Evaluation, use, and refinement of knowledge representations through acquisition modeling

Lisa Pearl
University of California, Irvine

February 21, 2015:
Learning in Generative Grammar: 50 years since the evaluation metric
GALANA: University of Maryland, College Park

Premise
The knowledge representation provided by Universal Grammar is what makes acquisition happen so fast and so well.

Knowledge representation
Universal Grammar

Acquisition

Why?

Knowledge representation
Universal Grammar

Premise
The knowledge representation provided by Universal Grammar is what makes acquisition happen so fast and so well.

Knowledge representation
Universal Grammar

What hypotheses are available and how are they defined?

Grammar = ...
...set of parameters
...set of violable constraints
...rules over phrasal nodes

Competing theories
So how do we choose when we have multiple theories about how knowledge is represented?

Ex: Parameters vs. violable constraints

Ex: Different implementations of dependencies

Competing theories
One answer: Use each one for acquisition. Does that knowledge representation make acquisition possible from the available data?

Is the hypothesis space helpfully constrained by the representation?

Well...how do we tell?

Is the acquisitional intake defined by the hypothesis space sufficient to get the job done?

Well...how do we tell?
A helpful tool: Computational modeling

We can computationally model a learner who incorporates the assumptions of a representation, set that learner up in a cognitively plausible learning scenario, and see if acquisition succeeds.

Is the acquisitional intake defined by the hypothesis space sufficient to get the job done?

Well...how do we tell?

A helpful tool: Computational modeling

We can computationally model a learner who incorporates the assumptions of a representation, set that learner up in a cognitively plausible learning scenario, and see if acquisition succeeds.

What we learn from computational modeling:
- Which representations allow acquisition to succeed
- What needs to be true about the learning scenario for a learner to succeed using those representations

The goal

This computational modeling feedback helps us refine our theories about both the knowledge representation and the acquisition process that uses that representation.

Today’s goal:
Computational acquisition modeling

Case studies
Metrical phonology

Syntax

What did you see?

Today’s goal:
Computational acquisition modeling

Case studies
Metrical phonology

Syntax

What did you see?

Metrical phonology: Target knowledge

Account for word-level stress patterns

Observable data: stress contour
OCTopus

Underlying representation?

Points of agreement:
- Use metrical feet:
  - Units ≥ syllables
  - but (often) smaller than words
  - (VC V VC)
  - ok a is
  - ok ta pas
  - oc to pus

- Look only at syllable rimes
- Divide word into syllables

Pearl, Ho, & Detrano 2014, under rev.
Metrical phonology: Target knowledge

Account for word-level stress patterns

Observable data: stress contour

Underlying representation?

Points of cross-linguistic variation:

- How to classify syllables
- What metrical feet are allowed
- How stress interacts with metrical feet

Points of disagreement:

Underlying grammar = ....?

Parameters with values set

Ranked violable constraints

- How stress interacts with metrical feet

Metrical phonology: Target knowledge

Observable data: stress contour

Underlying representation?

Points of cross-linguistic variation:

- How to classify syllables
- What metrical feet are allowed
- How stress interacts with metrical feet

Points of disagreement:

Underlying grammar = ....?

Parameters with values set

Ranked violable constraints

Metrical phonology: Target knowledge

Observable data: stress contour

Underlying representation?

Points of cross-linguistic variation:

- How to classify syllables
- What metrical feet are allowed
- How stress interacts with metrical feet

Points of disagreement:

Underlying grammar = ....?

Parameters with values set

Ranked violable constraints

Metrical phonology: Target knowledge

Observable data: stress contour

Underlying representation?

Points of cross-linguistic variation:

- How to classify syllables
- What metrical feet are allowed
- How stress interacts with metrical feet

Points of disagreement:

Underlying grammar = ....?

Parameters with values set

Ranked violable constraints
Hayes, Hayes 1995
8 parameters
Hypothesis space: 768 grammars
Parameter values used:
Bot, Em-RtCons, VC-H, PtDr-Rt, PL-Strong, MorTro, DF-Strong, WLER-Rt
...which are the values of the English grammar.

Premise: Many different candidates for a word's stress representation and contour are generated and then ranked according to which constraints are violated. Violating higher-ranked constraints is worse than violating lower-ranked constraints.

This means the "grammar" for a language is often a set of the possible rankings (grammars) that obey those orderings.

Official grammars for languages are often described as partial orderings of constraints.

Correct grammar builds compatible contour

This grammar, comprised of particular parameter values, generates an incorrect stress contour.

9 violable constraints

9 violable constraints

Grammar = ranked ordering of all constraints

Best candidate for the correct grammar has a compatible contour

Best candidate for the correct grammar has a compatible contour

Ex: The English "grammar" is compatible with 26 rankings.
Metrical phonology: Three knowledge representations

Constraint-ranking systems

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

- Nonfinality, Parse-a
- Foot binarity
- Trochaic
- Weight-to-Stress
- Align left, Align right
- *Sonorant nucleus

Best candidate for the correct grammar has a compatible contour

\\

Metrical phonology: Three knowledge representations

Constraint-ranking systems

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

- Nonfinality, Parse-a
- Foot binarity
- Trochaic
- Weight-to-Stress
- Align left, Align right
- *Sonorant nucleus

Only one candidate left, and it has a compatible contour.

Sample candidates

A sample grammar that is a version of the English “grammar”:

Best candidate for the correct grammar has a compatible contour

\\

Knowledge representation comparison

**HV:** 5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

**Hayes:** 8 parameters
Hypothesis space: 768 grammars

\\

Metrical phonology: Acquisitional intake

Acquisition goal: Identify the grammar that can account for the word-level stress patterns in the language

Observable data: stress contour

All representations: use metrical feet based on syllable rimes

---

(Images and diagrams are not transcribed in the natural text format.)
Metrical phonology: Acquisitional intake

Acquisition goal: Identify the grammar that can account for the word-level stress patterns in the language

Observable data: stress contour

All representations: use metrical feet based on syllable rimes

Parametric inference: Does this set any values?

Pearl, Ho, & Detrano 2014, under rev.

Metrical phonology: Acquisitional intake

Acquisition goal: Identify the grammar that can account for the word-level stress patterns in the language

Observable data: stress contour

All representations: use metrical feet based on syllable rimes

Parametric inference: Does this set any values?

OT inference: Does this implicate any constraint rankings?

OT: 9 violable constraints
Hypothesis space: 162,880 grammars
(English = 26 grammars)

Hayes: 8 parameters
Hypothesis space: 768 grammars

Note: These values/rankings are derived from stress patterns for English words in the adult lexicon.

