outperform the Ideal learner for first and last words.

Bigram Ideal, Unigram DPS, Bigram DMCMC learners: improvement only on first words.
Results: Exploring different performance measures

Unigrams vs. Bigrams, Token F-scores

Interesting:
Constrained unigram learners outperform the Ideal learner for first and last words.
Some constrained bigram learners are equivalent to the unconstrained learner for first and last words.

Search algorithm comparison

Model defines a distribution over hypotheses. We use Gibbs sampling to find a good hypothesis.
- Iterative procedure produces samples from the posterior distribution of hypotheses.

P(h|d)

- Ideal (Standard): A batch algorithm
- vs. DMCMC: incremental algorithm that uses the same sampling equation

Gibbs sampler

- Compares pairs of hypotheses differing by a single word boundary:
  - what's that
  - the doggie
  - yeah
  - where's the doggie

- Calculate the probabilities of the words that differ, given current analysis of all other words in the corpus.
- Sample a hypothesis according to the ratio of probabilities.
The unigram model

Assume word \( w_i \) is generated as follows:

1. Is \( w_i \) a novel lexical item?

\[
P(\text{yes}) = \frac{\alpha}{n + \alpha} \quad \text{Fewer word types = Higher probability}
\]
\[
P(\text{no}) = \frac{n}{n + \alpha}
\]

2. If novel, generate phonemic form \( x_1, \ldots, x_m \):

If not, choose lexical identity of \( w_i \) from previously occurring words:

\[
P(w_i = w) = \frac{n_w}{n} \quad \text{Power law = Higher probability}
\]

Notes

- Distribution over words is a Dirichlet Process (DP) with concentration parameter \( \alpha \) and base distribution \( P_0 \):

\[
P(w_i = w \mid w_1, \ldots, w_{i-1}) = \frac{n_w + \alpha P(w)}{i - 1 + \alpha}
\]

- Also (nearly) equivalent to Anderson’s (1990) Rational Model of Categorization.
**Bigram model**

Assume word $w_i$ is generated as follows:

1. Is $(w_{i-1}, w_i)$ a novel bigram?

   $$P(\text{yes}) = \frac{\beta}{n_{i-1} + \beta} \quad P(\text{no}) = \frac{n_{i-1}}{n_{i-1} + \beta}$$

2. If novel, generate $w_i$ using unigram model (almost).

   If not, choose lexical identity of $w_i$ from words previously occurring after $w_{i-1}$,

   $$P(w_i = w \mid w_{i-1} = w') = \frac{n_{i,w|w'}}{n_{i,w'}}$$

**Notes**

- Bigram model is a hierarchical Dirichlet process (Teh et al., 2005):

  $$P(w_i = w \mid w_{i-1} = w', w_{i-2} = w_{i-1}' \ldots w_1') = \frac{n_{i,w|x} + \beta P(w)}{i - 1 + \beta}$$

  $$P(w_i = w \mid w_{i-1} = w_{i-1}' \ldots w_1 = w_1') = \frac{b_i + \alpha P(w)}{b + \alpha}$$

**Results: The effect of number of samples**

- Even down to 500 samples per utterance, token F score is still above 60. Can still get reasonably high score with fairly few samples.

- Scores somewhat stable after about 4000 utterances trained on.
Results: The effect of number of samples

- Even down to 100 samples per utterance, token F score is still above 60. (Less samples required to get high score.)
- Jump in score occurs quickly, after only 500 utterances trained on.

Results: Standard vs. Decayed MCMC

- Ideal (Standard MCMC) learner continually outperforms DMCMC for lexicon.
- Unigram DMCMC only does well after about 4000 utterances have been trained on.
Unbiased acquirability model update functions

Naïve Parameter Learner (Yang 2002) [NParLearner]: Linear reward-penalty (Bush & Mosteller 1951)

- Learning rate $\gamma$:
  - small: small changes
  - large: large changes

- Parameter values $v_1$ vs. $v_2$
  - $p_i = p_i \cdot (1 - \gamma) p_i$
  - $p_i = 1 - p_i$
  - $p_1 = (1 - \gamma) p_i$

- Reward $r_1$
  - $p_1 = r_1$

Bayesian Learner [BayesLearner]: Bayesian update of binomial distribution (Chew 1971)

- Parameters $\alpha$, $\beta$
  - $\alpha \approx \beta$: initial bias at $p = 0.5$
  - $\alpha, \beta < 1$: initial bias toward endpoints ($p = 0.0, 1.0$)

- Here: $\alpha = \beta = 0.5$

Unambiguous data bias

Pearl (2008): A general class of probabilistic models learning from unambiguous data is guaranteed to succeed at acquiring the English grammar from English child-directed speech, provided the parameters are learned in certain orders.

Why learning from unambiguous data works: The unambiguous data favor the English grammar, so English becomes the optimal grammar.

However, they make up a small percentage of the available data (never more than 5%) so their effect can be washed away in the wake of ambiguous data if the ambiguous data are learned from as well and the parameters are not learned in an appropriate order.
<table>
<thead>
<tr>
<th>Is it just that children need more lexicon items?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis of adult-directed conversational speech</td>
</tr>
<tr>
<td>CALLFRIEND corpus (Canavan &amp; Zipperlen 1996), North American English portion: recorded telephone conversations between adults</td>
</tr>
<tr>
<td>82,487 word tokens, 4,417 word types</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parametric English grammar (somewhat better but not the best):</th>
</tr>
</thead>
<tbody>
<tr>
<td>63.7% token compatibility, 52.1% type compatibility</td>
</tr>
<tr>
<td>ranked 34th by tokens, 36th by types</td>
</tr>
<tr>
<td>Interesting: Best grammar in hypothesis space differs only by one parameter value (QI instead of English’s QS): 66.6% token compatibility, 56.3% type compatibility</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parametric English grammar is not the best for adult conversational speech either</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential explanation: linguists use items that appear infrequently in conversations when making their theories, under the assumption that these items are part of the adult knowledge state</td>
</tr>
</tbody>
</table>

Worth testing experimentally: the English adult knowledge state (do adults make the generalizations that linguists think they do, or are some of the crucial items exceptions that adults do not include in their generative system?)