

CRISSP Lecture Series



A program for experimental syntax:
data, theory, and biology

Jon Sprouse
University of Connecticut

Biology 1:

Experimental Syntax and the **acquisition**
of island effects.

KU Leuven - Brussels 03.17.15

Ground rule 1: Be explicit about types of knowledge

There are at least two dimensions that matter for debates about language acquisition mechanisms: domain-specificity and innateness. Viewing these dimensions at two factors with two levels, we get four types of knowledge:

	domain-specific	domain-general
innate	<p>Nativists are ok with stuff being here.</p> <p>Constructivists/empiricists want this cell empty.</p>	<p>Everybody needs these</p>
derived	<p>Everybody needs these</p>	<p>I am not sure what would be good examples of these, but I don't think anybody would object</p>

Different types of innate, domain-specific knowledge

One way to look at this typology is to say that the red square (innate, domain-specific knowledge) is Universal Grammar. In fact, that is something that I like to say.

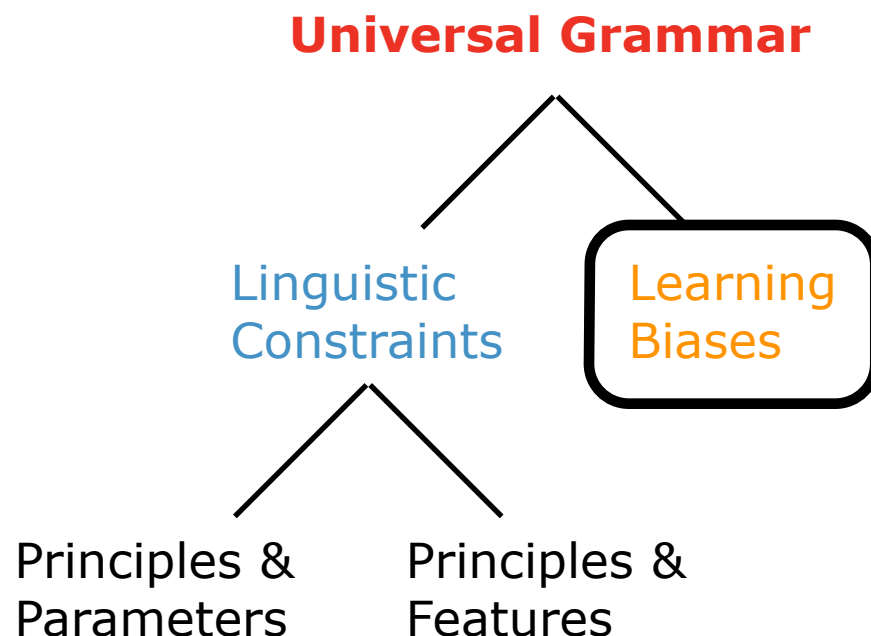
	domain-specific	domain-general
innate	Universal Grammar	
derived		

Different types of innate, domain-specific knowledge

But to be fair, there are any number of types of knowledge that can be innate and domain-specific. It is perhaps better to think of UG as a class of theories, with different instantiations that can vary based on the specific type of knowledge that is assumed to be innate:

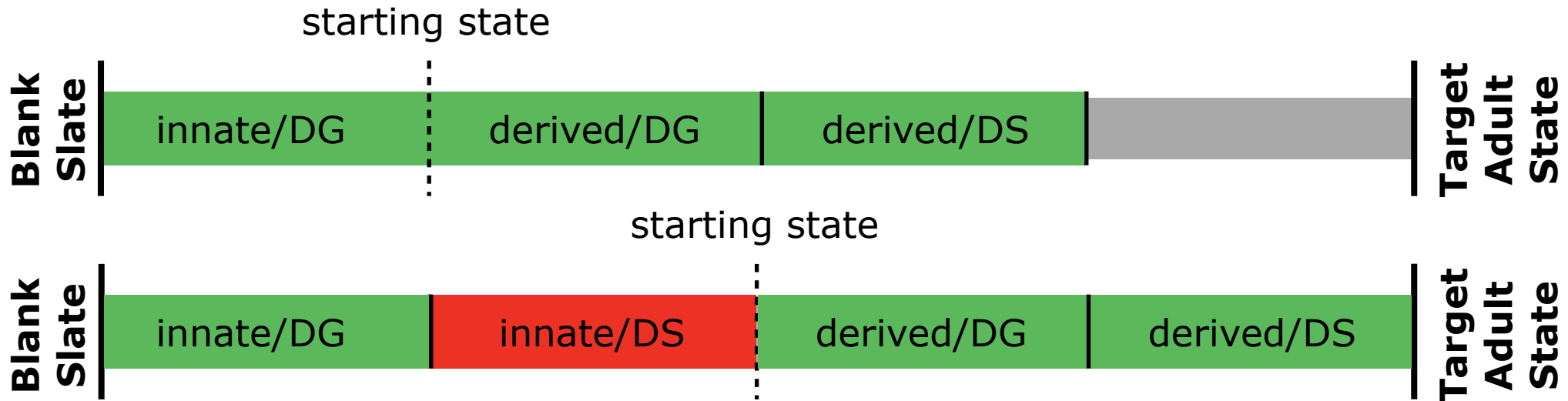
We can also sub-divide the UG class into theories that postulate innate linguistic constraints, and theories that postulate biases that allow for the construction of linguistic constraints.