Computational acquisition evaluation: English

English grammar

HVI: 5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

Hayes: 8 parameters
Hypothesis space: 768 grammars

Learning English metrical phonology: Non-trivial

Non-trivial because there are many data that are ambiguous for which parameter value or constraint ranking they implicate

This is generally a problem for acquisition (poverty of the stimulus = the data are compatible with many hypotheses).

Learning English metrical phonology: Non-trivial

Non-trivial because there are many irregularities. This is less common for acquisition – usually there aren’t a lot of exceptions to the system being acquired.
Learning English metrical phonology: Non-trivial

Non-trivial because there are many irregularities. This is less common for acquisition – usually there aren’t a lot of exceptions to the system being acquired.

Some causes of irregularity:

- **Interactions with morphology** (Chomsky & Halle 1968, Hayes 1982, Kiparsky 1979)

Example: Adding productive morphology doesn’t change the stress pattern, even though all grammars base their stress patterns on the syllables present in the word.

<table>
<thead>
<tr>
<th></th>
<th>PREtty</th>
<th>senSAtion</th>
<th>senSAtional</th>
</tr>
</thead>
<tbody>
<tr>
<td>EARly</td>
<td>PREttier</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Learning English metrical phonology: Non-trivial

These irregularities can cause multiple stress contours to be associated with a syllabic word form. This is problematic for the grammars in these knowledge representations...

```
Syllabic word form: V VV

KI tty  V vv  a WAY  UH OH

Generate one of these...

Syllable weight
Foot headedness
Extrametricality
Foot directionality
Stress analysis direction

```

Upshot of multiple stress contours: No one grammar can account for all the stressed words in the input.

But how big of a problem is this in English child-directed speech?

```
Syllabic word form: V VV

KI tty  V vv  a WAY  UH OH

Select one of these...

```

Learning English metrical phonology: Non-trivial

Some causes of irregularity:


Stress contours may be different across grammatical categories, even though the syllabic word form doesn’t change.

<table>
<thead>
<tr>
<th>NOUNS</th>
<th>VERBS</th>
<th>Syllabic word form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONduct</td>
<td>conDUCT</td>
<td>VC VCC</td>
</tr>
<tr>
<td>DEsert</td>
<td>deSERT</td>
<td>V VCC</td>
</tr>
<tr>
<td>SUSpect</td>
<td>suSPECT</td>
<td>V VCC</td>
</tr>
</tbody>
</table>

Learning English metrical phonology: Non-trivial

These irregularities can cause multiple stress contours to be associated with a syllabic word form. This is problematic for the grammars in these knowledge representations, since a grammar can only generate a single stress contour per syllabic word form...

```
Syllabic word form: V VV

KI tty  V vv  a WAY  UH OH

Select one of these...

```

Learning English metrical phonology: Non-trivial

Non-trivial because there are many irregularities. This is less common for acquisition – usually there aren’t a lot of exceptions to the system being acquired.
Learning English metrical phonology: Non-trivial

Analysis of Brent corpus (CHILDES database): 4780 word types (99,968 tokens) of American English speech directed at children between the ages of 6 and 12 months

<table>
<thead>
<tr>
<th>Syllabic word form</th>
<th>V VV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitty</td>
<td>V VV</td>
</tr>
<tr>
<td>VV</td>
<td>V VV</td>
</tr>
</tbody>
</table>

Multiple stress contours

HV: 73 of 123 syllabic word forms
Hayes: 86 of 149 syllabic word forms
OT: 166 of 452 syllabic word forms

This occurs a lot!

Learning English metrical phonology: Target state update

Acquisition success: Identify the grammar that can account for the word-level stress patterns in the language

Is this reasonable?

Probably.

A grammar is useful because it provides a compact representation of some aspect of the data. Even if it doesn’t cover all the data, covering some is helpful.

Learning English metrical phonology: Acquisition evaluation

How easily does a knowledge representation allow children to learn their specific language’s grammar, when given realistic data?

Learnability analysis provides a quantitative way to compare competing knowledge representations (Pearl 2011, Legate & Yang 2012)

Working premise: Rational learners

A learner trying to learn which grammar is the right one for the language will choose the grammar perceived to be the best.

Quantifying learnability

Once we define the acquisitional intake, we can then ask which grammar in the hypothesis space defined by the knowledge representation is best, assuming a rational learner that will choose the grammar compatible with the most data.
Quantifying learnability

Once we define the **acquisitional intake**, we can then ask which grammar in the hypothesis space defined by the knowledge representation is **best**, assuming a rational learner that will choose the grammar **compatible with the most data**.

**Compatibility** with a data point: A grammar is compatible with a data point if the grammar can account for that data point.

Here: Matching stress contour.

Intuition: More compatibility is better.
A grammar that can account for 70% of the data is better than a grammar that can only account for 55% of the data.

Example:
A grammar that can account for 70% of the data has a raw compatibility of 0.70.

**Learnability potential** for a knowledge representation: The amount of data the best grammar is compatible with. This is how much of the data that knowledge representation is capable of accounting for with any of its grammars.

Example: If the best grammar can account for 70% of the data, this knowledge representation has a learnability potential of 0.70.


---

**English learnability: Knowledge representations**

So what’s the **best** any grammar in a given knowledge representation actually does, given realistic child-directed data?

**Learnability potential** = proportion of data the best grammar (relative compatibility = 1.00) can account for

**Raw compatibility of best grammar**

- HV: 0.668 types
- Hayes: 0.683 types
- OT: 0.657 types

Implication:
The best grammar in any of these knowledge representations is pretty useful to have. It allows a learner to account for a good proportion of the input, even if there’s a significant chunk that can’t be accounted for.

**Raw compatibility of best grammar**

- HV: 0.668 types
- Hayes: 0.683 types
- OT: 0.657 types
Implication:
The best grammar in any of these knowledge representations is pretty useful to have. It allows a learner to account for a good proportion of the input, even if there’s a significant chunk that can’t be accounted for.

But...is that really the best they can do?

Data filters

Updated working assumption: The learner will try to learn a grammar that can account for all the productive data encountered (Legate & Yang 2012).

Acquisitional intake = only productive data

(Liti & Gagliardi 2015)

Productive data filter for metrical phonology

Updated working assumption: The learner will try to learn a grammar that can account for all the productive data encountered (Legate & Yang 2012).

Syllabic word form:
V VV

Productive data filter

Principled way to implement this = Tolerance Principle

English learnability: Knowledge representations

So what’s the best any grammar in a given knowledge representation actually does, given realistic child-directed data and a productive data filter?

