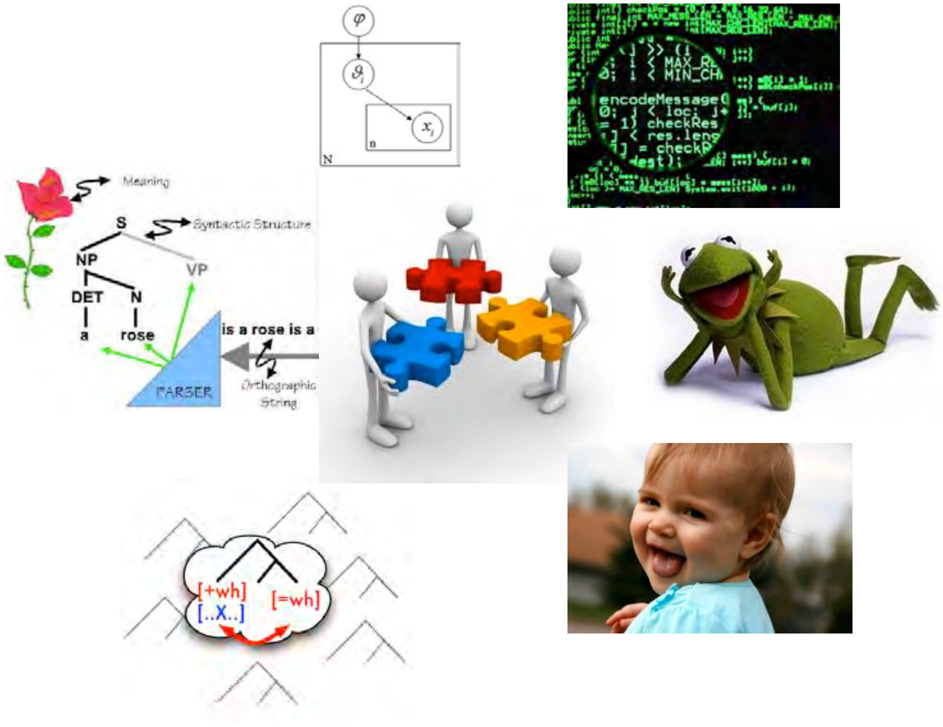
Two good ways to use computational methods to understand language (Acquisition edition)

Lisa Pearl

University of California, Irvine

Computation of
Language
Laboratory
UC Irvine



May 4, 2012: Mayfest 2012

University of Maryland, College Park

Method: “a systematic procedure, technique, or mode of inquiry employed by …a particular discipline or art” – Merriam Webster Online Dictionary

…to tell us something we didn’t know before.

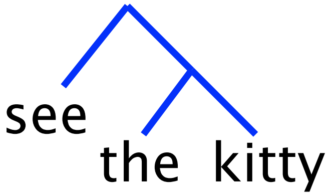
Method: “a systematic procedure, technique, or mode of inquiry employed by …a particular discipline or art” – Merriam Webster Online Dictionary

…to tell us something we didn’t know before.

Theoretical methods:

What knowledge of language is (and what children have to learn)

SEE the KItty



si ðə kiɾi

$$\begin{bmatrix}+\text{stop}\\+\text{consonant}\\+\text{alveolar}\end{bmatrix} \rightarrow [\text{ɾ}] \;/\; \begin{bmatrix}+\text{vowel}\\+\text{stressed}\end{bmatrix} \_\_ \begin{bmatrix}+\text{vowel}\\-\text{stressed}\end{bmatrix}$$

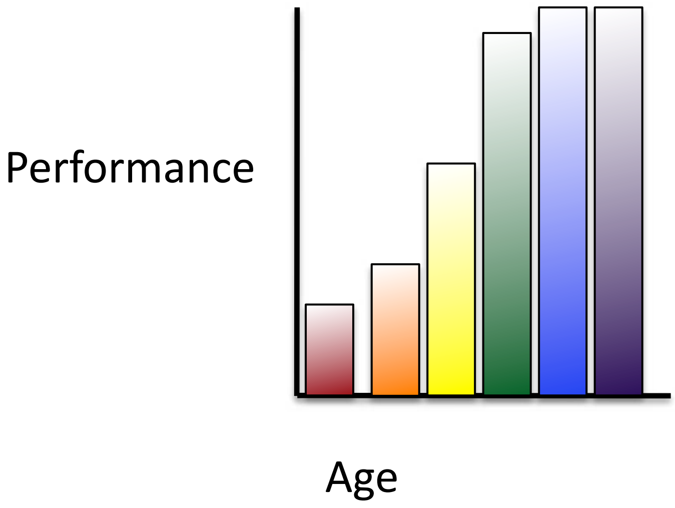
see’(the kitty)(x_{listener})

Method: "a systematic procedure, technique, or mode of inquiry employed by …a particular discipline or art" – Merriam Webster Online Dictionary

…to tell us something we didn't know before.

Experimental methods:

When knowledge is acquired & plausible capabilities about **how**



$$\frac{p(ki\quad tty)}{p(ki)}$$

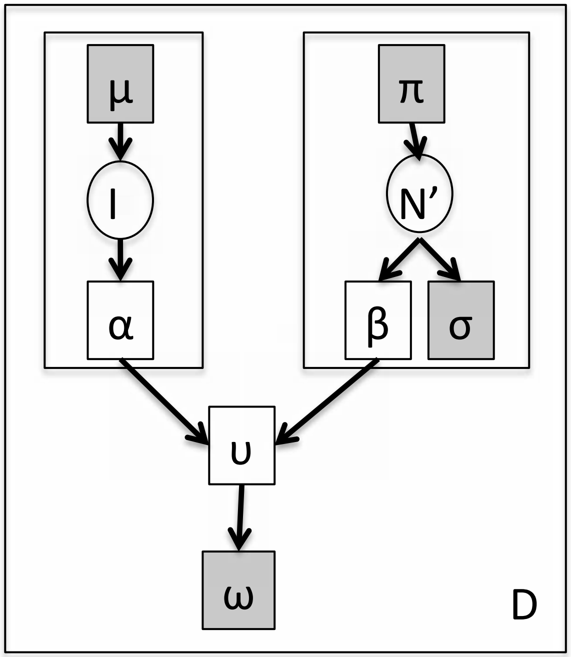
$p(\mathrm{H1}\,|$ $)$

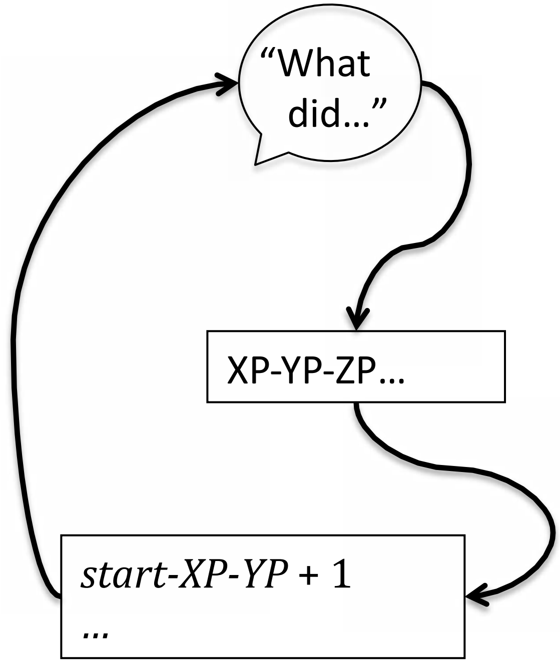
$\propto\ p($ $|\ \mathrm{H1})\ p(\mathrm{H1})$

Method: "a systematic procedure, technique, or mode of inquiry employed by …a particular discipline or art" – Merriam Webster Online Dictionary

…to tell us something we didn't know before.

Computational methods:
Strategies that are both useful and useable for **how** children acquire knowledge





Computational methods often rely on the results of theoretical and experimental methods, and can be used to inform both theory and the learning process.

Computational methods

Theoretical methods

Experimental methods

Road map: Two good ways

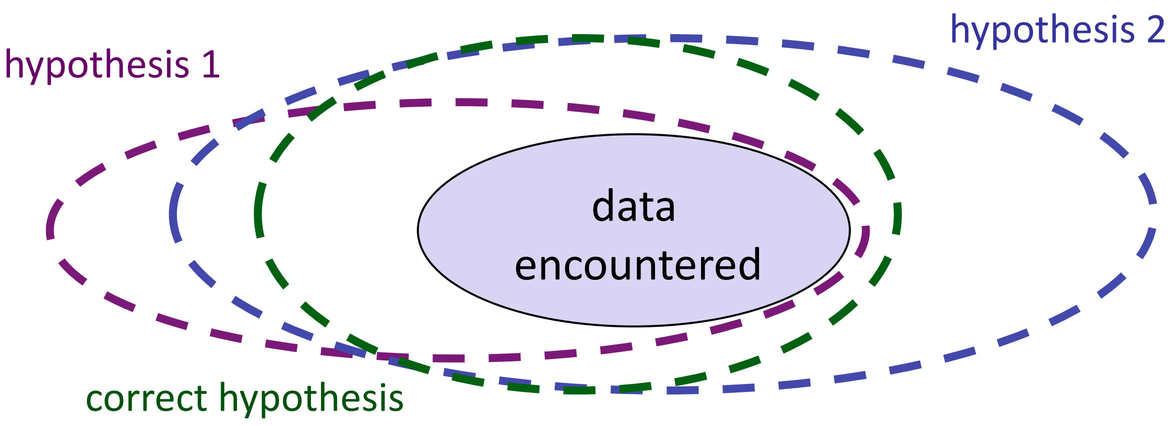
- Informing theory: Arguments from acquisition
  - Investigating Universal Grammar
  - Testing theories of knowledge representation
- Informing the learning process: Useful, useable, and better than adults?
  - Comparing ideal and non-ideal approaches to discover how “less is more”

Road map: Two good ways

- Informing theory: Arguments from acquisition
  - Investigating Universal Grammar
  - Testing theories of knowledge representation
- Informing the learning process: Useful, useable, and better than adults?
  - Comparing ideal and non-ideal approaches to discover how “less is more”

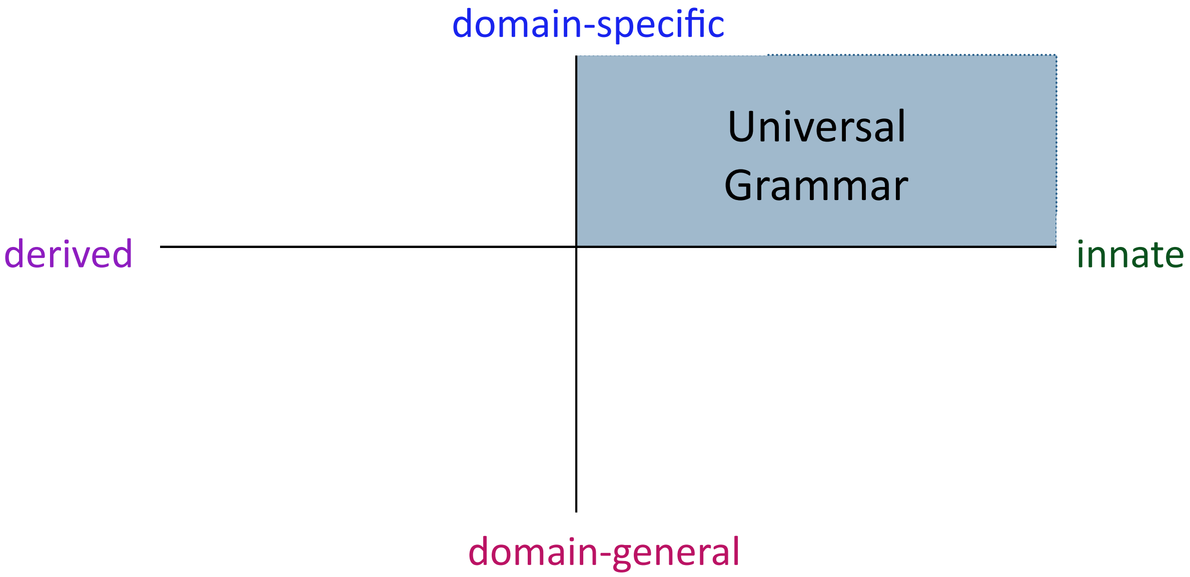
Informing Theory: Arguments from Acquisition

One explicit motivation for Universal Grammar is that it explains how children solve the induction problem inherent in language acquisition.



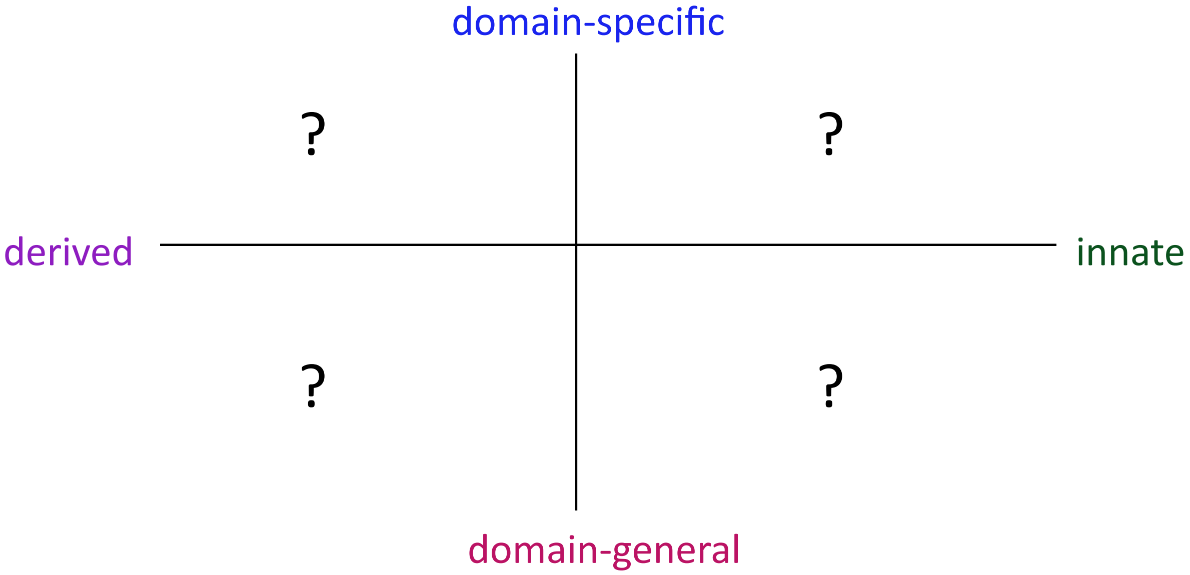
Informing Theory: Arguments from Acquisition

Specifically, Universal Grammar consists of the necessary learning biases that are both innate and domain-specific (Chomsky 1965, Chomsky 1975).