Today I am going to focus on the bias-type of theory simply because I think it is more palatable to non-generative grammarians that might want to debate with.

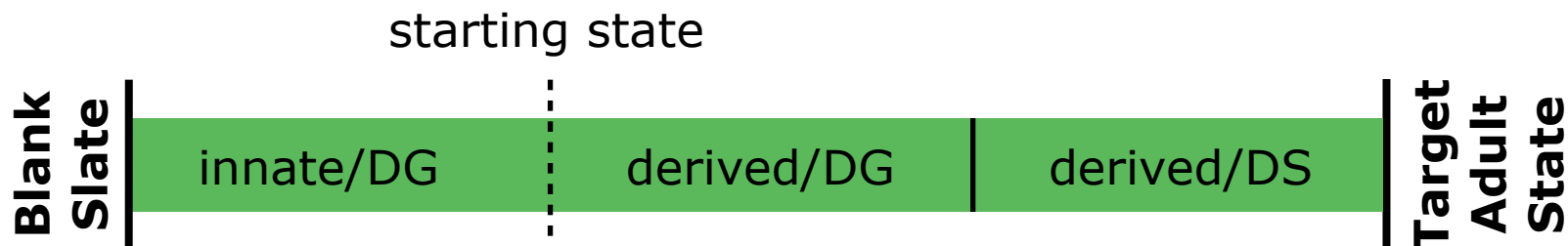


Ground rule 2: Target states can't vary

Debates about UG are debates about **starting states**. That means that the target state has to be held constant, like this:

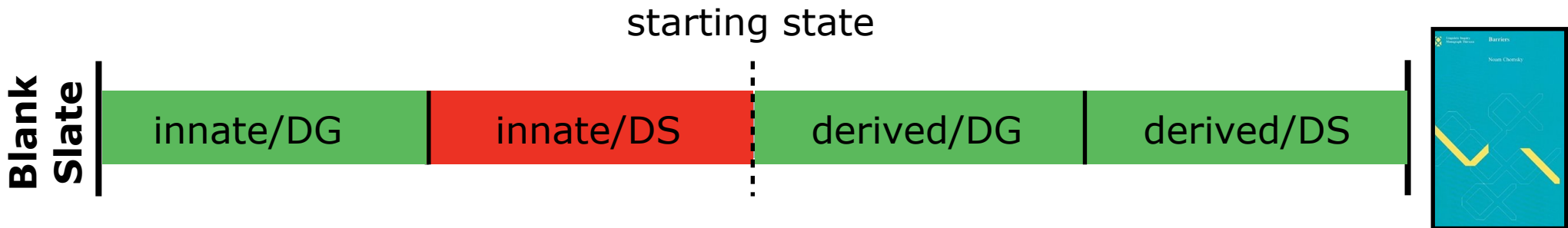


If you allow the target state to vary too, then of course it may be possible to learn "the target state" from a different starting state!

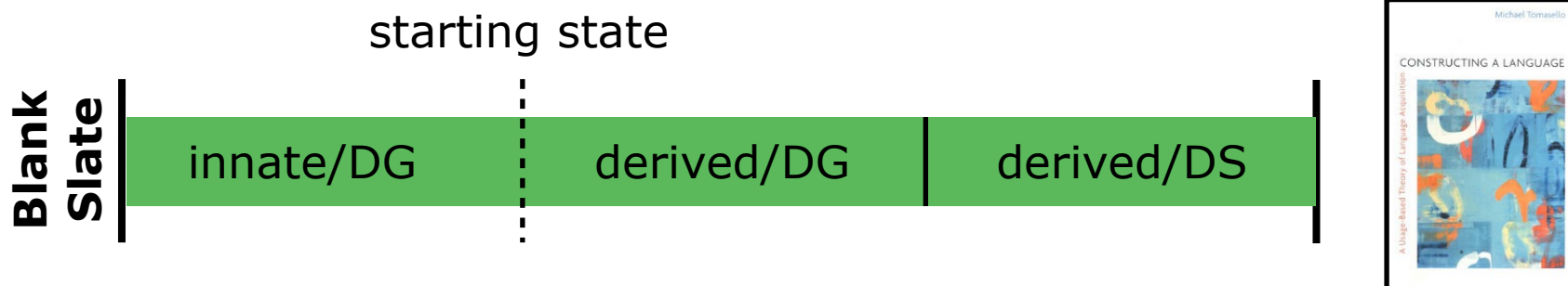


What does this mean in practice?