Learnability potential = proportion of data the best grammar (relative compatibility = 1.00) can account for

Raw compatibility of best grammar
HV: 0.949 productive types
Hayes: 0.933 productive types
OT: 0.843 productive types

84-95% of the productive word types

94-95% of the productive word types

Raw compatibility of best grammar
HV: 0.949 productive types
Hayes: 0.933 productive types
OT: 0.843 productive types

84-95% of the productive word types

acquisitional intake

acquisitional intake
Relative compatibility of English grammar

HV: 0.798 out of 362,880 grammars
Hayes: 0.817 out of 362,880 grammars
OT: 0.676 out of 768 grammars

Better than many...but many are still better
HV: 51 are better
Hayes: 249 are better
OT: 66,407 are better

Implication:

There are still many other grammars in the hypothesis space that are more compatible with the data. It would be easier to pick one of these other more compatible grammars.

Relative compatibility of English grammar

HV: 0.673 out of 156 grammars
Hayes: 0.676 out of 768 grammars
OT: 0.817 out of 362,880 grammars

Better than many...but many are still better
HV: 51 are better
Hayes: 249 are better
OT: 66,407 are better

Implication:

There are still many other grammars in the hypothesis space that are more compatible with the data. It would be easier to pick one of these other more compatible grammars.

Relative compatibility of English grammar

HV: 0.622 out of 156 grammars
Hayes: 0.680 out of 768 grammars
OT: 0.798 out of 362,880 grammars

Little if any improvement...and sometimes worse:
HV: 59 are better
Hayes: 246 are better
OT: 73,302 are better

Implication:

There are still many other grammars in the hypothesis space that are more compatible with the data. It would be easier to pick one of these other more compatible grammars.

Relative compatibility of English grammar

HV: 0.673 out of 156 grammars
Hayes: 0.676 out of 768 grammars
OT: 0.817 out of 362,880 grammars

Better than many...but many are still better
HV: 51 are better
Hayes: 249 are better
OT: 66,407 are better

Implication:

There are still many other grammars in the hypothesis space that are more compatible with the data. It would be easier to pick one of these other more compatible grammars.

Relative compatibility of the English grammar = proportion of grammars in the hypothesis space that is the best English grammar.

What about if children have a productive data filter on their acquisitional intake?

Relative compatibility of English grammar

HV: 0.673 out of 156 grammars
Hayes: 0.676 out of 768 grammars
OT: 0.817 out of 362,880 grammars

Better than many...but many are still better
HV: 51 are better
Hayes: 249 are better
OT: 66,407 are better
What values/constraint rankings do the grammars use that are more compatible with the data than the official English grammar?

Parametric: HV

It turns out that many high compatibility grammars use a different Quantity Sensitivity value: Quantity Insensitive (QI), rather than Quantity Sensitive (QS).

So what happens if we swap the English definition's quantity sensitivity value?

QS → QI

Relative compatibility over all data = 0.94!

But relative compatibility over productive data = 0.71...

Upshot: For the HV knowledge representation, the learning problem could be ameliorated by simply switching one parameter value as long as children aren't using a productive data filter.
What values/constraint rankings do the grammars use that are more compatible with the data than the official English grammar?

It turns out that many high compatibility grammars use a different Foot Inventory value: Syllabic Trochees (Syl-Tro) rather than Moraic Trochees (Mor-Tro).

This allows them to handle words like baby and kitty, which have a final unstressed heavy syllable.

For the Hayes knowledge representation, the learning problem could be ameliorated by simply switching one parameter value especially if children are using a productive data filter.
What values/constraint rankings do the grammars use that are more compatible with the data than the official English grammar?

It turns out that all high compatibility grammars use a different ordering of Non-Finality (Non-Fin) and Weight-to-Stress VV (WSP-VV): Non-Fin is ranked higher than WSP-VV.

So what happens if we swap the English definition’s ordering of these constraints?

Non-Fin >> WSP-VV

Relative compatibility over all data = 0.99!
However, the current English grammar definitions in each representation are especially if children are using a productive data filter.

All three knowledge representations are useful for acquisition which is much more easily learnable.

But each representation has a grammar that is very close to the current English grammar definition (change one parameter value or one constraint ordering) which is much more easily learnable.

Metrical phonology learnability:

Take away

Open questions:

Theory of acquisition:
Are children using a productive data filter? This affects how much data any representation can account for and which particular representation has an English-like grammar that is easily learnable from realistic child-directed English data.

Theory of representation for English:
Are these English-like grammars the ones children have? Are these the ones adults have? Can verify experimentally.

Are these English-like grammars (more?) compatible with adult-directed English data? Can verify computationally.

If so, this supports these grammars as the actual English grammar in each representation.

Metrical phonology: Big picture

This approach allows us to evaluate metrical phonology representations by using them for acquisition. We can then refine our theories of acquisition and representation.

Today’s goal:
Computational acquisition modeling

Case studies

Metrical phonology

Syntax

Wh...
Syntax: Syntactic islands

- **Why?** Central to UG-based syntactic theories.
- **What?** Dependencies can exist between two non-adjacent items. They do not appear to be constrained by length (Chomsky 1965, Ross 1967), but rather by whether the dependency crosses certain structures (called “syntactic islands”).

*What does Jack think ___?*
*What does Jack think that Lily said that Sarah heard that Jareth believed ___?*

Some example islands

- Complex NP island:  
  *What did you make [the claim that Jack bought ___]?*  
  - Subject island:  
    *What do you think [the joke about ___] offended Jack?*
  - Whether island:  
    *What do you wonder [whether Jack bought ___]?*  
  - Adjunct island:  
    *What do you worry [if Jack buys ___]?*

Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

*What does Jack think ___?  
What does Jack think that Lily said that Sarah heard that Jareth believed ___?*

Complex NP island:
*What did you make [the claim that Jack bought ___]?  
What do you think [the joke about ___] offended Jack?  
What do you wonder [whether Jack bought ___]?  
What do you worry [if Jack buys ___]?*

Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012) collected magnitude estimation judgments for four different islands, using a factorial definition that controlled for two salient properties of island-crossing dependencies:

- **length** of dependency (matrix vs. embedded)
- **presence of an island** structure (non-island vs. island)

**Subject islands**

<table>
<thead>
<tr>
<th>Who ___ claimed that Lily forgot the necklace?</th>
<th>matrix</th>
<th>non-island</th>
</tr>
</thead>
<tbody>
<tr>
<td>What did the teacher claim that Lily forgot ___?</td>
<td>embedded</td>
<td>non-island</td>
</tr>
<tr>
<td>Who ___ made the claim that Lily forgot the necklace?</td>
<td>matrix</td>
<td>island</td>
</tr>
<tr>
<td>*What did the teacher make the claim that Lily forgot ___?</td>
<td>embedded</td>
<td>island</td>
</tr>
</tbody>
</table>