Informing Theory: Arguments from Acquisition

Open question: For any given piece of linguistic knowledge, what biases are necessary to learn it from child-directed data? Are any of them necessarily both innate and domain-specific?



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 __?

Pearl & Sprouse submitted

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").

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 __]?

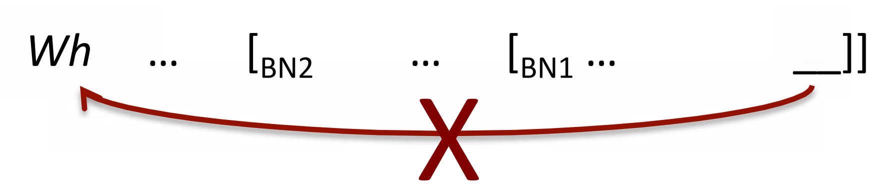
Pearl & Sprouse submitted

Syntactic islands

- **Predominant theory in generative syntax:**
  syntactic islands require innate, domain-specific learning biases

Example: Subjacency (Chomsky 1973, Huang 1982, Lasnik & Saito 1984)
A dependency cannot cross two or more bounding nodes.

Bounding nodes: language-specific (CP, IP, and/or NP)



Wh … $[_{BN2}$ … $[_{BN1}$ … __]]

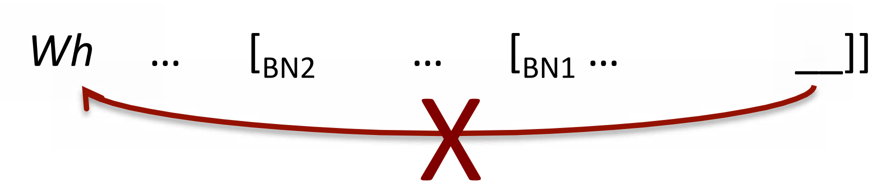
Pearl & Sprouse submitted

Syntactic islands

- **Predominant theory in generative syntax:**

  syntactic islands require innate, domain-specific learning biases

Subjacency learning biases:
(1) Innate, domain-specific knowledge of hypothesis space: Exclude hypotheses that allow dependencies crossing 2+ bounding nodes.



Wh … $[_{\text{BN2}}$ … $[_{\text{BN1}}$ … __]]

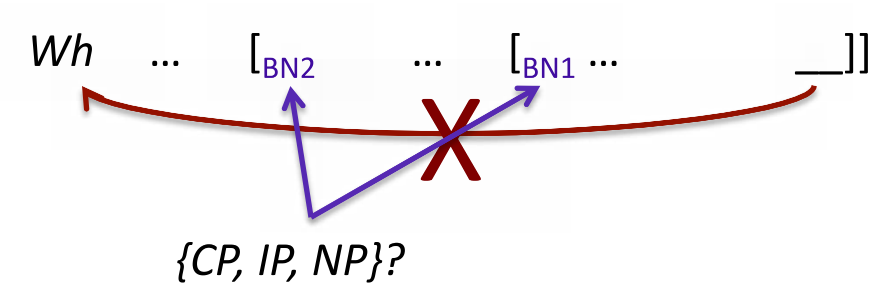
Pearl & Sprouse submitted

Syntactic islands

- **Predominant theory in generative syntax:**
  syntactic islands require innate, domain-specific learning biases

Subjacency learning biases:
(1) Innate, domain-specific knowledge of hypothesis space: Exclude hypotheses that allow dependencies crossing 2+ bounding nodes.

(2) Innate, domain-specific knowledge of hypothesis space: Hypothesis space consists of bounding nodes for all languages, and the child must identify the ones applicable to his language.

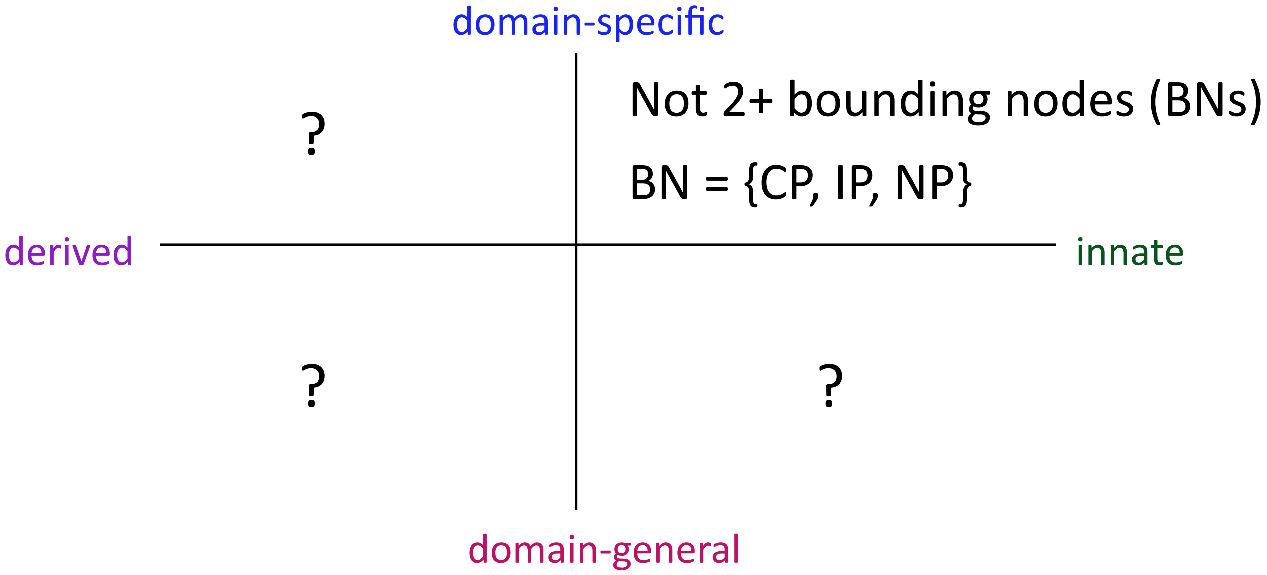


Pearl & Sprouse submitted

Syntactic islands

- **Predominant theory in generative syntax:**

  syntactic islands require innate, domain-specific learning biases…in addition to whatever else they might require.



domain-specific

?

Not 2+ bounding nodes (BNs)

BN = {CP, IP, NP}

derived

innate

?

?

domain-general

Pearl & Sprouse submitted

Syntactic islands

- **How do we test this?**

(1) Explicitly define the target knowledge state, using adult acceptability judgments.

(2) Identify the data available in the input, using realistic samples. (Is there an induction problem?)

(3) Implement a probabilistic learner that can learn about syntactic islands and see what kind of learning biases it requires.

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

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)

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

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)

Complex NP islands

Who __ claimed that Lily forgot the necklace? matrix | non-island
What did the teacher claim that Lily forgot __? embedded | non-island
Who __ made the claim that Lily forgot the necklace? matrix | island
*What did the teacher make the claim that Lily forgot __? embedded | island

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

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

Who __ thinks the necklace is expensive? matrix | non-island

What does Jack think __ is expensive? embedded | non-island

Who __ thinks the necklace for Lily is expensive? matrix | island

*Who does Jack think the necklace for __ is expensive? embedded | island

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

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

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

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)

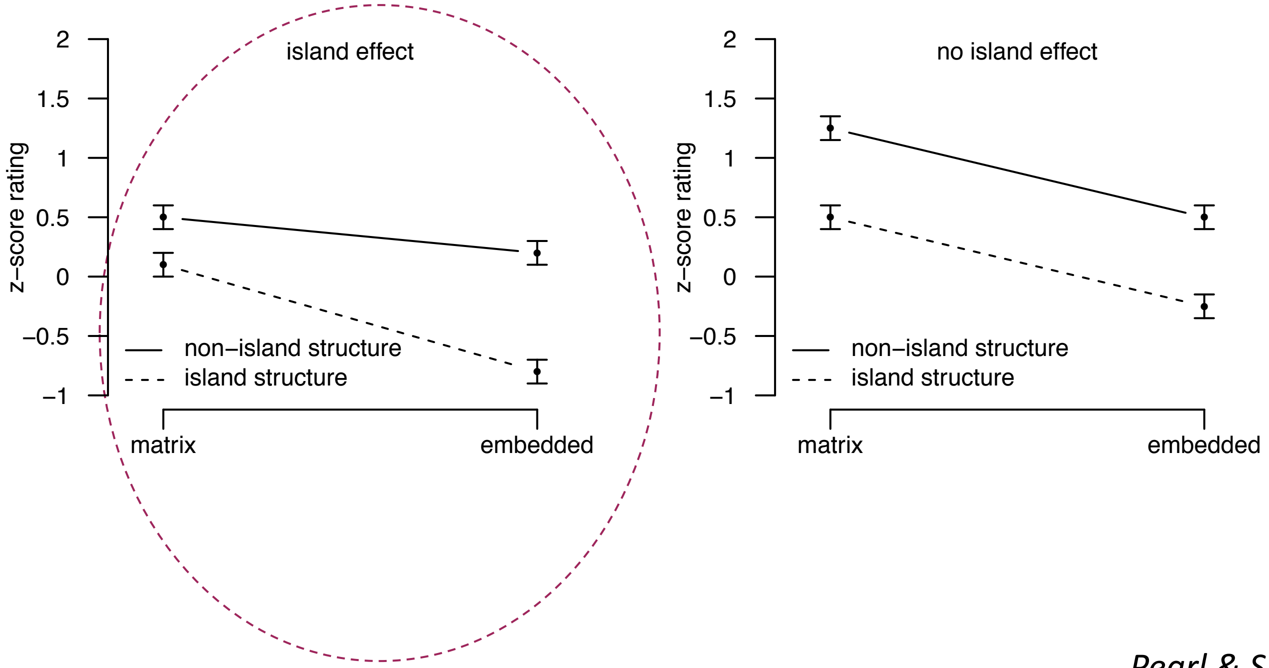
Adjunct islands

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

Pearl & Sprouse submitted

The target state:
Adult knowledge of syntactic islands

Syntactic island = superadditive interaction of the two factors (additional unacceptability that arises when the two factors are combined, above and beyond the independent contribution of each factor).



Pearl & Sprouse submitted

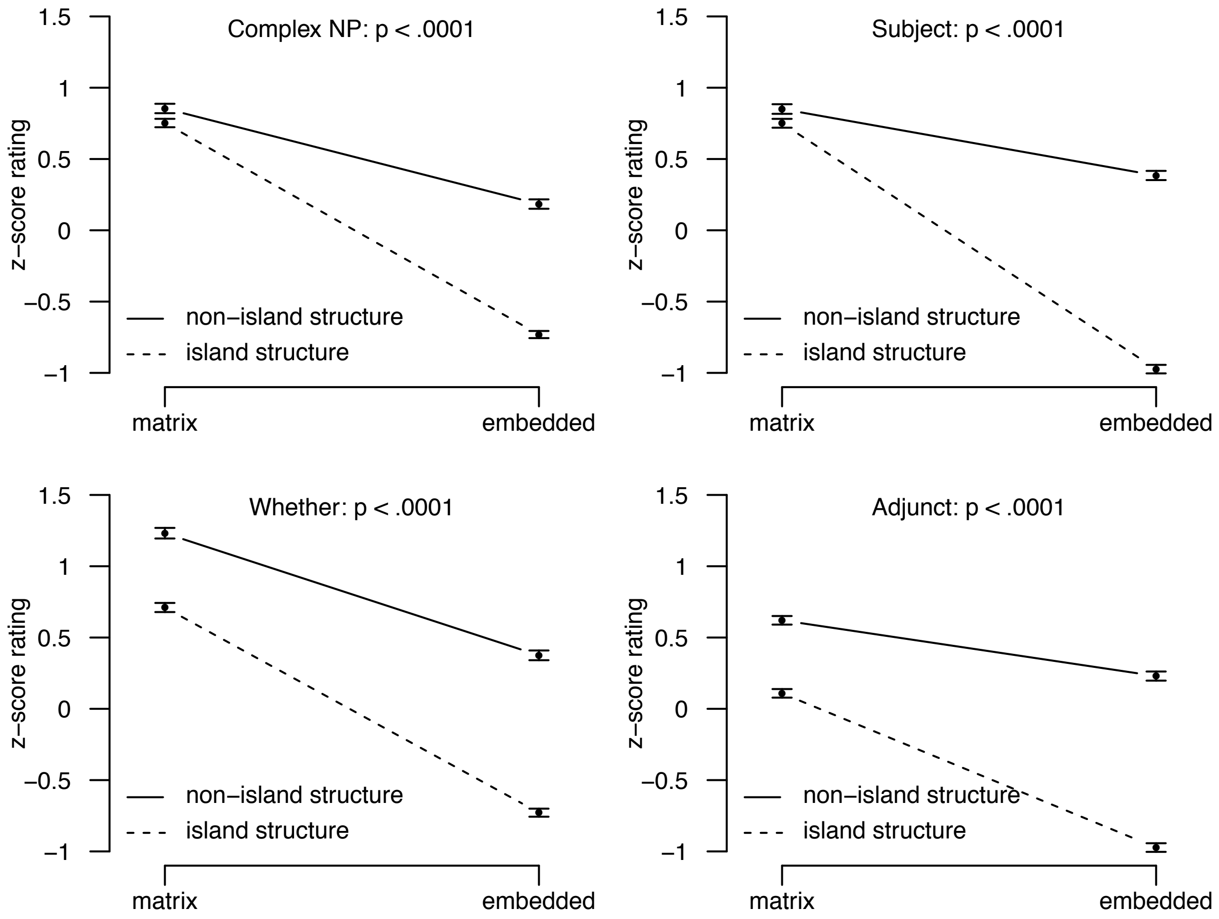
The target state:
Adult knowledge of syntactic islands

Sprouse et al. (2012)'s data on the four island types (173 subjects)

Superadditivity present for all islands tested

=

Knowledge that dependencies cannot cross these island structures is part of the adult knowledge state



Pearl & Sprouse submitted

The input: Assessing the induction problem

Data from five corpora of child-directed speech (Brown-Adam, Brown-Eve, Brown-Sarah, Suppes, Valian) from CHILDES (MacWhinney 2000): speech to 25 children between the ages of one and five years old.

Total words: 813,036

Utterances containing a *wh*-dependency: 31,247

Sprouse et al. (2012) stimuli types:

ungrammatical

	MATRIX + NON-ISLAND	EMBEDDED + NON-ISLAND	MATRIX + ISLAND	*EMBEDDED + ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

Pearl & Sprouse submitted

The input: Assessing the induction problem

These kinds of utterances are fairly rare in general - the most frequent appears about 0.9% of the time (295 of 31,247.)

Sprouse et al. (2012) stimuli types (**out of 31,247**):

ungrammatical

	MATRIX + NON-ISLAND	EMBEDDED + NON-ISLAND	MATRIX + ISLAND	*EMBEDDED + ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

Pearl & Sprouse submitted

The input: Assessing the induction problem

Being grammatical doesn’t necessarily mean an utterance will appear in the input at all.

Sprouse et al. (2012) stimuli types (**out of 31,247**):

ungrammatical

	MATRIX + NON-ISLAND	EMBEDDED + NON-ISLAND	MATRIX + ISLAND	*EMBEDDED + ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

Pearl & Sprouse submitted

The input: Assessing the induction problem

Unless the child is sensitive to very small frequencies, it’s difficult to tell the difference between grammatical and ungrammatical dependencies sometimes…

Sprouse et al. (2012) stimuli types (**out of 31,247**):

ungrammatical

	MATRIX + NON-ISLAND	EMBEDDED + NON-ISLAND	MATRIX + ISLAND	*EMBEDDED + ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

Pearl & Sprouse submitted

The input: Assessing the induction problem

…and impossible to tell no matter what the rest of the time. This looks like an induction problem for the language learner if we're looking for direct evidence in the input.

Sprouse et al. (2012) stimuli types (**out of 31,247**):

ungrammatical

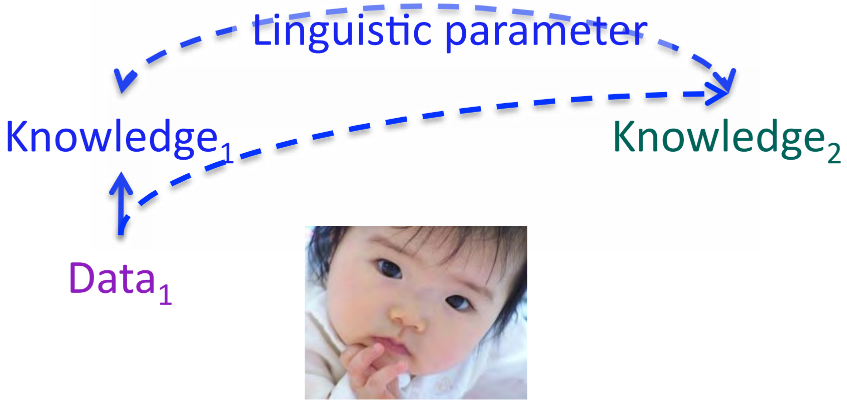
	MATRIX + NON-ISLAND	EMBEDDED + NON-ISLAND	MATRIX + ISLAND	*EMBEDDED + ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

Pearl & Sprouse submitted

Building a computational learner

Idea: Use indirect positive evidence, too.

Similar in spirit to linguistic parameters: Data are deemed informative, even if they are not data about the specific phenomenon of interest.



Here: Dependencies other than the ones of interest (the Sprouse et al. 2012 stimuli) are useful to learn from.

Pearl & Sprouse submitted

Building a computational learner

Learning Bias: Children track the occurrence of structures that can be derived from phrase structure trees during parsing - container nodes.

[$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ like ___]]]?
IP VP

Container node sequence: IP-VP

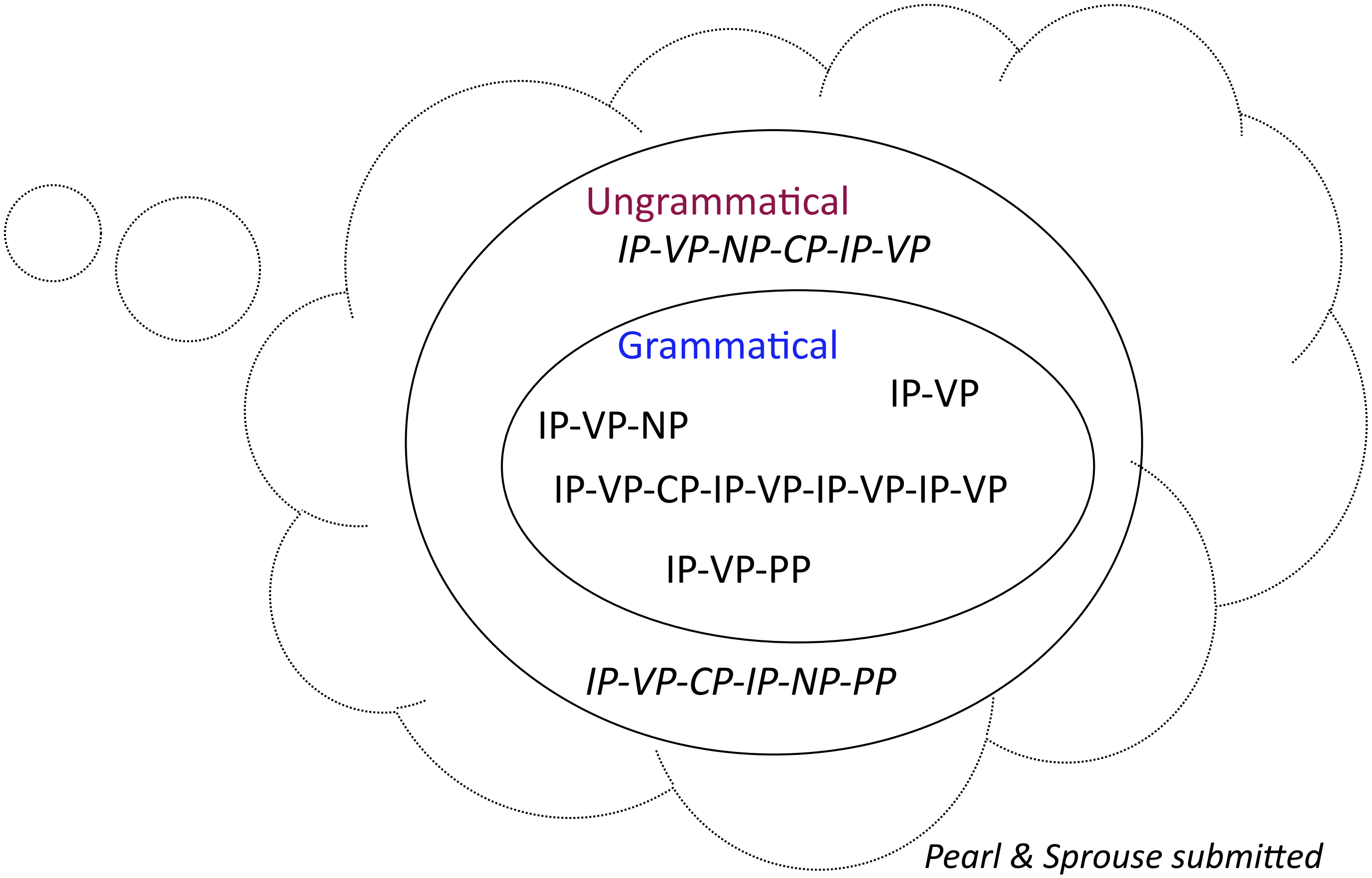
[$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ think [$_{CP}$ [$_{IP}$ [$_{NP}$ the gift] [$_{VP}$ was [$_{PP}$ from ___]]]]]]]]?
IP VP CP IP VP PP

Container node sequence: IP-VP-CP-IP-VP-PP

Pearl & Sprouse submitted

Building a computational learner

Children's hypotheses are about what container node sequences are grammatical for dependencies in the language.



Ungrammatical

IP-VP-NP-CP-IP-VP

Grammatical

IP-VP

IP-VP-NP

IP-VP-CP-IP-VP-IP-VP-IP-VP

IP-VP-PP

IP-VP-CP-IP-NP-PP

Pearl & Sprouse submitted

What does the target knowledge look like?

Sprouse et al. (2012) stimuli:

Complex NP islands		**Subject islands**
IP	matrix | non-island	IP
IP-VP-CP-IP-VP	embedded | non-island	IP-VP-CP-IP
IP	matrix | island	IP
*IP-VP-NP-CP-IP-VP	embedded | island	*IP-VP-CP-IP-NP-PP

All the ungrammatical dependencies are distinct from all the grammatical dependencies for these syntactic islands.

Pearl & Sprouse submitted

What does the target knowledge look like?

Sprouse et al. (2012) stimuli:

Whether islands			**Adjunct islands**
IP	matrix	non-island	IP
IP-VP-CP-IP-VP	embedded	non-island	IP-VP-CP-IP-VP
IP	matrix	island	IP
*IP-VP-CP-IP-VP	embedded	island	*IP-VP-CP-IP-VP

Pearl & Sprouse submitted

What does the target knowledge look like?

Sprouse et al. (2012) stimuli:

Whether islands			**Adjunct islands**
IP	matrix	non-island	IP
IP-VP-CP-IP-VP	embedded	non-island	IP-VP-CP-IP-VP
IP	matrix	island	IP
*IP-VP-CP-IP-VP	embedded	island	*IP-VP-CP-IP-VP

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

Pearl & Sprouse submitted

Building a computational learner

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

CP_{null}, CP_{that}, $CP_{whether}$, CP_{if}, etc.

The learner can then distinguish between these structures:

IP-VP-$CP_{null/that}$-IP-VP
IP-VP-$CP_{whether/if}$-IP-VP

Pearl & Sprouse submitted

What does the target knowledge look like?

Sprouse et al. (2012) stimuli:

Complex NP islands			**Subject islands**
IP	matrix	non-island	IP
IP-VP-CP$_{that}$-IP-VP	embedded	non-island	IP-VP-CP$_{null}$-IP
IP	matrix	island	IP
*IP-VP-NP-CP$_{that}$-IP-VP	embedded	island	*IP-VP-CP$_{null}$-IP-NP-PP

All the ungrammatical dependencies are still distinct from all the grammatical dependencies for these syntactic islands.

Pearl & Sprouse submitted

What does the target knowledge look like?

Sprouse et al. (2012) stimuli:

Whether islands			**Adjunct islands**
IP	matrix	non-island	IP
IP-VP-CP$_{that}$-IP-VP	embedded	non-island	IP-VP-CP$_{that}$-IP-VP
IP	matrix	island	IP
*IP-VP-CP$_{whether}$-IP-VP	embedded	island	*IP-VP-CP$_{if}$-IP-VP

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

Pearl & Sprouse submitted

Building a computational learner

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence's probability is the smoothed product of its trigrams.

Pearl & Sprouse submitted

Building a computational learner

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence's probability is the smoothed product of its trigrams.

$[_{CP}$ Who did $[_{IP}$ she $[_{VP}$ think $[_{CP}$ $[_{IP}$ $[_{NP}$ the gift] $[_{VP}$ was $[_{PP}$ from __]]]]]]]]?

IP VP CP_{null} IP VP PP

start-IP-VP-CP_{null}-IP-VP-PP-end =

start-IP-VP

IP-VP-CP_{null}

VP-CP_{null}-IP

CP_{null}-IP-VP

IP-VP-PP

VP-PP-end

Pearl & Sprouse submitted

Building a computational learner

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence's probability is the smoothed product of its trigrams.

$[_{CP}$ Who did $[_{IP}$ she $[_{VP}$ think $[_{CP}$ $[_{IP}$ $[_{NP}$ the gift] $[_{VP}$ was $[_{PP}$ from __]]]]]]]]?
IP VP CP$_{null}$ IP VP PP

start-IP-VP-CP$_{null}$-IP-VP-PP-end =
start-IP-VP
IP-VP-CP$_{null}$
VP-CP$_{null}$-IP
CP$_{null}$-IP-VP
IP-VP-PP
VP-PP-end

Probability(IP-VP-CP$_{null}$-IP-VP-PP) = p(start-IP-VP-CP$_{null}$-IP-VP-PP-end)
= p(start-IP-VP) * p(IP-VP-CP$_{null}$)*p(VP-CP$_{null}$-IP)*p(CP$_{null}$-IP-VP)
*p(IP-VP-PP)*p(VP-PP-end)

Pearl & Sprouse submitted

Building a computational learner

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence's probability is the smoothed product of its trigrams.

What this does:

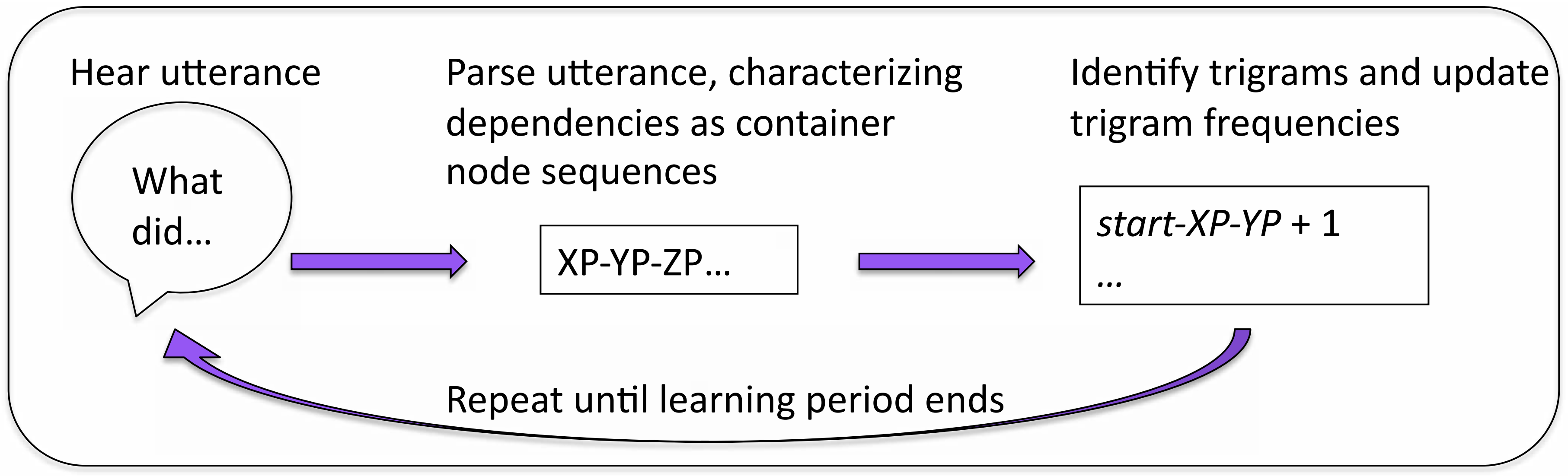
- longer dependencies are less probable than shorter dependencies, all other things being equal

- individual trigram frequency matters: short dependencies made of infrequent trigrams will be less probable than longer dependencies made of frequent trigrams

Effect: the frequencies observed in the input can temper the detrimental effect of dependency length.