**Complex NP islands**

<table>
<thead>
<tr>
<th>Who ___ thinks the necklace is expensive?</th>
<th>matrix</th>
<th>non-island</th>
</tr>
</thead>
<tbody>
<tr>
<td>What does Jack think ___ is expensive?</td>
<td>embedded</td>
<td>non-island</td>
</tr>
<tr>
<td>Who ___ thinks the necklace for Lily is expensive?</td>
<td>matrix</td>
<td>island</td>
</tr>
<tr>
<td>*Who does Jack think the necklace for ___ is expensive?</td>
<td>embedded</td>
<td>island</td>
</tr>
</tbody>
</table>

Some example islands

- Complex NP island:  
  *What did you make [the claim that Jack bought ___]?*  
  - Subject island:  
    *What do you think [the joke about ___] offended Jack?*
  - Whether island:  
    *What do you wonder [whether Jack bought ___]?*  
  - Adjunct island:  
    *What do you worry [if Jack buys ___]?*
Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012) collected magnitude estimation judgments for four different islands, using a factorial definition that controlled for two salient properties of island-crossing dependencies:
- length of dependency (matrix vs. embedded)
- presence of an island structure (non-island vs. island)

Whether islands

Who __ thinks that Jack stole the necklace? matrix | non-island
What does the teacher think that Jack stole __ ? matrix | non-island
Who __ wonders whether Jack stole the necklace? matrix | island
*What does the teacher wonder whether Jack stole __ ? embedded | island

Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012) collected magnitude estimation judgments for four different islands, using a factorial definition that controlled for two salient properties of island-crossing dependencies:
- length of dependency (matrix vs. embedded)
- presence of an island structure (non-island vs. island)

Adjunct islands

Who __ thinks that Lily forgot the necklace? matrix | non-island
What does the teacher think that Lily forgot __ ? embedded | non-island
Who __ worries if Lily forgot the necklace? matrix | island
*What does the teacher worry if Lily forgot __ ? embedded | island

Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

(From Sprouse et al. 2012 data on the four island types, with 173 subjects)

Superadditivity: present for all islands tested = Knowledge that dependencies cannot cross these island structures is part of adult knowledge about syntactic islands

Importance for acquisition: This is one kind of target behavior that we’d like a learner to produce.

Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

Bounding nodes are language-specific.
(CP, IP, and/or NP – must learn which ones are relevant for language)

Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

Subjacency--ish (Pearl & Sprouse 2013a, 2013b)
(2) A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

Container node: phrase structure node that contains dependency

What do you like in this picture?
Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

\[
\text{Wh} \quad \begin{array}{c}
\text{[BN1} \\
\vdots \\
\text{BN2]} \\
\end{array}
\]

Subjacency-ish (Pearl & Sprouse 2013a, 2013b)
(2) A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

\[
\text{Wh} \quad \begin{array}{c}
\text{[CN1} \\
\vdots \\
\text{CN5]} \\
\end{array}
\]

Low probability regions are language-specific (defined by sequences of container nodes that must be learned)

Focus on evaluating this one today

Subjacency-ish: Target knowledge

Can the grammatical dependencies be distinguished from the ungrammatical ones?

Sprouse et al. (2012) stimuli:

**Complex NP islands**

- `begin-IP-end`
- `begin-IP-VP-CP-NP-IP-VP-end`

**Subject islands**

- `begin-IP-end`
- `begin-IP-VP-CP-NP-IP-VP-end`

All the ungrammatical dependencies are distinct from all the grammatical dependencies for these syntactic islands.
Can the grammatical dependencies be distinguished from the ungrammatical ones?

Subjacency-ish: Target knowledge

Sprouse et al. (2012) stimuli:

### Complex NP islands
- begin-IP-end
- begin-IP-VP-CP-IP-NP-CP
- begin-IP-VP-CP-IP-VP
- begin-IP-VP-CP-IP-NP-PP

### Subject islands
- matrix | non-island begin-IP-end
- embedded | non-island begin-IP-VP-CP-IP-VP
- embedded | island begin-IP-VP-CP-IP-VP
- embedded | island *begin-IP-VP-CP-IP-VP

Uh oh - the ungrammatical dependencies look identical to some of the grammatical dependencies for these syntactic islands.

Subjacency-ish: Dependency representation

One solution:
Have CP container nodes be more specified for the learner:
Use the lexical head to subcategorize the CP container node.

CP\_\_\_\_\_CP\_\_\_\_CP\_\_\_\_CP\_\_\_\_CP etc.

The learner can then distinguish between these structures:

- IP-VP-CP\_\_\_\_\_IP-VP
- IP-VP-CP\_\_\_\_\_IP-VP

Subjacency-ish: Target knowledge

Can the grammatical dependencies be distinguished from the ungrammatical ones?

Sprouse et al. (2012) stimuli:

### Whether islands
- begin-IP-end
- begin-IP-VP-CP-IP-VP-end
- begin-IP-VP-CP-end
- begin-IP-VP-CP-IP-NP-PP-end

### Adjunct islands
- matrix | non-island begin-IP-end
- embedded | non-island begin-IP-VP-CP-IP-VP-end
- embedded | island begin-IP-VP-CP-IP-VP
- embedded | island *begin-IP-VP-CP-IP-VP

Now the ungrammatical dependencies are distinct from all the grammatical dependencies for these syntactic islands, too.

Subjacency-ish: Acquisitional intake

Children must learn which local pieces of structure are low probability for a wh-dependency. They learn this from the wh-dependencies in their intake, which are defined over the container nodes of the wh-dependency.

\[
\begin{align*}
\text{[wh did [IP VP [IP VP think [IP CP she [IP VP the gift [IP VP was [IP PP from [IP PP]]]]]]]?} \\
\text{Encoding of dependency: begin-IP-VP-CP-IP-VP-PP-end}
\end{align*}
\]

Subjacency-ish: Realistic acquisitional intake

Child-directed speech (Brown-Adam, Brown-Eve, Suppes, Valian) from CHILDES: 101,838 utterances containing 20,923 wh-dependencies

What kind of dependencies are present?