Pearl & Sprouse submitted

Learning process



Hear utterance

What did…

Parse utterance, characterizing dependencies as container node sequences

XP-YP-ZP…

Identify trigrams and update trigram frequencies

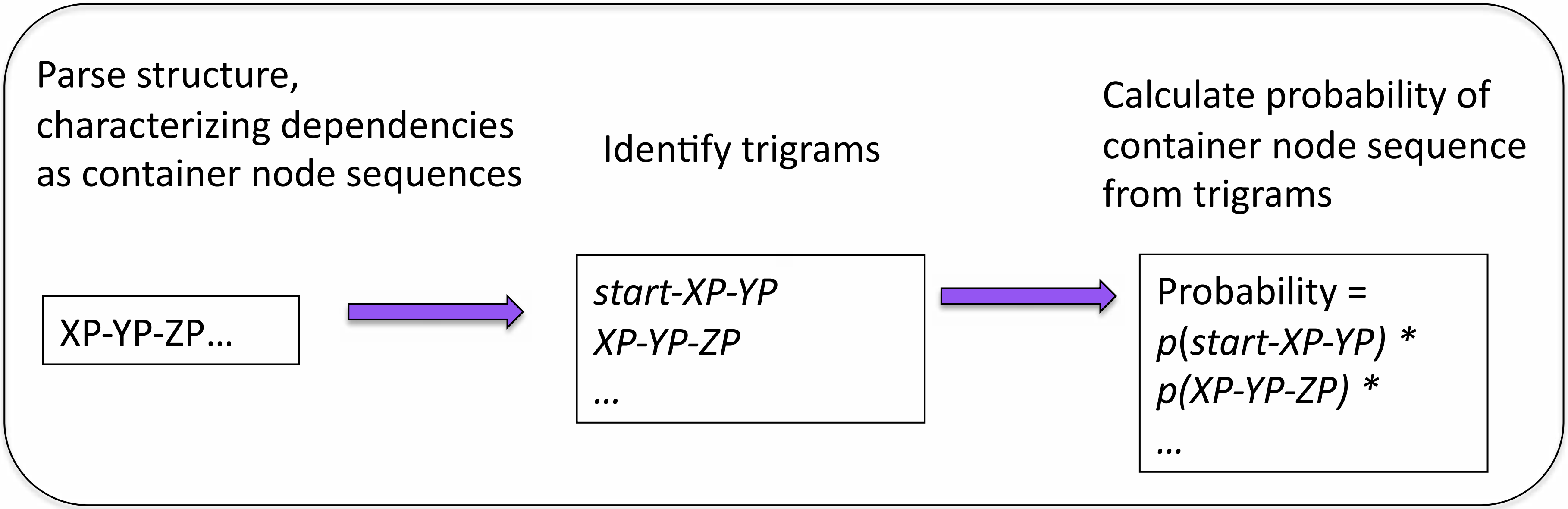
start-XP-YP + 1

…

Repeat until learning period ends

Pearl & Sprouse submitted

Generating grammaticality preferences



Parse structure, characterizing dependencies as container node sequences

XP-YP-ZP…

Identify trigrams

start-XP-YP
XP-YP-ZP
…

Calculate probability of container node sequence from trigrams

Probability =
p(start-XP-YP) *
p(XP-YP-ZP) *
…

Pearl & Sprouse submitted

Building a computational learner: Empirical grounding

Child-directed speech (Brown-Adam, Brown-Eve, Suppes, Valian) from CHILDES:

What kind of dependencies are present?

76.7%	IP-VP	*What did you see __?*
12.8%	IP	*What __ happened?*
5.6%	IP-VP-IP-VP	*What did she want to do __?*
2.5%	IP-VP-PP	*What did she read from __?*
1.1%	IP-VP-CP_{null}-IP-VP	*What did she think he said __?*

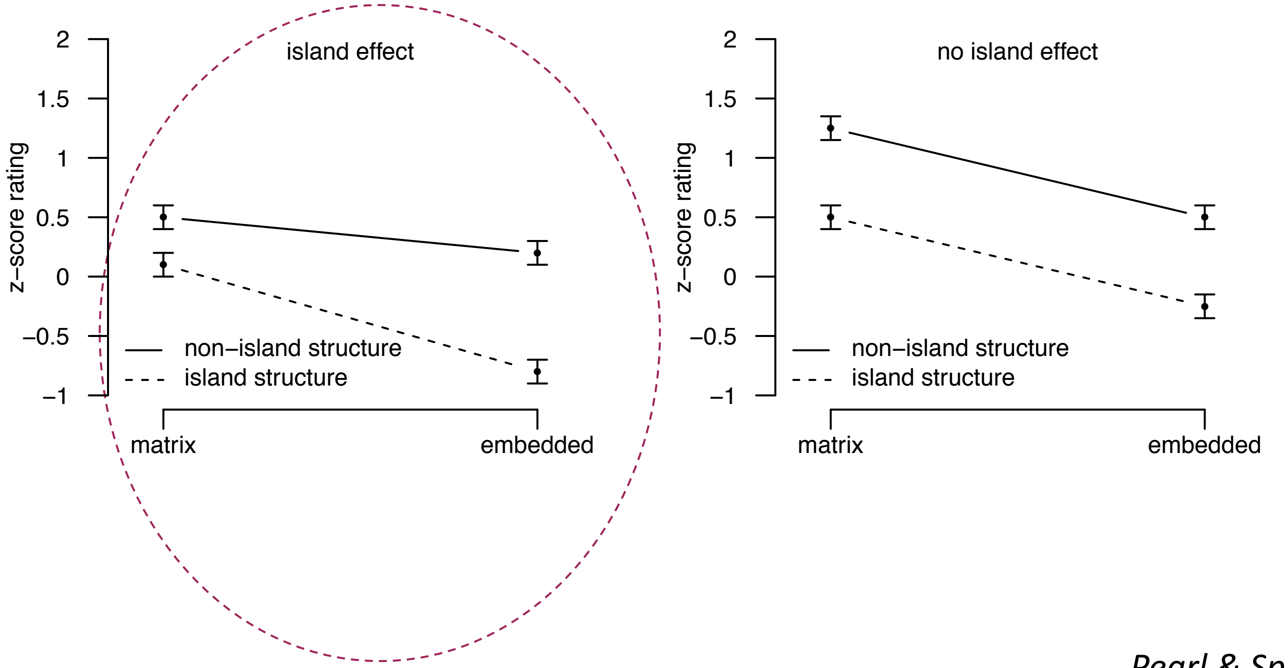
…

Pearl & Sprouse submitted

Success metrics

Compare learned grammaticality preferences to Sprouse et al. (2012) judgment data.

Then, for each island, we plot the predicted grammaticality preferences from the modeled learner on an interaction plot, using log probability of the dependency on the y-axis. Non-parallel lines indicate knowledge of islands.

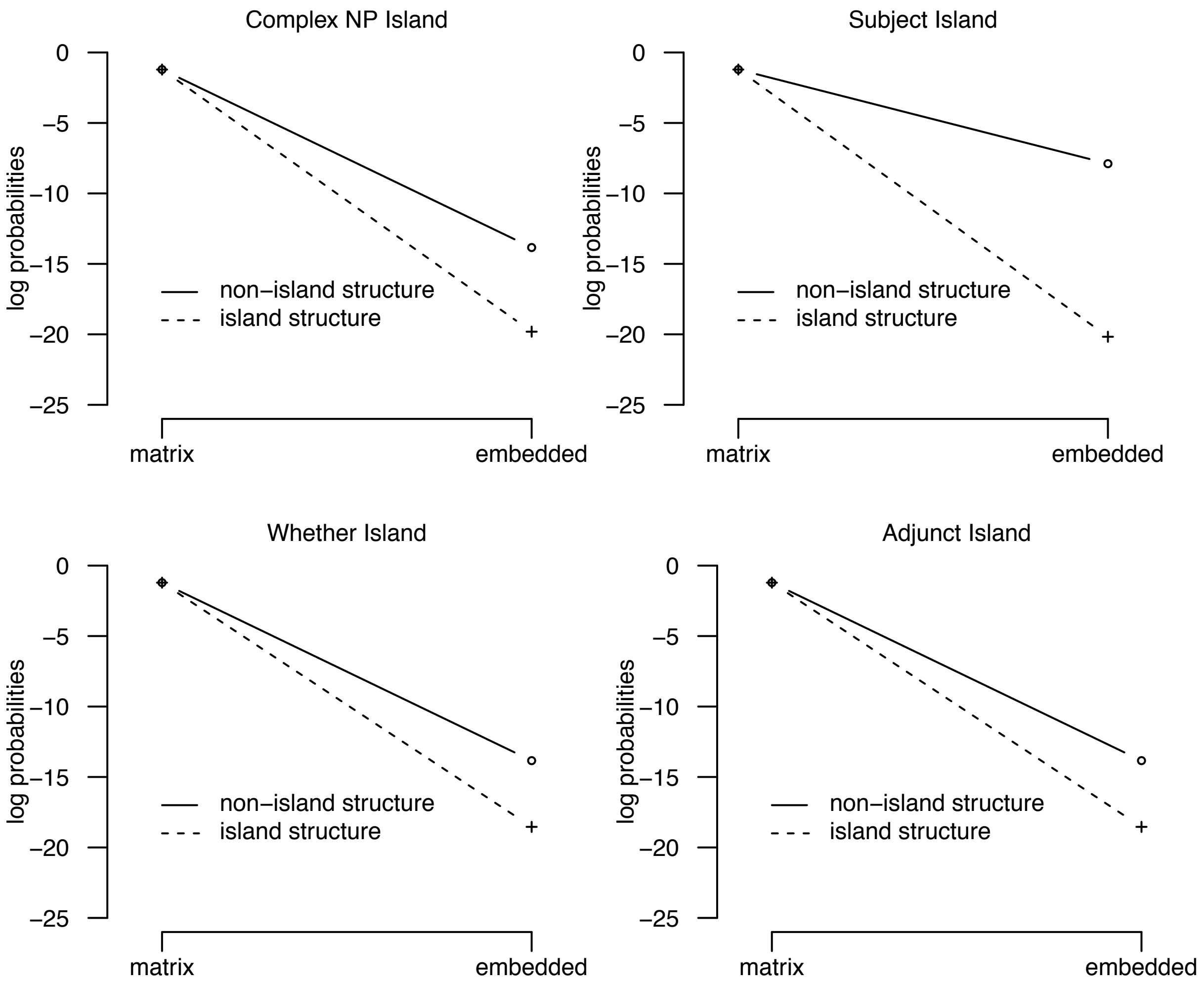


Pearl & Sprouse submitted

Learning results

Superadditivity observed for all four islands:

This learner has knowledge of these syntactic islands!



Pearl & Sprouse submitted

Proposed learning biases

Only one learning bias is potentially both innate and domain-specific.

	Innate	Derived	Domain-specific	Domain-general
Attend to container nodes	?	?	*	
Extract container node trigrams	*			*
Update trigram probabilities	*			*
Calculate dependency probability from trigrams	*			*

Pearl & Sprouse submitted

Container nodes

What kind of bias is this?

Identifying container nodes
- applies to language data: domain-specific
- derived from ability to parse utterances

Pearl & Sprouse submitted

Container nodes

What kind of bias is this?

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

Pearl & Sprouse submitted

Specifying CP container nodes

What kind of learning bias is this?

About a linguistic representation: domain-specific

Innate vs. derived?

Pearl & Sprouse submitted

Specifying CP container nodes

What kind of learning bias is this?

About a linguistic representation: domain-specific

Innate vs. derived?

- Could be specified innately

Pearl & Sprouse submitted

Specifying CP container nodes

What kind of learning bias is this?

About a linguistic representation: domain-specific

Innate vs. derived?

- Could be specified innately

- Could be derived from prior linguistic experience:

  - Uncontroversial to assume children learn to distinguish different types of CPs since the lexical content of CPs has substantial consequences for the semantics of a sentence.

  - Also, adult speakers are sensitive to the distribution of *that* versus null complementizers (Jaeger 2010).

Pearl & Sprouse submitted

Main implications of this learner

(1) Even though there is an induction problem for these syntactic islands, it may not require Universal Grammar learning biases to solve it.

	Innate	Derived	Domain-specific	Domain-general
Attend to container nodes	?	?	*	
Extract container node trigrams	*			*
Update trigram probabilities	*			*
Calculate dependency probability from trigrams	*			*

Pearl & Sprouse submitted

Main implications of this learner

(2) Even if a Universal Grammar learning bias is required, it is different from the biases previously proposed.

In particular, while it also specifies a particular linguistic representation, there is no bias defining the "theory". This falls out from the other non-UG learning biases.

	Innate	Derived	Domain-specific	Domain-general
Attend to container nodes	?	?	*	
vs.				
Attend to bounding nodes (BNs)	*		*	
Dependencies crossing 2+ BNs are not allowed	*		*	

Pearl & Sprouse submitted

Making an argument from acquisition

Universal Grammar: a theory of linguistic knowledge that is explicitly motivated by the induction problems during acquisition.

How to use computational methods effectively:

- Identify induction problems.
- Test learning strategies comprised of many learning biases to solve these induction problems.
- When these strategies work, examine the nature of the learning biases that define them.

Road map: Two good ways

- Informing theory: Arguments from acquisition
  - ✓ Investigating Universal Grammar

  - Testing theories of knowledge representation

- Informing the learning process: Useful, useable, and better than adults?
  - Comparing ideal and non-ideal approaches to discover how “less is more”

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.

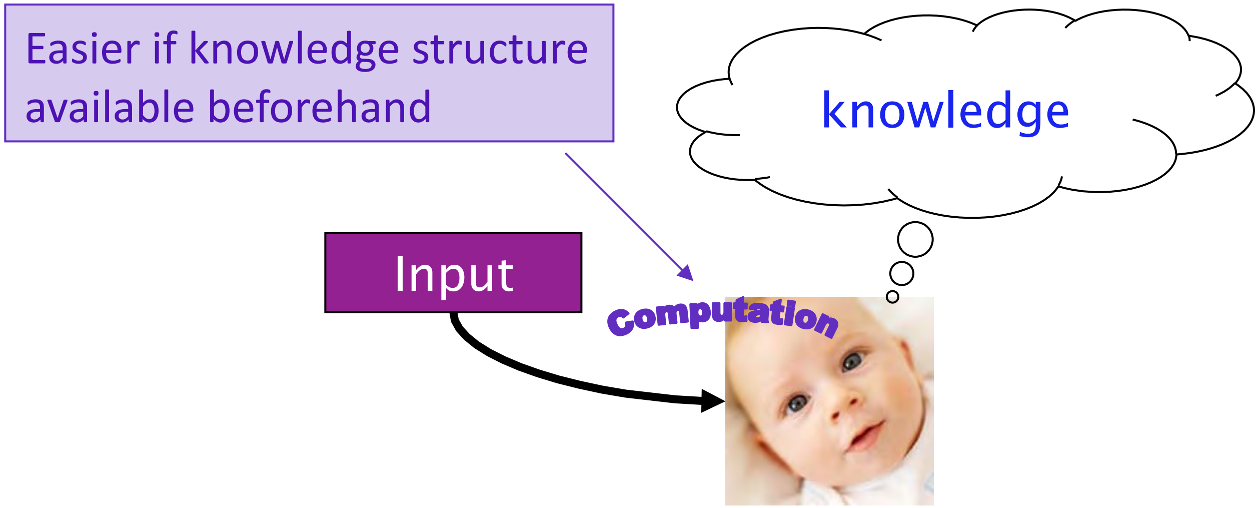
Argument from constrained cross-linguistic variation

Pearl 2011

Knowledge representation motivations

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

Argument from acquisition



Easier if knowledge structure available beforehand

knowledge

Input

Computation

Pearl 2011

Knowledge representation motivations

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

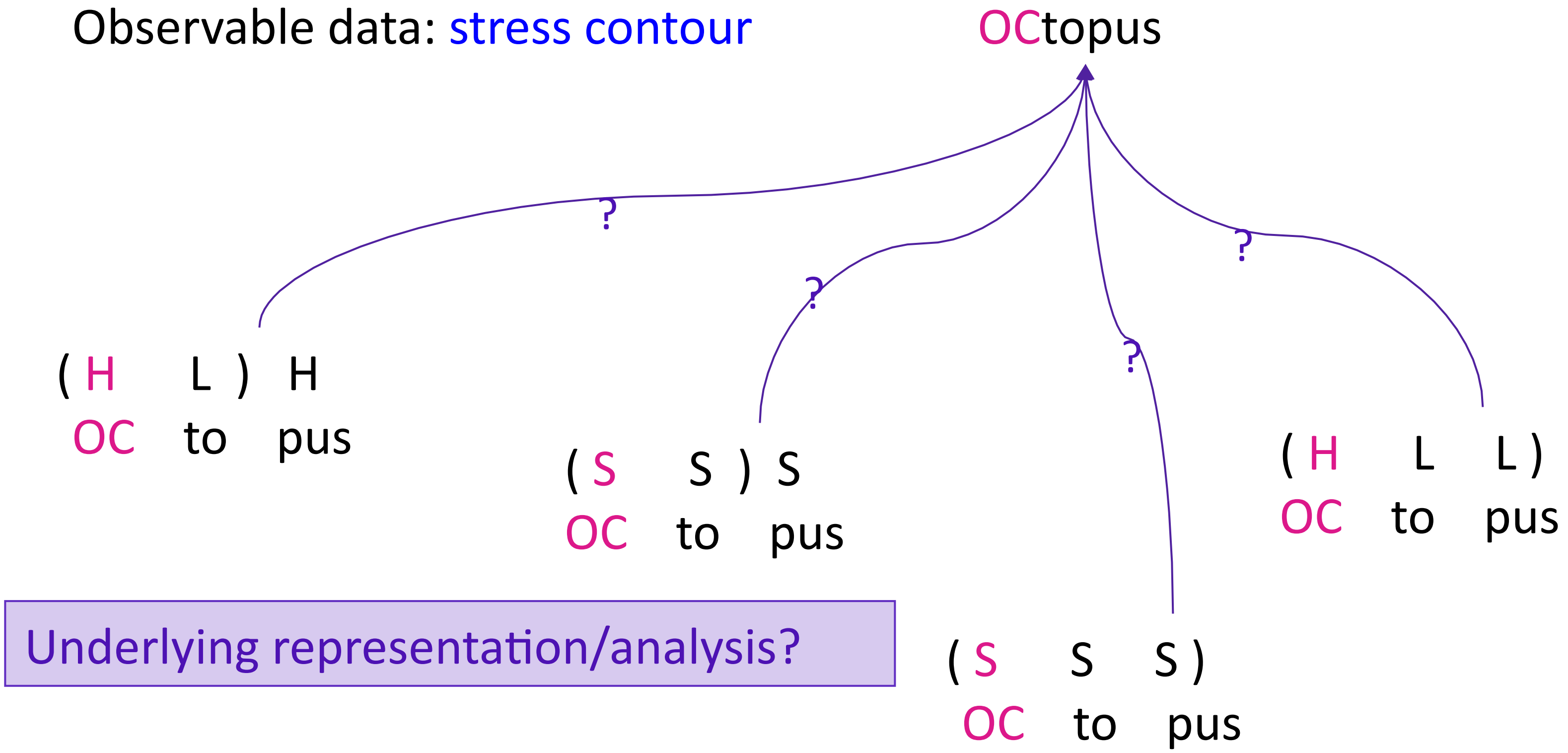
Argument from acquisition

- Using computational and quantitative methods along with available empirical data, we can explicitly test different proposals for knowledge representation.

Pearl 2011

Case study:
A generative system of metrical phonology

Observable data: stress contour



OCtopus

?

(H L) H

OC to pus

?

(S S) S

OC to pus

?

(S S S)

OC to pus

?

(H L L)

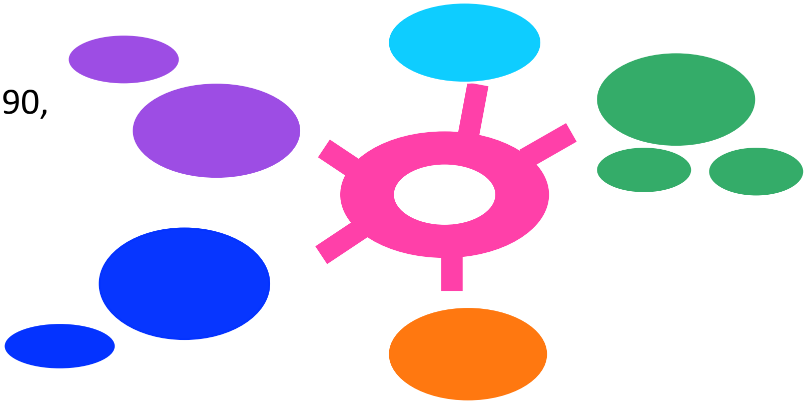
OC to pus

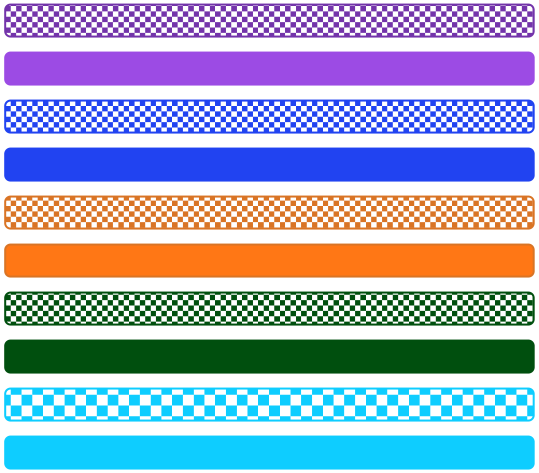
Underlying representation/analysis?

Pearl 2011

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

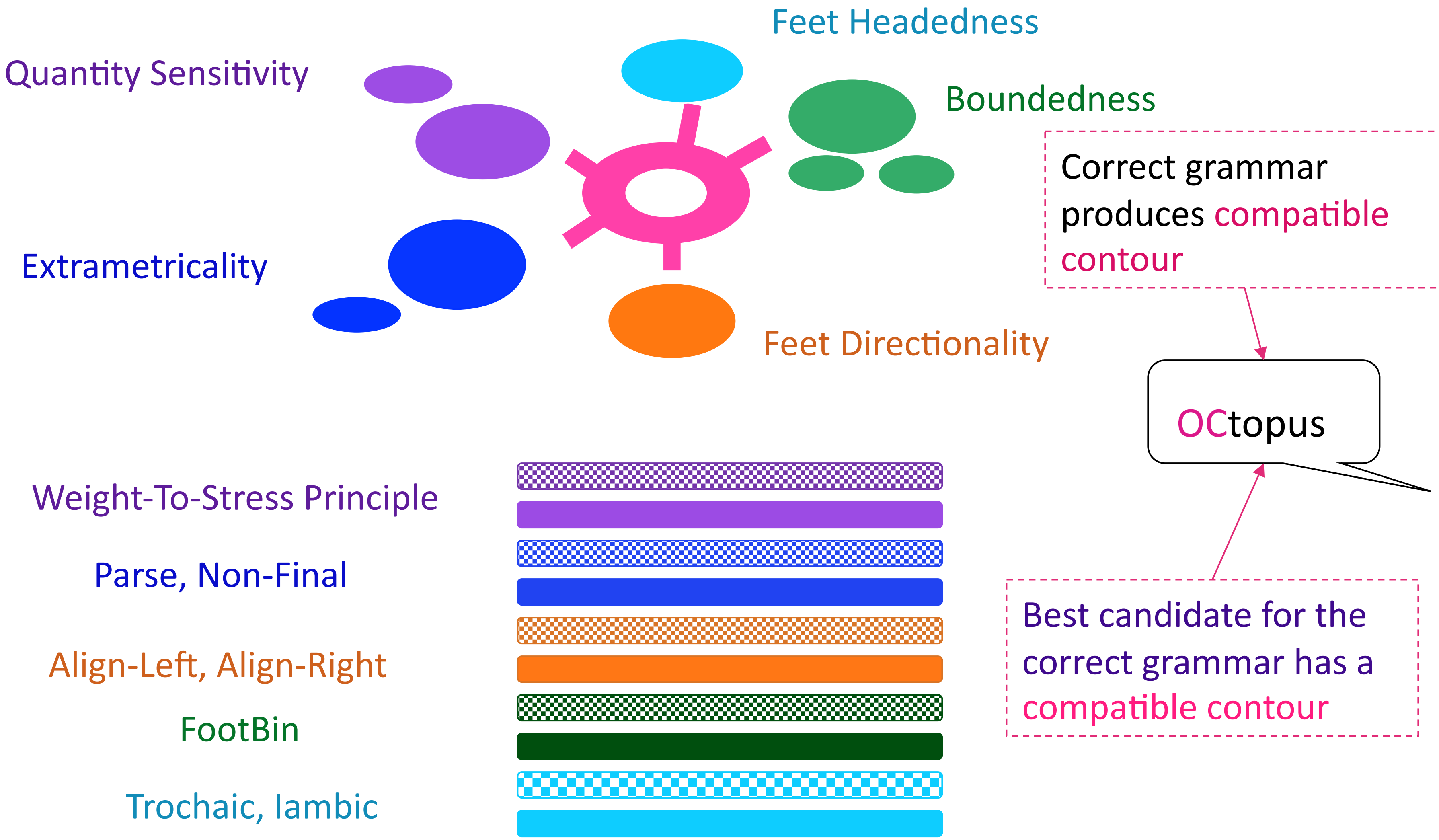




- Optimality theoretic system: 10 constraints (Hammond 1999, Prince & Smolensky 1993, Tesar & Smolensky 2000)
- Hypothesis space: 10! grammars (3,628,800)

Pearl 2011

Comparing knowledge representations



Feet Headedness

Quantity Sensitivity

Boundedness

Correct grammar produces compatible contour

Extrametricality

Feet Directionality

OCtopus

Weight-To-Stress Principle

Parse, Non-Final

Align-Left, Align-Right

FootBin

Trochaic, Iambic

Best candidate for the correct grammar has a compatible contour

Pearl 2011

Non-trivial language: English

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

OCtopus ?

or

or

- This is generally a problem for acquisition.

Pearl 2011

Non-trivial language: English

- 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.

Pearl 2011

Non-trivial language: English

- 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.

Analysis of child-directed speech (8 -15 months) from Brent corpus (Brent & Siskind 2001) from CHILDES (MacWhinney 2000): 504,084 tokens, 7390 types

For words with 2 or more syllables:

- 174 unique syllable-rime type combinations (ex: closed-closed (VC VC))

Pearl 2011

Non-trivial language: English

- 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.

Analysis of child-directed speech (8 -15 months) from Brent corpus (Brent & Siskind 2001) from CHILDES (MacWhinney 2000): 504,084 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*

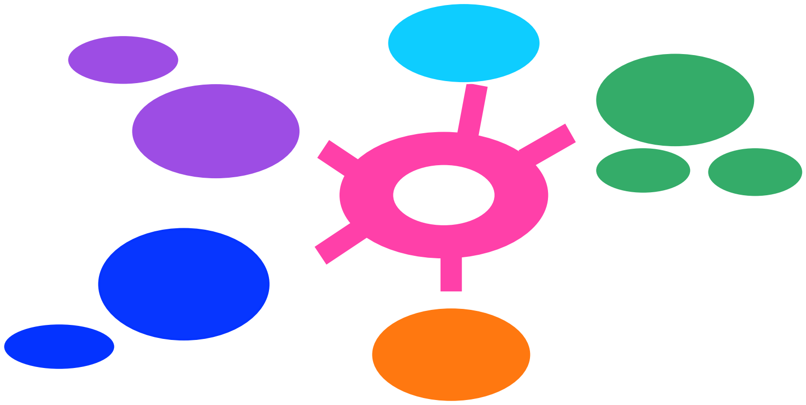
  AN swer

  SOME WHERE

Pearl 2011

Cognitively inspired learners using parameters

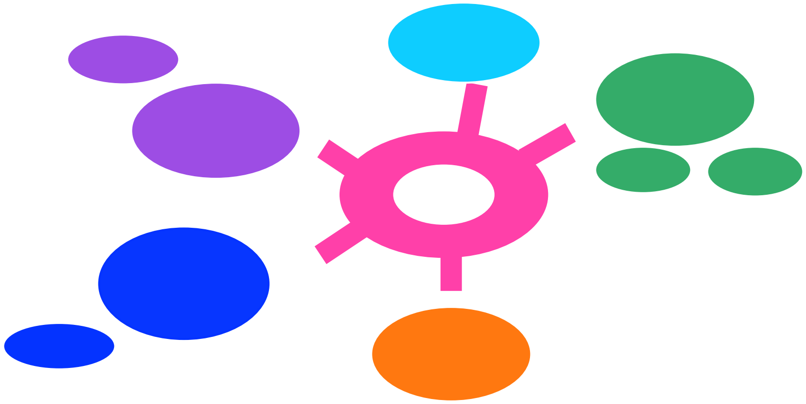
- Target state = grammar for English (Halle & Vergnaud 1987, Dresher & Kaye 1990, Dresher 1999) derived from cross-linguistic variation and adult linguistic knowledge



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

Pearl 2011

Cognitively inspired learners using parameters



- Only one cognitively plausible learner of the many variants tried was ever successful at converging on the adult English grammar when given realistic child-directed input, and then only once every 3000 runs! This seemed like very poor performance.