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Dependency Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.7%</td>
<td>begin-IP-VP-end</td>
</tr>
<tr>
<td>12.8%</td>
<td>begin-IP-end</td>
</tr>
<tr>
<td>5.6%</td>
<td>begin-IP-VP-IP-VP-end</td>
</tr>
<tr>
<td>2.5%</td>
<td>begin-IP-VP-PP-end</td>
</tr>
<tr>
<td>1.1%</td>
<td>begin-IP-VP-CP-IP-VP-end</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Pearl & Sprouse 2013a, 2013b, under review
Subjacency-ish: Modeling acquisition

Because wh-dependencies are perceived as sequences of container nodes, local pieces of dependency structure can be characterized by container node trigrams.

[Diagram showing dependency structure and container node trigrams]

Assumption: learning period for modeled learners is 3 years (ex: between 2 and 5 years old for modeling children’s acquisition of islands), so they would hear one million utterances.

Total learning period: 200,000 wh-dependency data points (wh-dependencies make up approximately 20% of the input)
Subjacency-ish: Success metric

For each set of island stimuli from Sprouse et al. (2012), we generate grammaticality preferences for the modeled learner based on the dependency's probability.

We can then plot the log probability of the dependency on the y-axis of the interaction plot.

Non-parallel lines indicate superadditivity, which indicates knowledge of islands.

Subjacency-ish: Take away

Representation validation
If dependencies are represented as container node sequences, acquisition works well for these four syntactic islands. The learner can leverage probabilities of container node trigrams.

Subjacency-ish vs. Subjacency: What’s in UG?

UG = innate + domain-specific

Subjacency

Wh – \{Wh, Wh, Wh, Wh, Wh, Wh, Wh, Wh, Wh\}

Attend to container nodes & subcategorize by CP
Low probability items are dispreferred

Adjunct

Wh – \{Wh, Wh, Wh, Wh, Wh, Wh, Wh, Wh, Wh\}

Attend to bounding nodes (BNs)
Dependencies crossing 2+ BNs are not allowed

Computational acquisition modeling:
Big picture

Informing theories of representation & acquisition

Metrical phonology:

- Can identify learning assumptions (like productive data filters) that benefit children using different knowledge representations
- Can identify language-specific grammars within these representations that are easier to learn from realistic data than the current versions
Computational acquisition modeling: Big picture

Informing theories of representation & acquisition

Syntax

❖ Can validate representations that make it easy to learn syntactic islands, and provide alternative proposals for what's in UG
❖ Can provide concrete demonstrations of learning strategies using these representations that succeed on realistic input data

What did you see?

Three knowledge representations

Parametric systems

HV: Halle & Vergnaud 1987, Dresher 1999, Pearl 2011
5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

Correct grammar builds compatible contour

Grammar = Set of parameter & sub-parameter values

Parametric systems

HV: Halle & Vergnaud 1987, Dresher 1999, Pearl 2011
5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

Correct grammar builds compatible contour

Three knowledge representations

Thank you!

Jon Sprouse  Tim Ho  Zephyr Detrano

Pranav Anand  Misha Becker  Bob Berwick  Adrian Brasoveanu
Alex Clark  Sandy Chung  Bob Frank  Norbert Hornstein
Joanna Lee  Jeff Lidz  Jim McCloskey  Armin Mester
Joseph Nunn  Colin Phillips  William Sakas  Virginia Valian
Matt Wagers  Charles Yang

GALANA selection committee

Audiences at:
Berkeley Linguistics Society Annual Meeting 2014
UC Santa Cruz Linguistics colloquium 2014
Logic & Philosophy of Science 2013 colloquium, UC Irvine
Institute for Mathematical Behavioral Sciences 2013, UC Irvine
Johns Hopkins University Cognitive Science colloquium 2013
New York University Linguistics colloquium 2012
UMaryland Mayfest 2012
Input & Syntactic Acquisition Workshop 2012

Extra material

This work was supported in part by NSF grants BCS-0843896 and BCS-1347028.
Three knowledge representations

Parametric systems
HV: Halle & Vergnaud 1987, Dresher 1999, Pearl 2011
5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

Three knowledge representations

Parametric systems
HV: Halle & Vergnaud 1987, Dresher 1999, Pearl 2011
5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

Quantity sensitivity
Are syllables all identical, or are they differentiated by syllable weight (into Heavy and Light syllables)?

OCtopus
Correct grammar builds compatible contour

Foot directionality
Are feet constructed from the left or from the right?

Boundedness
How big are metrical feet?

Foot headedness
Which syllable in a foot is stressed?

Parameter values used:
Quantity sensitive, VC syllables = Heavy, Extrametricality on rightmost syllable, Feet built from the right, Foot = 2 syllables, Leftmost syllable in foot stressed
Three knowledge representations

**Parametric systems**

- Hayes: Hayes 1995
  - 8 parameters
  - Hypothesis space: 768 grammars

**Correct grammar builds compatible contour**

- OCtopus

**Parameter values used:**

- QS-VC-H, Em-Rt, FtDir-Rt, B-2-Syl, FtHd-Left

...which are the values of the English grammar.

---

Three knowledge representations

**Parametric systems**

- Hayes: Hayes 1995
  - 8 parameters
  - Hypothesis space: 768 grammars

**Correct grammar builds compatible contour**

- OCtopus

- Stress analysis direction
- Word layer end rule
- Degenerate feet
- Extrametricality
- Syllable weight
- Foot directionality
- Parsing locality

---

Three knowledge representations

**Parametric systems**

- Hayes: Hayes 1995
  - 8 parameters
  - Hypothesis space: 768 grammars

**Correct grammar builds compatible contour**

- OCtopus

- Extrametricality
- Syllable weight
- Foot directionality
- Parsing locality

---

Three knowledge representations

**Parametric systems**

- Hayes: Hayes 1995
  - 8 parameters
  - Hypothesis space: 768 grammars

**Correct grammar builds compatible contour**

- OCtopus

- Extrametricality

Are syllables on the edge (or parts of syllables) excluded from metrical feet?

---

Three knowledge representations

**Parametric systems**

- Hayes: Hayes 1995
  - 8 parameters
  - Hypothesis space: 768 grammars

**Correct grammar builds compatible contour**

- OCtopus

- Syllable weight

Syllables are distinguished into Heavy and Light. Are syllables ending in VC (like oc) Heavy or Light?
Three knowledge representations

Parametric systems

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Correct grammar builds compatible contour

OCtopus

Foot directionality
Are metrical feet constructed from the left or the right?

Three knowledge representations

Parametric systems

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Correct grammar builds compatible contour

OCtopus

Foot inventory
How big are metrical feet?
Where does the stress fall within them?

Three knowledge representations

Parametric systems

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Correct grammar builds compatible contour

OCtopus

Word layer end rule
Where does word-level stress go if there are multiple stressed syllables? Can leftover Light syllables have word-level stress?