Pearl 2011

Where the problem lies

Premise: The 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 with empirical data).

Is this true?

Pearl 2011

Where the problem lies

- English grammar compatibility with data:
  - Generates contours matching 73.0% observable data tokens (62.1% types)
  - Note: not expected to be at 100% because of irregularities in English data

- Average compatibility of grammars selected by cognitively plausible learners using realistic input:
  - 73.6% by tokens (63.3% by types)

Pearl 2011

Where the problem lies

This isn’t true for the kind of data children encounter!

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

- English grammar compared to other 155 grammars in the hypothesis space
  - Ranked 52nd by tokens, 56th by types
  - English grammar is barely in the top third - unsurprising that modeled learners rarely select this grammar, given the child-directed speech data!

Pearl 2011

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

What about a child learner using the OT knowledge representation?

OT system test

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

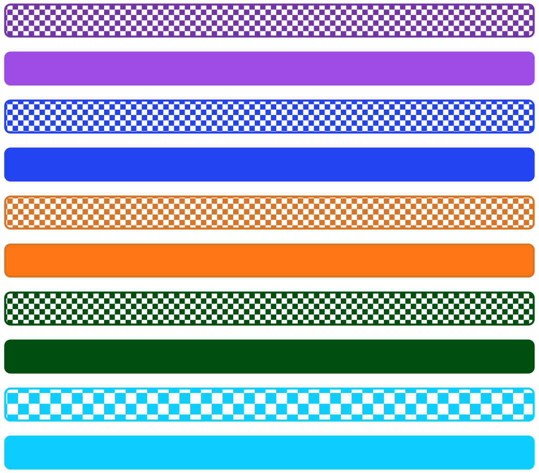
Weight-To-Stress Principle: VV, VC

Parse, Non-Final

Align-Left, Align-Right

FootBin: syllables, moras

Trochaic, Iambic



OT system test

- Adult English grammar (Hammond 1999, Pater 2000):
  - Combination of constraint orderings, such as 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

	C1	C2	C3	C4
(OC to) pus			*	*
oc (TO pus)	*		*	
(oc TO) pus		*	*	

Parameters vs. OT comparison

	Parameters	OT
Grammars in hypothesis space	156	3,628,800
Best grammar type compatibility	70.3%	67.5%
% of hypothesis space (best) English grammar scores lower than [types]	31.1%	34.8%
(Best) English grammar compatibility [types]	62.1%	26.6%

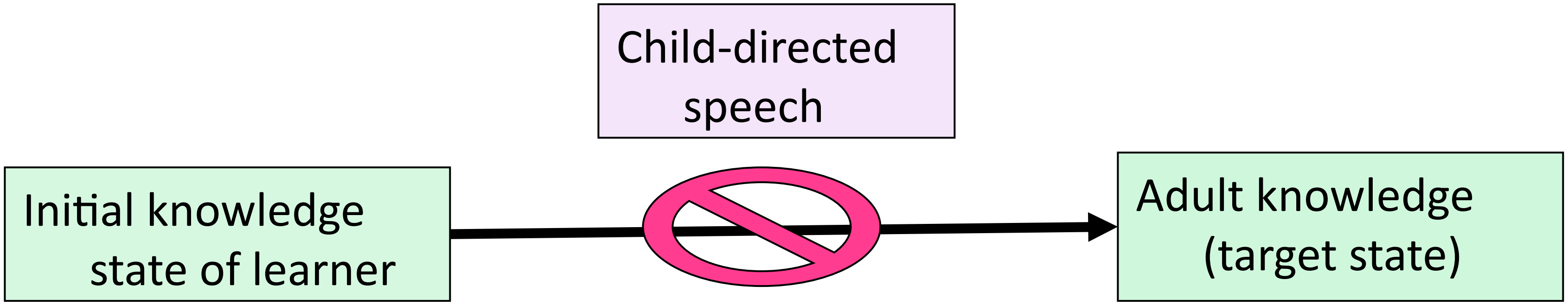
Comparable, except the best English grammar compatibility is very low for OT, compared to the English grammar in the parametric system. Also, the hypothesis space size is much larger for OT.

Problem for both learners

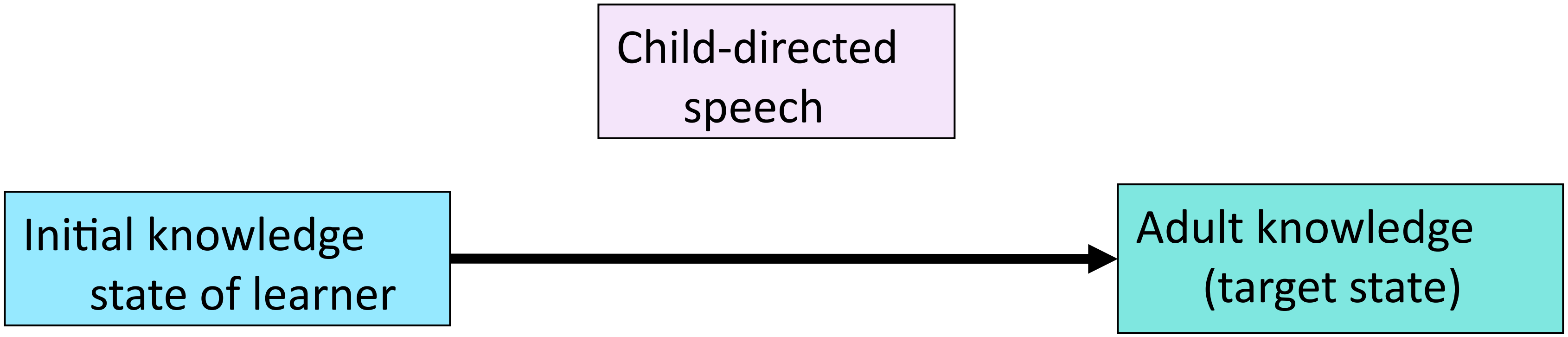
- Parametric child learner has a learnability problem: can't get to adult target state given the data available to children

- 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

Getting out of the learnability problem: 3 options


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

Option 1: Change the initial state & the target state


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

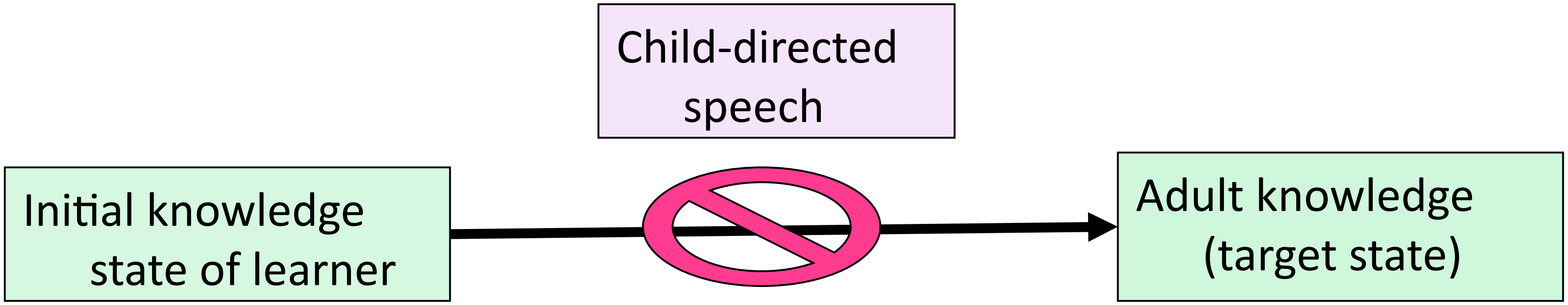
Pearl 2011

A different initial & target state for knowledge

Theoretical + computational/quantitative investigations for metrical phonology: Perhaps different parameters, constraints, or other representations make the adult English grammar more acquirable from child-directed speech (ex: Hayes 1995, Heinz 2007)

Pearl 2011

Getting out of the learnability problem: 3 options

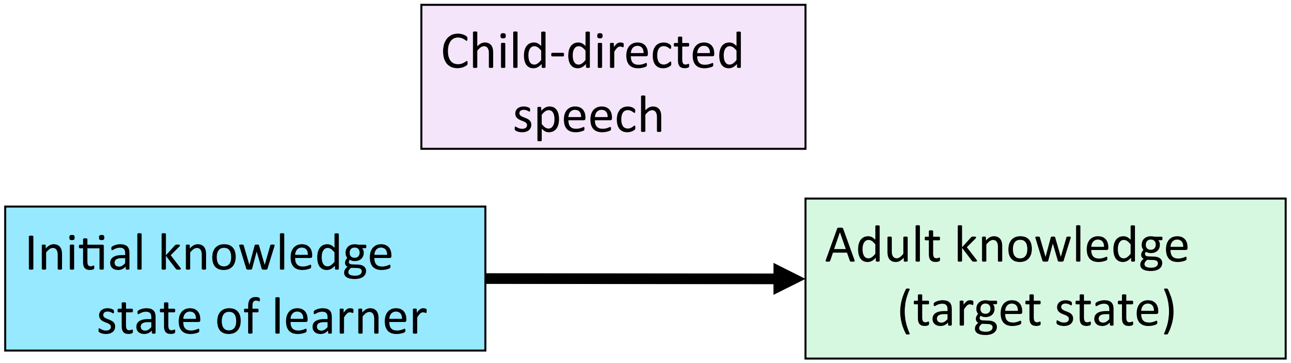


Child-directed speech

Initial knowledge state of learner

Adult knowledge (target state)

Option 2: change the initial state



Child-directed speech

Initial knowledge state of learner

Adult knowledge (target state)

Pearl 2011

A different (richer) initial state for learning

- Maybe young children have additional boosts from useful learning biases
  - Pearl 2008 (computational): learners biased to learn only from unambiguous data can learn the parametric system examined here from child-directed speech data, as long as the parameters are set in a particular order.

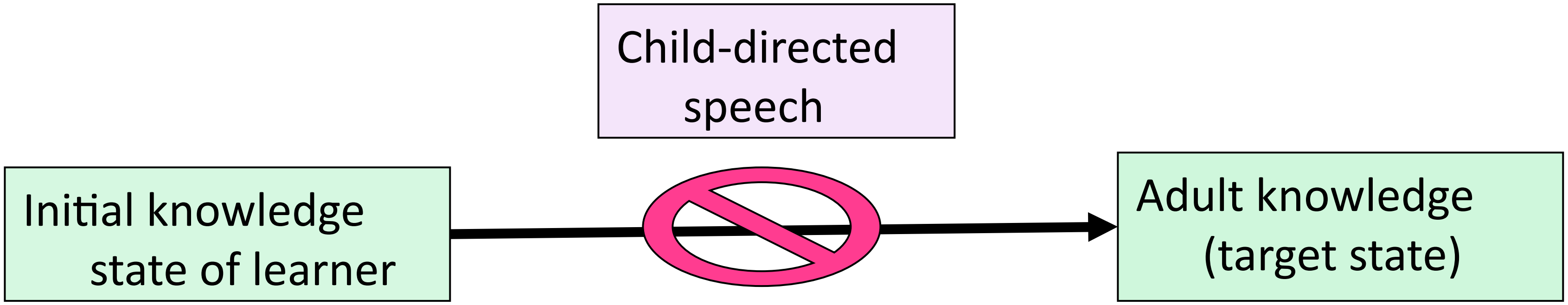
Pearl 2011

A different (richer) initial state for learning

- Maybe young children have additional boosts from useful learning biases
  - Pearl 2008 (computational): learners biased to learn only from unambiguous data can learn the parametric system examined here from child-directed speech data, as long as the parameters are set in a particular order.
  - Required learning biases at the initial state:
    - Use unambiguous data (and have a method for identifying these data for each parameter value)
    - Follow parameter-setting order constraints (and potentially have a method for deriving these constraints)

Pearl 2011

Getting out of the learnability problem: 3 options

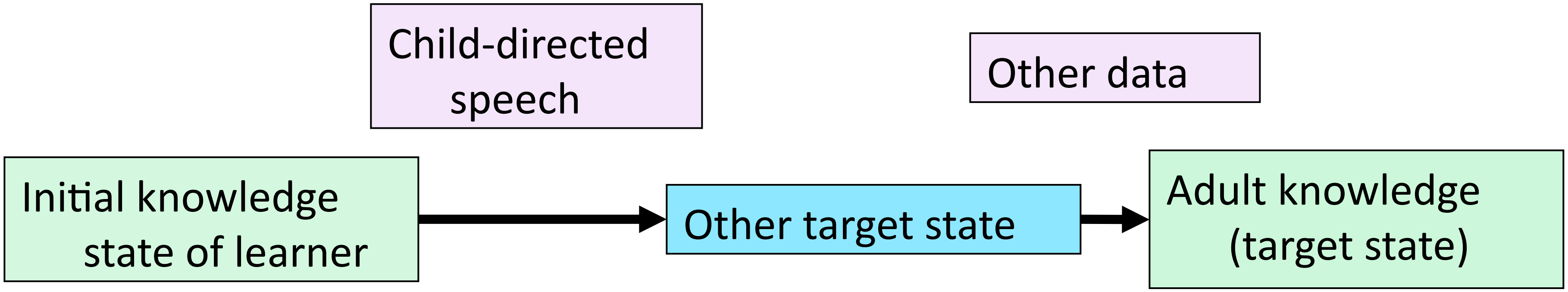


Child-directed speech

Initial knowledge state of learner

Adult knowledge (target state)

Option 3: change the (immediate) target state



Child-directed speech

Other data

Initial knowledge state of learner

Other target state

Adult knowledge (target state)

Pearl 2011

The learning trajectory: Knowledge change over time

- Idea: These knowledge representations are fine. It's just that there's an intermediate target state.

- Maybe young children don't acquire the adult English grammar until later, after they are exposed to more word types and realize the connection between stress contour and the English morphological system (connection to English morphological system: Chomsky & Halle 1968, Kiparsky 1979, Hayes 1982)

Brown 1973: morphological inflections not used regularly till 36 months

Pearl 2011

The learning trajectory: Knowledge change over time

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.

Experimental support: elicitation task with English 34-month-olds used items that were compatible with the parametric grammars modeled learners often chose here (Kehoe 1998) .

Pearl 2011

Making arguments from acquisition

Different theoretical proposals can be motivated and tested via computational and quantitative methods + empirical child-directed speech data

At the same time, we may need to draw on experimental work to make sure children are acquiring these representations when we think they are.

Road map: Two good ways

- Informing theory: Arguments from acquisition
  - ✓ Investigating Universal Grammar
  - ✓ Testing theories of knowledge representation
- Informing the learning process: Useful, useable, and better than adults?
  - Comparing ideal and non-ideal approaches to discover how “less is more”

Investigating learning strategies

For any potential strategy:

Is it useful?

What is possible to learn from the available data?

- Ideal/rational models, computational level approach
- What data representations are useful? What assumptions are useful?

Investigating learning strategies

For any potential strategy:

Is it useful?

Is it useable?

What is possible for children to learn from the available data?

- Constrained/process models, algorithmic level approach

- Are these representations and assumptions still useful if cognitive resources are limited?

Investigating learning strategies

For any potential strategy:

Is it useful?

Is it useable?

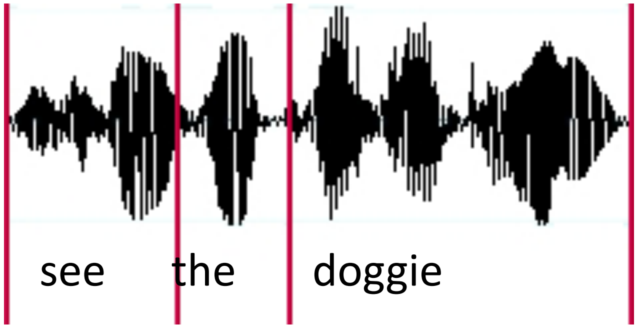
Does it work better when cognitive resources are constrained?

"Less is more" hypothesis of Newport (1990): Children do better precisely because they have more limited cognitive abilities.

- Also adults (sometimes) when their abilities are inhibited (Cochran et al. 1999, Kersten et al. 2001 but see Perfors 2011)

- What learning strategies have this property?

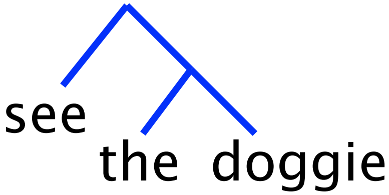
Case study:
Word segmentation



- A big deal: basis for more complex linguistic knowledge

SEE the DOGgie

phonology

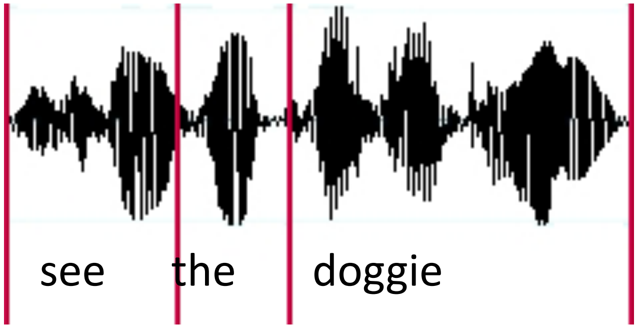


syntax

see’(the doggie)($x_{listener}$)

semantics

Case study:
Word segmentation



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

whatsthat
thedoggie
yeah
wheresthedoggie

→

whats that
the doggie
yeah
wheres the doggie

Word segmentation strategies

- Language-dependent cues: phonotactics, allophonic variation, metrical (stress) patterns, effects of coarticulation

Problem: Since these vary cross-linguistically, need to know some words in the language to figure them out. But these cues are used to help identify words in the first place…

Word segmentation strategies

- Language-independent cue: probability of sequences of units like phonemes or syllables

- Potential: Early bootstrapping
  - Thiessen & Saffran 2003: statistical information used very early

Bayesian inference:
A strategy that can use sequence probabilities

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

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}}\,\underbrace{P(h)}_{\text{prior}}$$

- Ideal learner: Is this information useful?
- Constrained learner: Is this information useable? Is there any evidence it's better when the learner is constrained?

Bayesian segmentation

(Goldwater et al. 2009)

Data: unsegmented corpus (transcriptions)

Hypotheses: sequences of word tokens

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}} \underbrace{P(h)}_{\text{prior}}$$

= 1 if concatenating words forms corpus,

= 0 otherwise.

Corpus: “lookatthedoggie”

P(*d*|*h*) =1	P(*d*|*h*) = 0
loo k atth ed oggie	*i like penguins*
lookat thedoggie	*look at thekitty*
look at the doggie	*a b c*

Bayesian segmentation

(Goldwater et al. 2009)

Data: unsegmented corpus (transcriptions)

Hypotheses: sequences of word tokens

Optimal solution is the segmentation with highest posterior probability.

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}} \underbrace{P(h)}_{\text{prior}}$$

= 1 if concatenating words forms corpus,

= 0 otherwise.

Encodes assumptions or biases in the learner:

- prefer short words
- prefer fewer words

Bayesian segmentation:
Ideal vs. Constrained

Learner assumptions:

- Basic unit of representation = phoneme
- Words are either independent units (unigram assumption)

  or

  Words are units that predict other words (bigram assumption)

Pearl, Goldwater, & Steyvers 2011

Bayesian segmentation: Ideal vs. Constrained

Bayesian learners examined:

Ideal	Constrained
perfect memory	decaying memory
large processing capabilities	limited processing capabilities
batch data processing	incremental data processing

Pearl, Goldwater, & Steyvers 2011

Bayesian segmentation:
Ideal vs. Constrained

Find a "less is more" effect for some constrained learners who have a unigram assumption, learning from English data.

Correct token identification: 64% constrained vs. 54% ideal

Why?

Their cognitive limitations caused them not to notice frequently occurring predictable sequences of short words like "at the". So, they didn't try to make them one word ("atthe") – an undersegmentation error that the ideal learners often made.

Pearl, Goldwater, & Steyvers 2011

Bayesian segmentation:
Ideal vs. Constrained

Cognitive plausibility: Make the learning process we're modeling look more like the learning process children are using.

Maybe we should revisit some of our modeling assumptions:

Basic unit of representation = phoneme?

Phillips & Pearl 2012, in prep

Perceptual units for infants

Word segmentation timeline:

Statistical learning at the beginning of segmentation, before 7.5 months

What representations do infants have at this point?

- Phonemes around ~10 months (Werker & Tees 1984)
- Syllables around 3 months (Eimas 1999, Jusczyk & Derrah 1987)

Phillips & Pearl 2012, in prep

Bayesian segmentation:
Ideal vs. Constrained

Updated learner assumptions:

- Basic unit of representation = syllable
- Words are either independent units (unigram assumption)

  or

  Words are units that predict other words (bigram assumption)

Phillips & Pearl 2012, in prep

Bayesian learners

Ideal learner:

- Process data in a batch (perfect memory)
- Have enough processing resources to exhaustively search potential segmentations
- Select optimal segmentation

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Dynamic Programming + Maximization [DPM]):

- Process data incrementally
- Have enough processing resources to exhaustively search potential segmentations
- Select optimal segmentation

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Dynamic Programming + Sampling [DPS]):

- Process data incrementally
- Have enough processing resources to exhaustively search potential segmentations
- Select segmentation probabilistically

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Decayed Markov Chain Monte Carlo [DMCMC]):

- Process data incrementally
- Have limited processing resources and decaying memory, so cannot do exhaustive search
- Select segmentation probabilistically

Phillips & Pearl 2012, in prep

Bayesian learning over syllables

Word token F-scores

	Unigram	Bigram
Ideal	53.1	77.1
DPM	58.8	75.1
DPS	63.7	77.8
DMCMC	55.1	86.3

$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

Recall:

#found / #true

Results averaged over 5 randomly generated test sets (~2800 utterances) that were separate from the training sets (~25200 utterances), all generated from the Pearl-Brent derived corpus.

Phillips & Pearl 2012, in prep

Bayesian learning over syllables

Word token F-scores

	Unigram	**Bigram**
Ideal	53.1	77.1
DPM	**58.8**	75.1
DPS	**63.7**	77.8
DMCMC	**55.1**	86.3

$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

Recall:

#found / #true

A learner who assumes words are not predictive of other words performs significantly better when its abilities are constrained.

More robust effect than Pearl et al. 2011 observed for unigram learner: All three constrained learners do better.

Phillips & Pearl 2012, in prep

Bayesian learning over syllables

Word token F-scores

	Unigram	Bigram
Ideal	53.1	77.1
DPM	**58.8**	75.1
DPS	**63.7**	77.8
DMCMC	**55.1**	**86.3**

$$F = \frac{2 * \text{Prec} * \text{Rec}}{\text{Prec} + \text{Rec}}$$

Precision:

#correct / #found

Recall:

#found / #true

One constrained learner who assumes words are predictive of other words performs significantly better than the ideal learner.

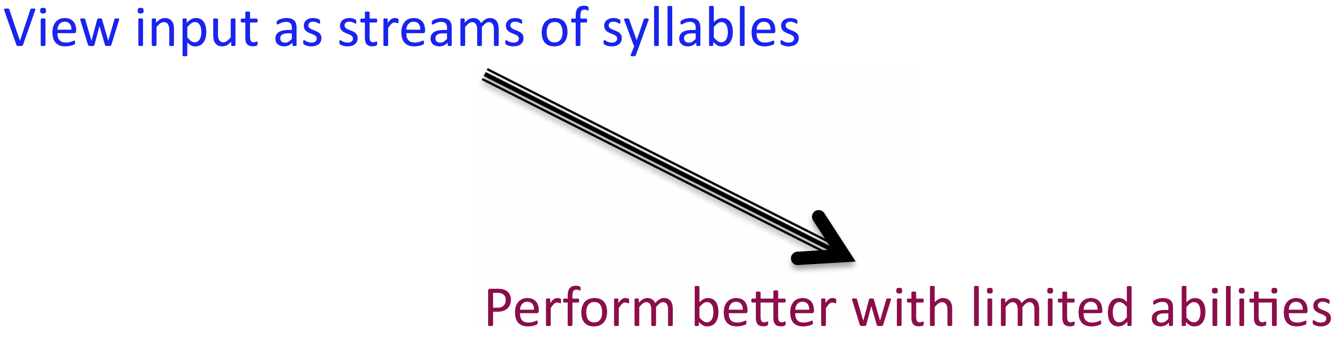
New effect: Pearl et al. 2011 did not observe this effect in bigram learners.

Phillips & Pearl 2012, in prep

The utility of cognitively plausible modeling assumptions

In learners with either the unigram or the bigram assumption, we find what looks like a “less is more” effect.

By trying to make the model represent the input the way we think children do, we have reproduced behavior that we think children have.



View input as streams of syllables

Perform better with limited abilities

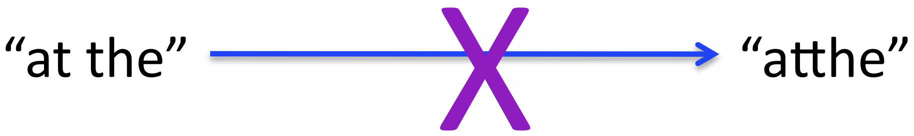
Phillips & Pearl 2012, in prep

What’s causing “less is more”?

Still under investigation, but…

Unigram learners could be benefiting in a similar way to the learners in Pearl et al. 2011:

Constrained learners don’t create the undersegmentation errors that ideal learners do for frequently occurring sequences of short words. (They don’t notice them as much.)



“at the”

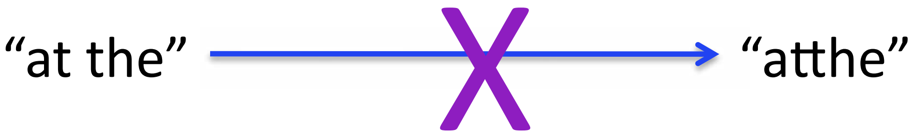
“atthe”

Phillips & Pearl 2012, in prep

What’s causing “less is more”?

Still under investigation, but…

Bigram learners wouldn’t make this error though, because they have a way to represent predictable sequences. But the DMCMC bigram learner is significantly outperforming the ideal bigram learner…



“at the”

X

“atthe”

Phillips & Pearl 2012, in prep

What’s causing “less is more”?

Still under investigation, but…

If we look at the recall scores for these bigram learners, we notice that token recall is higher for the DMCMC learner while lexicon recall (word types) is higher for the ideal learner.

	Token recall	**Lexicon recall**
Ideal Bigram	72.5	**79.7**
DMCMC Bigram	**85.5**	76.8

Phillips & Pearl 2012, in prep

What's causing "less is more"?

Still under investigation, but…

One interpretation: The constrained learner is correctly segmenting more frequent words (with more tokens per word) while the ideal learner is correctly segmenting more word types.

	Token recall	**Lexicon recall**
Ideal Bigram	72.5	**79.7**
DMCMC Bigram	**85.5**	76.8

Constrained learner does well on more "important" words that occur more often?

Phillips & Pearl 2012, in prep

Understanding the learning process

Case study: Bayesian inference as an initial strategy for word segmentation

✓ Is it useful?

Ideal learners using this strategy perform fairly well, given realistic child-directed speech data.

Phillips & Pearl 2012, in prep

Understanding the learning process

Case study: Bayesian inference as an initial strategy for word segmentation

✓ Is it useful?

✓ Is it useable?

Constrained learners can still use this strategy and do quite well.

Phillips & Pearl 2012, in prep

Understanding the learning process

Case study: Bayesian inference as an initial strategy for word segmentation

✓ Is it useful?

✓ Is it useable?

✓ Does it work better when cognitive resources are constrained?

By representing the input in a way infants are likely to do, we find a stronger "less is more" effect, with constrained learners outperforming ideal learners.

Phillips & Pearl 2012, in prep

Recap: Two good ways to use computational methods

Make arguments from acquisition for theory.

Identify learning strategies that are useful, useable, and can explain surprisingly superior child learning.