Three knowledge representations

Parametric systems

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Correct grammar builds compatible contour

OCtopus

Parameter values used:
Bottom-up, Extrametricality on rightmost consonant, VC syllables = Heavy, Feet built from the right, Light syllables not skipped in between feet, Foot = Moraic trochee (2 moras with stress on leftmost), Single light edge syllables not allowed to have stress, Rightmost syllable gets main stress

Three knowledge representations

Parametric systems

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Correct grammar builds compatible contour

OCtopus

Parsing locality
Is one Light syllable skipped between metrical feet?
Three knowledge representations

**Parametric systems**

Hayes: Hayes 1995
8 parameters
Hypothesis space: 768 grammars

Parameter values used:
Bot, Em-RtCons, VC-H, PtDe-Rt,
PL-Strong, MorTro, DF-Strong, WLER-Rt

...which are the values of the English grammar.

Three knowledge representations

**Constraint-ranking systems**

9 violable constraints

Premise: Many different candidates for a word's stress representation and contour are generated and then ranked according to which constraints are violated. Violating higher-ranked constraints is worse than violating lower-ranked constraints.

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(OC to) pus</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(oc to) pus</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
</tr>
</tbody>
</table>

Best candidate for the correct grammar has a compatible contour

This means the "grammar" for a language is often a set of the possible rankings (grammars) that obey those orderings.

Ex: The English "grammar" is compatible with 26 rankings.
Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Best candidate for the correct grammar has a compatible contour

Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Best candidate for the correct grammar has a compatible contour

Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Three knowledge representations
Constraint-ranking systems
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Best candidate for the correct grammar has a compatible contour
Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

<table>
<thead>
<tr>
<th>Weight-to-Stress (VC)</th>
<th>Should all VC syllables be stressed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>(oc TO (pus))</td>
</tr>
<tr>
<td>✓</td>
<td>(OC to (pus))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Align left</th>
<th>Should metrical feet include the leftmost syllable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>(OC to (pus))</td>
</tr>
<tr>
<td>✓</td>
<td>(oc TO (pus))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sonorant nucleus</th>
<th>Should syllables not have sonorants (m, n, p, t, l, r) as the nucleus?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>(your SELF)</td>
</tr>
<tr>
<td>✓</td>
<td>(yr SELF)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonfinality</th>
<th>Parse-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

<table>
<thead>
<tr>
<th>Nonfinality</th>
<th>Parse-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

<table>
<thead>
<tr>
<th>Nonfinality</th>
<th>Parse-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

<table>
<thead>
<tr>
<th>Nonfinality</th>
<th>Parse-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

<table>
<thead>
<tr>
<th>Nonfinality</th>
<th>Parse-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Next important: VV syllables are stressed.

Sample candidates

- (OC) to pus
- (OC) to (PUS)
- (OC) to (PUS)
- (OC) to (PUS)

A sample grammar that is a version of the English "grammar":

Best candidate for the correct grammar has a compatible contour

Three knowledge representations

**Constraint-ranking systems**

**OT:** Hammond 1999, Pater 2000, Tesar & Smolensky 2000
9 violable constraints
Hypothesis space: 9! rankings = 362,880 grammars
Principle (Rooting): All words must have stress

Next important: The final syllable is not included in a foot.

Sample candidates

- (OC) to pus
- (OC) to (PUS)
- (OC) to (PUS)
- (OC) to (PUS)

A sample grammar that is a version of the English "grammar":

Best candidate for the correct grammar has a compatible contour

Knowledge representation comparison

**HV:** 5 parameters & 3 sub-parameters
Hypothesis space: 156 grammars

**Hayes:** 8 parameters
Hypothesis space: 768 grammars

English instantiations

**HV:** 5 parameters & 4 sub-parameters
Hypothesis space: 156 grammars

**Hayes:** 8 parameters
Hypothesis space: 768 grammars

OT: 9 violable constraints
Hypothesis space: 362,880 grammars

(English + 26 grammars)
Updated working assumption: The learner will try to learn a grammar that can account for all the productive data encountered (Legate & Yang 2012).

Why would this occur?

Perhaps the learner realizes that some data are unproductive, and therefore likely irregular and unpredictable. The goal then becomes to learn a grammar that can account for all the data that are predictable.

How would this occur?

A formal way for identifying if there is a dominant rule for a set of items is the Tolerance Principle (Yang 2005, Legate & Yang 2012). This is used to estimate how many exceptions a rule can tolerate in a set before it’s no longer useful for the learner to have the rule. If there are too many exceptions, it’s better not to have a rule and learn patterns on an individual item basis instead of having a rule that keeps getting violated.

The number of exceptions a rule can tolerate for a set of N items is:

\[ \frac{N}{\ln(N)} \]

(Yang 2005, Legate & Yang 2012)
The Tolerance Principle in action

For every syllable word form with multiple stress contours, the learner could assess whether any of those contours is the dominant one (the “rule” for that syllable word form), using the Tolerance Principle.

V VV

If one contour is dominant, the learner should focus on accounting for that pattern, since it’s regular and productive. The grammar should be able to generate it. The other contours can be ignored for purposes of learning the grammar.

If no contour is dominant, the learner should ignore this syllable word form for the purposes of learning the grammar since there is no obvious regularity to account for.

Productive data filter in action

Parametric: HV & Hayes, with inflectional knowledge
HV: 123 syllable word forms   Hayes: 149 syllable word forms
   3 syllable distinctions   4 syllable distinctions

V VV

These items are good for the HV English grammar.
These items are good for the Hayes English grammar.

The Tolerance Principle looks at the word types with each stress pattern. Each represents an individual item that might follow the regular stress pattern rule (if there is one).

How many exceptions are allowed? $506 / \ln(506) = 81$
If this is the dominant pattern, way too many exceptions:

162 + 325 > 81

Learner conclusion: No dominant stress pattern, so none of these syllable word form data should be used to learn the English grammar.

This will end up helping both grammars, since they won’t be penalized for the patterns they can’t account for.

However, the Hayes grammar is helped a little more, since it couldn’t account for the most frequent stress pattern before, while the HV grammar could.
Productive data filter in action

Constraint-based: OT, with inflectional knowledge
452 syllable word forms
8 syllable distinctions

These items are bad for all English grammars.

These items are good for most English grammars (21/26).

How many exceptions are allowed? $355 / \ln(355) \approx 60$

These items are good for a few English grammars (5/26).