Computational methods

Theoretical methods

Experimental methods

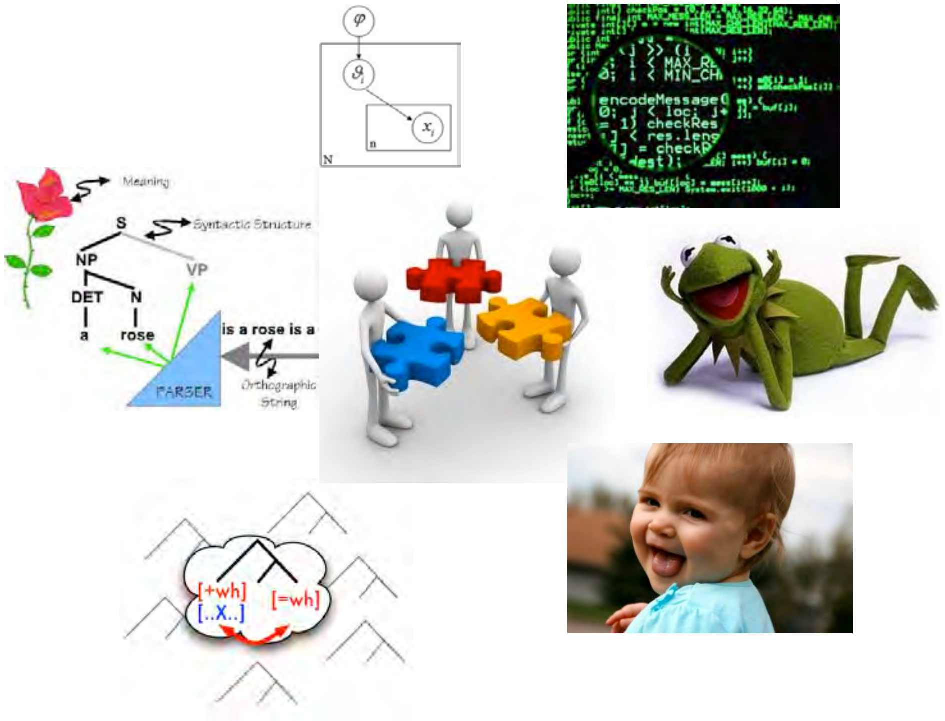
Thank you!

Jon Sprouse

Diogo Almeida Misha Becker Bob Berwick
Bob Frank Michael Frank Heather Goad
Norbert Hornstein Bill Idsardi Roger Levy
Amy Perfors Colin Phillips William Sakas
Amy Weinberg Charles Yang

Lawrence Phillips

Ivano Caponigro Alexander Clark
Sharon Goldwater Tom Griffiths
Jeff Lidz Diane Lillo-Martin
Mark Steyvers Virginia Valian



Computation of Language Laboratory

UC Irvine

Extra Material

Building a computational learner

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence's probability is the smoothed product of its trigrams.

What kind of bias is this?

- have enough memory to hold the utterance and its dependency in mind: innate and domain-general

- have enough memory to hold three units in mind (Mintz 2006, Wang & Mintz 2008, Saffran et al. 1996, Aslin et al. 1996, Saffran et al. 1999, Graf Estes et al. 2007, Saffran et al. 2008, Pelucchi et al. 2009a, 2009b): innate and domain-general

- track trigrams of units: innate and domain-general

Pearl & Sprouse submitted

Building a computational learner: Empirical grounding

Hart & Risley 1995: Children hear approximately 1 million utterances in their first three years.

Assumption: learning period for modeled learners is 3 years (ex: between 2 and 5 years old for modeling children's acquisition), 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)

Pearl & Sprouse submitted

OT system test

- 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:

  74.6% of data tokens (67.5% 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?

OT system test

- Grammars with higher compatibility than best English grammar:

  1,157,538 (token compatibility)

  1,263,130 (type compatibility)

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

	Parameters	OT
Grammars in hypothesis space	156	3,628,800
Best grammar compatibility	76.5% (tokens) 70.3% (types)	74.6% (tokens) 67.5% (types)

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

Parameters vs. OT comparison

	Parameters	OT
Grammars in hypothesis space	156	3,628,800
Best grammar compatibility	76.5% (tokens) 70.3% (types)	74.6% (tokens) 67.5% (types)
% of hypothesis space (best) English grammar scores lower than	28.3% (tokens) 31.1% (types)	31.9% (tokens) 34.8% (types)

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

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

However, the best English grammar compatibility is very low for OT, compared to the English grammar in the parametric system.

Bayesian learners

Ideal learner:

- Process data in a batch (perfect memory)
- Have enough processing resources to exhaustively search potential segmentations
- Select optimal segmentation

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Dynamic Programming + Maximization [DPM]):

- Process data incrementally
- Have enough processing resources to exhaustively search potential segmentations
- Select optimal segmentation

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Dynamic Programming + Maximization [DPM]):

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.

	did you wanna sit down
0.33	dId yu wa/n6 sIt dQn
0.21	dId/yu wa/n6 sIt dQn
0.15	dId/yu wa n6 sIt dQn
…	…

Bayesian learners

Constrained learner (Dynamic Programming + Sampling [DPS]):

- Process data incrementally
- Have enough processing resources to exhaustively search potential segmentations
- Select segmentation probabilistically

Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Dynamic Programming + Sampling [DPS]):

For each utterance:

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

		did you wanna sit down
	0.33	dId yu wa/n6 sIt dQn
→	0.21	dId/yu wa/n6 sIt dQn
	0.15	dId/yu wa n6 sIt dQn
	…	…

Bayesian learners

Constrained learner (Decayed Markov Chain Monte Carlo [DMCMC]):

- Process data incrementally
- Have limited processing resources and decaying memory, so cannot do exhaustive search
- Select segmentation probabilistically

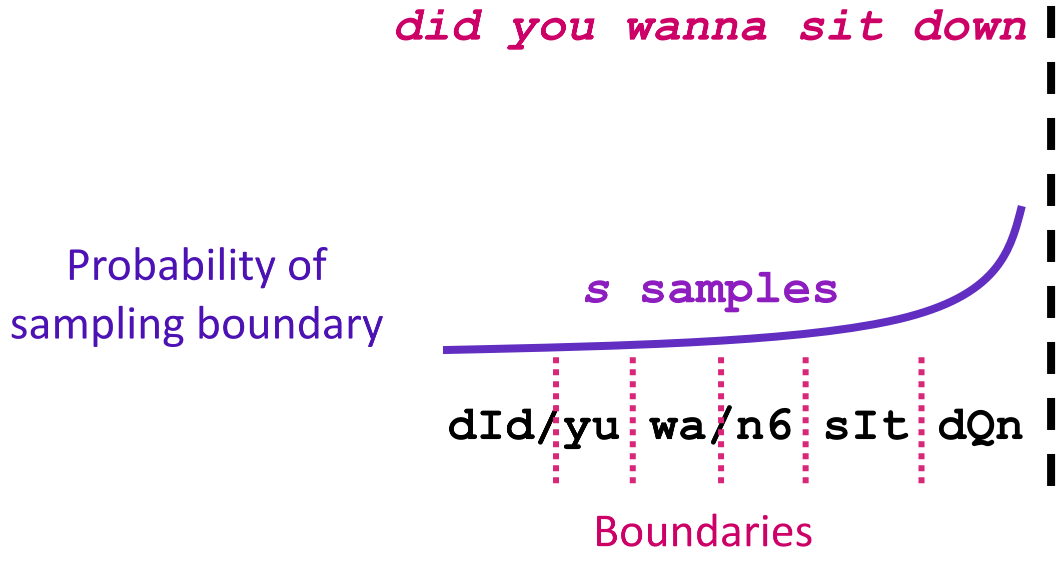
Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Decayed Markov Chain Monte Carlo [DMCMC]):

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 each boundary sample.



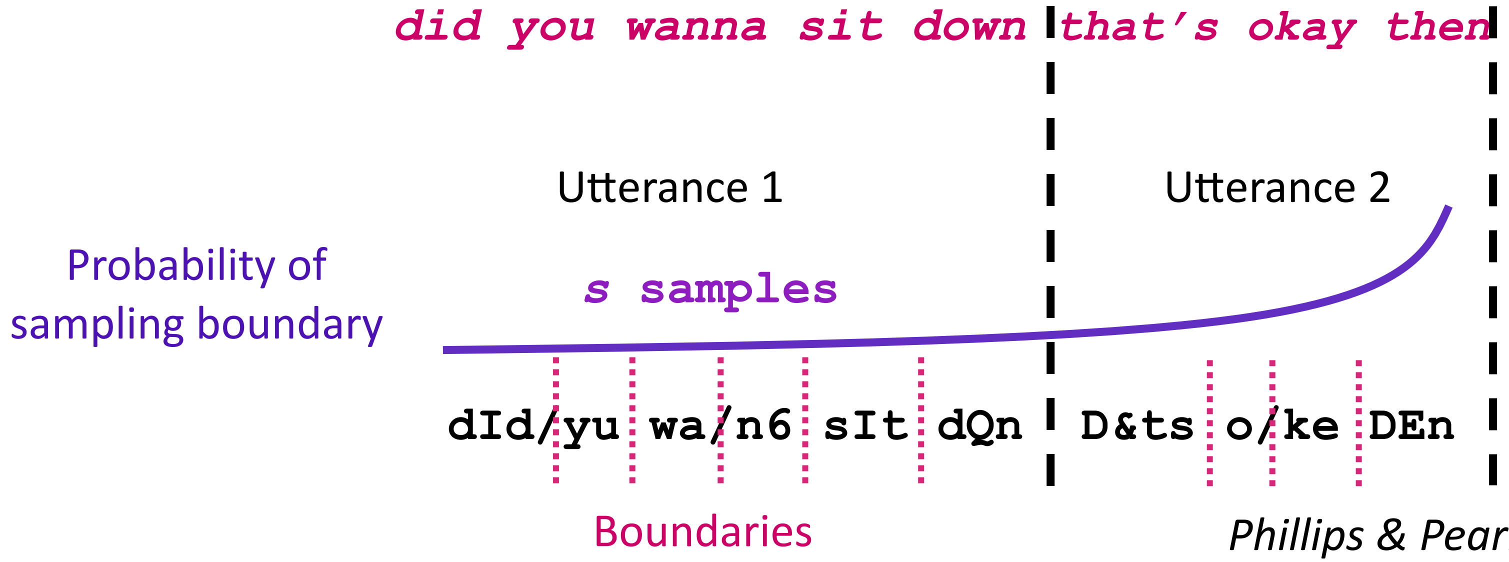
Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Decayed Markov Chain Monte Carlo [DMCMC]):

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 each boundary sample.



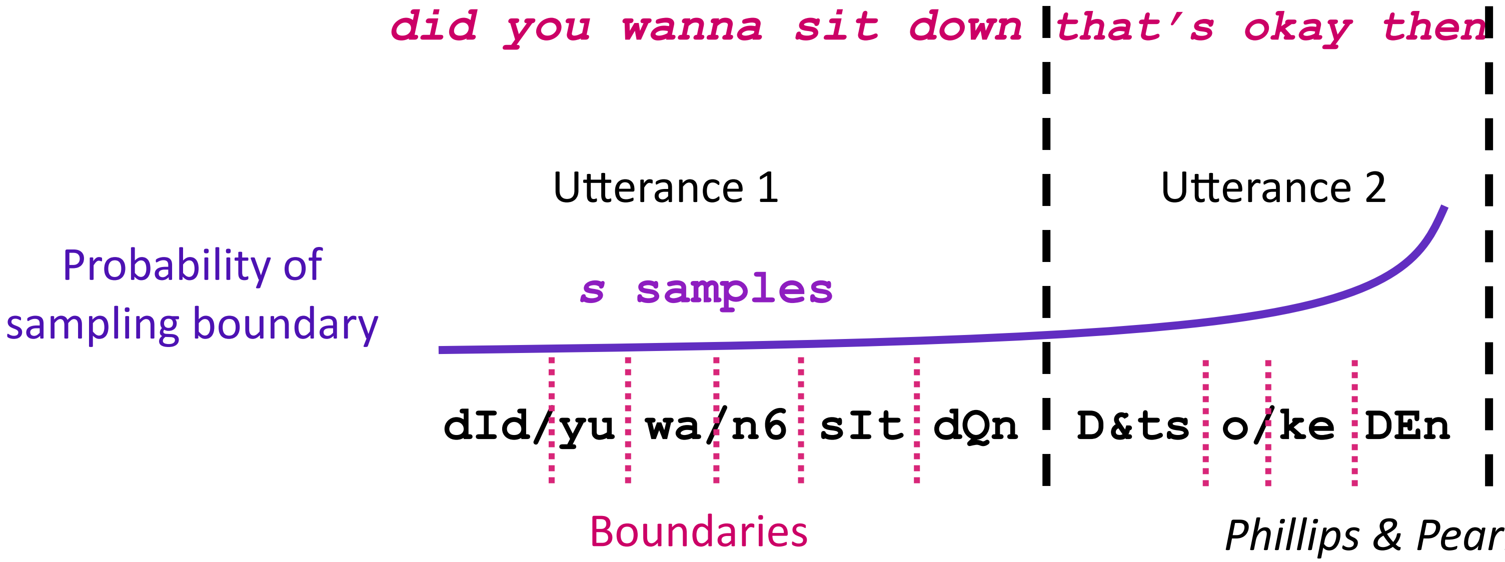
Phillips & Pearl 2012, in prep

Bayesian learners

Constrained learner (Decayed Markov Chain Monte Carlo [DMCMC]):

For all DMCMC learners:

d = 1.5 (~77% chance of sampling a boundary in the current utterance)
s = 20000 samples per utterance (78% fewer samples than ideal learner)



did you wanna sit down that's okay then

Utterance 1

Utterance 2

Probability of sampling boundary

***s* samples**

dId/yu wa/n6 sIt dQn D&ts o/ke DEn

Boundaries

Phillips & Pearl 2012, in prep

Learner input

- Pearl-Brent corpus (9 months or younger section)
  - 28,391 utterances of phonemically transcribed child-directed speech (96,920 tokens, 3,213 types), which was then syllabified.
  - Average utterance length: 3.4 words, 4.2 syllables

Example input:

```
dId/yu/sIt/dQn
dId/yu/wa/n6/sIt/dQn
D&ts/o/ke/DEn
kAm/h)
...
```

≈

```
did/you/sit/down
did/you/wa/nna/sit/down
thats/o/kay/then
come/here
...
```

Phillips & Pearl 2012, in prep