How many exceptions are allowed? $355 / \ln(355) \approx 60$

If this is the dominant pattern, too many exceptions:
$316 + 14 > 60$
How many exceptions are allowed? $355 / \ln(355) \approx 60$

Under the OT syllable representation, there is a dominant stress pattern for this word form. Therefore, this pattern should be accounted for by the English grammar.

Unfortunately, this is the only pattern the English grammars cannot account for...this means a learner using the productivity filter would have even more trouble learning the current English OT grammar constraints.

The learnability problem:

Change the (immediate) target state. Assume there is a transitory state in learning that the learner reaches and then leaves once additional knowledge is acquired.
Learning English metrical phonology

One solution: The learner has derived additional knowledge that helps guide learning.

General knowledge: Interactions with morphology
(Chomsky & Halle 1968, Hayes 1982, Kiparsky 1979)

Specific knowledge: Adding productive morphology doesn’t change the stress pattern, even though all grammars base their stress patterns on the syllables present in the word.

- *Early*
- *Prettier*
- *Sensation*
- *Sationally*
- *Early*
- *Prettiest*
- *Sensational*
- *Sationally*

English children seem to use inflectional morphology productively around 3 (Brown 1973) – so they may be aware it doesn’t get stressed, based on their prior linguistic experience.

So how does the (best) English grammar compare to the other grammars defined by the knowledge representation, once the learner knows inflectional morphology is stressless?

Relative compatibility of the English grammar = proportion of grammars in the hypothesis space the (best) English grammar is better than

<table>
<thead>
<tr>
<th>Representation</th>
<th>Relative Compatibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>HV</td>
<td>0.712 by types out of 156 grammars</td>
</tr>
<tr>
<td>Hayes</td>
<td>0.704 by types out of 768 grammars</td>
</tr>
<tr>
<td>OT</td>
<td>0.786 by types out of 362,880 grammars</td>
</tr>
</tbody>
</table>

Better than many…but many are still better

Implication:
There remain many other grammars in the hypothesis space that are more compatible with the data, even though the learner knows inflectional morphology is stressless. It would be easier to pick one of these other more compatible grammars.

Relative compatibility of English grammar
HV: 0.712 by types out of 156 grammars
Hayes: 0.704 by types out of 768 grammars
OT: 0.786 by types out of 362,880 grammars

Better than many…but many are still better

Continuing conclusion:
The same learnability issues persist for the English grammar in all three knowledge representations, even when the learner has some knowledge of the interactions between morphology and metrical phonology.

Parametric: HV
Parametric: Hayes
Constraint-based: OT
### HV vs. Hayes on most frequent word forms

<table>
<thead>
<tr>
<th>Stressed wordform</th>
<th># Types</th>
<th>Examples</th>
<th>HV</th>
<th>Hayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp</td>
<td>592</td>
<td>water, doing, going</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Xp</td>
<td>472</td>
<td>little, getting, coming</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Li</td>
<td>334</td>
<td>baby, sweetie, mommy</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Xi</td>
<td>309</td>
<td>kitty, daddy, very</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ap</td>
<td>235</td>
<td>goodness, handsome, helper</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ll</td>
<td>188</td>
<td>okay, byebye, TV</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Ai</td>
<td>172</td>
<td>window, birdie, only</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>La</td>
<td>171</td>
<td>peanuts, secrets, highest</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Xa</td>
<td>170</td>
<td>biggest, buckets, hiccups</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>xl</td>
<td>145</td>
<td>below, today, hurray</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### The impact of morphological knowledge

*Example: What happens to words of the La stressed word form when the child gets morphological knowledge? (for the Hayes grammar, which can’t account for it without morphological knowledge)*

**Before morphological knowledge**

- 171 La (island, giant, moment)

**After morphological knowledge**

- 57 La (54 of the 171 + 3 added from Lp form)
  - Hayes still can’t account for these

- 100 Lp (father’s→father pockets→pocket slobbered→slobberer)

- 17 L (cutest→cute nicest→nice weirdest→weird)

- Hayes can now account for these

**In this case, knowing inflectional morphology is stressless helps!**

### Proposed learning biases/capabilities

Several learning biases/capabilities are potentially both innate and domain-specific.

<table>
<thead>
<tr>
<th>Innate</th>
<th>Derived</th>
<th>Domain-specific</th>
<th>Domain-general</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn from all wh-dependencies</td>
<td>?</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>Parse data into phrase structure trees</td>
<td>?</td>
<td>?</td>
<td>*</td>
</tr>
<tr>
<td>Attend to container nodes &amp; subcategorize by CP</td>
<td>?</td>
<td>?</td>
<td>*</td>
</tr>
<tr>
<td>Extract &amp; track container node trigrams</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calculate dependency probability from trigrams</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Pearl & Sprouse 2013a, 2013b, under review*
### Parse data into phrase structure trees

<table>
<thead>
<tr>
<th>Innate</th>
<th>Derived</th>
<th>Domain-specific</th>
<th>Domain-general</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

- **Learn from all wh-dependencies**
  - **Domain-specific**, since this is language data.
  - May seem reasonable to attend to wh-dependency data when learning about wh-dependencies (and so this would be **derived**)

- **Learn from all wh-dependencies**
  - **Domain-specific**, since this is language data.
  - May seem reasonable to attend to wh-dependency data when learning about wh-dependencies (and so this would be **derived**)

- **Attend to container nodes & subcategorize by CP**
  - **Domain-specific**, since the structure is specific to language.
  - May be possible to bootstrap this information (acquiring syntactic categories: Mintz 2003, 2006; acquisition of hierarchical structure given syntactic categories as input: Klein & Manning 2002). If so, this would be **derived**...

  ...but it’s currently **unclear** if all the necessary phrase structure knowledge can be bootstrapped.

  **Important:** The need for this capability is not specific to learning islands – it’s (presumably) needed for learning any kind of syntactic knowledge.
### Identifying container nodes
- Applies to language data: domain-specific
- Derived from ability to parse utterances

### Attending to container nodes (among all the other data out there)
- Applies to language data: domain-specific
- Innate vs. derived?
  - Could be specified innately (like bounding nodes)
  - Could be derived from a bias to use representations that are already being used for parsing

### About a linguistic representation: domain-specific
- Innate vs. derived?
  - Could be specified innately

...but still have to know this is the right thing to subcategorize.
Extract & track container node trigrams

Applied in different cognitive domains: domain-general


...though why trigrams instead of some other n-gram?

Calculate dependency probability from trigrams

Main implications of this learner

(2) Even if Universal Grammar learning biases are required, they are different from (and less specific than) the biases previously proposed.

In particular, while one bias also specifies a particular linguistic representation, there is no bias defining the “constraint”. This falls out from the other non-UG learning biases.

Calculate dependency probability from trigrams

Why learning from container node trigrams works

For each island-spanning dependency, there is at least one extremely low probability container node trigram in the dependency.

Complex NP island

Subject island

Whether island

Adjunct island

These trigrams are never observed in the input – which is crucially different than being observed rarely. Thus, these islands are worse than dependencies involving trigrams that are rarely seen (e.g., dependencies with CPnull) and even longer dependencies that involve more frequent trigrams (e.g., triply embedded object dependencies using CPnull).

The empirical necessity of trigrams

Not unigrams
A unigram model will successfully learn Whether and Adjunct islands, as there are container nodes in these dependencies that never appear in grammatical dependencies (CPwhether and CPAdjunct)….but it will fail to learn Complex NP and Subject islands, as all of the container nodes in these islands are shared with grammatical dependencies.

Complex NP:

Subject:

Whether:

Adjunct:
The empirical necessity of trigrams

Not bigrams
At least for Subject islands, there is no bigram that occurs in a Subject island violation but not in any grammatical dependencies. The most likely candidate for such a bigram is IP-NP...However, sentences such as *What, again, about Jack impresses you? or *What did you say about the movie scared you? suggest that a gap can arise inside of NPs, as long as the extraction is of the head noun (what), not of the noun complement of the preposition.

Complex NP: IP-VP-NP_CP\_true\_IP-VP
Subject: *IP-VP-CP\_true\_IP-NP-PP
Whether: IP-VP-CP\_whether\_IP-VP
Adjunct: IP-VP-CP\_IP-VP

Parasitic gaps

The learner can’t handle parasitic gaps, which are dependencies that span an island (and so should be ungrammatical) but which are somehow rescued by another dependency in the utterance.

*Which book did you laugh [before reading ___]?  
Which book did you judge ___true [before reading ___parasitic]?  

Adjunct island

*What did [the attempt to repair ___] ultimately damage the car?  
What did [the attempt to repair ___parasitic] ultimately damage ___true?  

Complex NP island

Parasitic gaps

Why not? The current learner would judge the parasitic gap as ungrammatical since it is inside an island, irrespective of what other dependencies are in the utterance.

*Which book did you laugh [before reading ___]?  
Which book did you judge ___true [before reading ___parasitic]?  

Adjunct island

*What did [the attempt to repair ___] ultimately damage the car?  
What did [the attempt to repair ___parasitic] ultimately damage ___true?  

Complex NP island

This may be able to be addressed in a learner that is able to combine information from multiple dependencies in an utterance (perhaps because the learner has observed multiple dependencies resolved in utterances in the input).

Across-the-board constructions

A similar problem occurs for across-the-board constructions.

*Which book did you [[read ___] and [then review ___]]?  
dependency for both gaps: IP-VP-VP

*Which book did you [[read the paper] and [then review ___]]?  
dependency for gap: IP-VP-VP

*Which book did you [[read ___] and [then review the paper]]?  
dependency for gap: IP-VP-VP

Again, this may be able to be addressed in a learner that is able to combine information from multiple dependencies in an utterance (perhaps because the learner has observed multiple dependencies resolved in utterances in the input).

Some cross-linguistic issues

High probability trigrams that may be ungrammatical

Rizzi (1982): reports situations in Italian where simply doubling a grammatical sequence of trigrams leads to ungrammaticality...

But these involve the same trigrams, so the learner in Pearl & Sprouse (forthcoming) will treat both the same (either grammatical or ungrammatical). If humans do have different judgments of these, then this cannot be accounted for by this learning algorithm.

Complementizer that

That-trace effects

*Who do you think that ___ read the book?  
Who do you think ___ read the book?  

The current learning strategy captures this distinction.
Complementizer *that*

That-trace effects

…but the current learning strategy will also generate a preference for object gaps without *that* compared to object gaps with *that*. (object *that*-trace effect)

What do you think that he read __? [prefers this one]

What do you think he read __?

Interestingly, Cowart 1997 finds an object *that*-trace effect, but it is much smaller than the subject *that*-trace effect.

The model generates an asymmetrical dispreference when using adult-directed corpora, which contain more instances of *that* (5.40 versus 2.81). This could be taken to be a developmental prediction of the current algorithm: Children may disprefer object gaps in embedded *that*-CP clauses more than adults, and this dispreference will weaken as they are exposed to additional tokens of *that* in utterances containing dependencies.

Pearl & Sprouse 2013a

Now what?

This learning strategy for *wh*-dependencies makes some developmental predictions – can we verify these experimentally?

“*that*-trace” effect prediction:
Children initially disprefer all dependencies containing *that*, even ones adults allow

Subject extraction
*Who do you think *that* read the book?*

Who do you think __ read the book?

Object extraction
*What do you think *that* he read __?*

What do you think he read __?

Now what?

This learning strategy for *wh*-dependencies makes some developmental predictions – can we verify these experimentally?

“*that*-trace” effect prediction:
Children initially disprefer all dependencies containing *that*, even ones adults allow

Subject extraction
*Who do you think __ read the book?*

Who do you think __ read the book?

Object extraction
*What do you think *that* he read __?*

What do you think he read __?

Now what?

How does this learning strategy for *wh*-dependencies measure up cross-linguistically?

Island effects vary.
Ex: Italian does not have a subject island effect when the *wh*-dependency is part of a relative clause, though it does when the *wh*-dependency is part of a question. (Sprouse et al. submitted)

Would the input naturally lead our kind of learner to this distinction?

Now what?

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

Related phenomena: The distribution of gaps

*Parasitic gaps*: Dependencies that span an island (and so should be ungrammatical) but which are somehow rescued by another dependency in the utterance.

*Which book did you laugh before reading __?*

Before reading __?

*Which book did you judge __ true before reading __?*

Before reading __?
Now what?

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

Related phenomena: The distribution of gaps

**Across-the-board (ATB) extraction**: Similar situation.

*Which book did you [read __ and (then review __)]?*  
dependency for both gaps: IP-VP-VP

*Which book did you [read the paper and (then review __)]?*  
dependency for gap: IP-VP-VP

*Which book did you [read __ and (then review the paper)]?*  
dependency for gap: IP-VP-VP

Now what?

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

**Semi-related phenomena: Binding dependencies**

There don’t appear to be the same restrictions on binding dependencies that there are on wh-dependencies.

*The boy thought the joke about himself was really funny.*

*Who did the boy think [the joke about __] was really funny?*  
Subject island

Pearl & Sprouse 2013a