Linguists who believe in UG tend to assume that the target state is a relatively complex grammar (e.g., one that includes the Subjacency Condition or PIC). That is why innate/domain-specific knowledge is necessary to reach it.



Linguists who don't believe in UG tend to assume that the target state is a relatively simple grammar (e.g., one that does not include Subjacency/PIC). Therefore it is not surprising that they believe that the target state can be reached without innate/domain-specific knowledge.

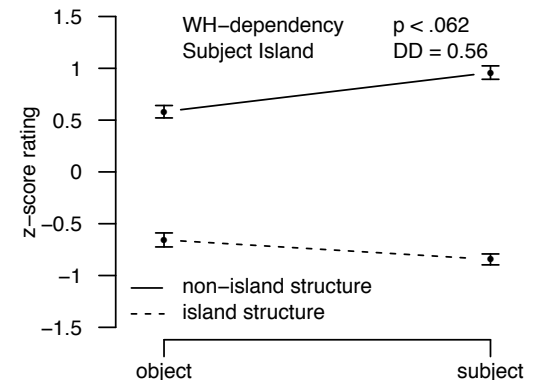
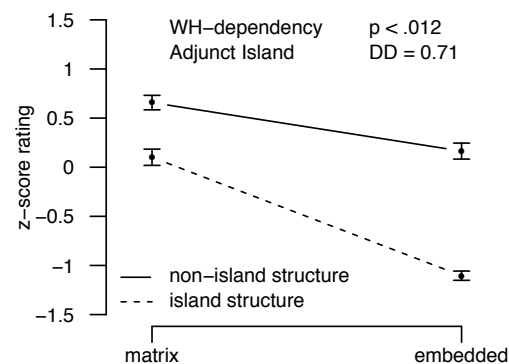
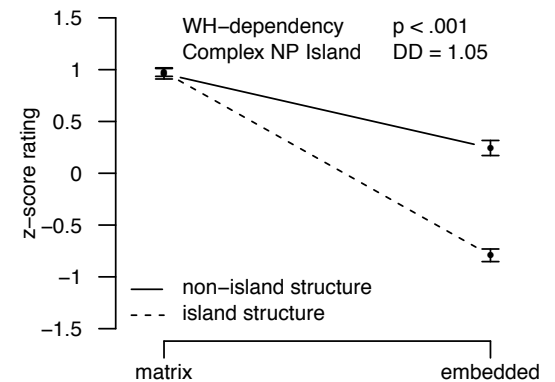
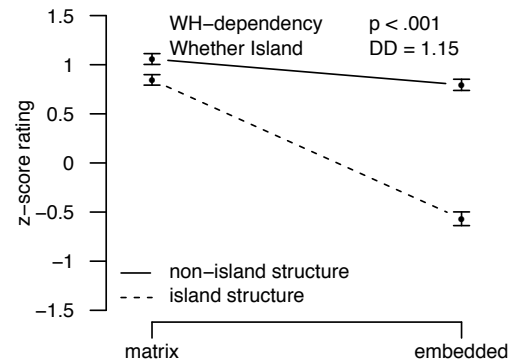


In order to have a conversation, we have to agree on the target state.

Let's use behavior

Since we are unlikely to get everybody to agree on a target state when we talk in terms of the grammar, another option is to define the target state as achieving the behavior that human adults exhibit. Such as acceptability judgments:

We can try to build a learner that acquires the super-additive pattern that we see when an island constraint is at work.



If the learner can achieve this behavior, then it is a possible theory of acquisition (though not necessarily correct).

Ground rule 3: Use realistic input

Debates often put a lot of focus on the type of learner:

Ideal learner:

An ideal learner has perfect knowledge of the input. That is, it knows all of the input simultaneously, and can manipulate properties of the input as necessary to achieve the desired target state. This entails that it has perfect memory, recall, parsing, etc. Ideal learners are like logical arguments: if an ideal learner can't learn a phenomenon (using certain mechanisms) then those mechanisms could not possibly be a solution to the learning problem.

Realistic learner:

A realistic learner attempts to better mimic the process of language learning in humans. This typically means input that is not perfectly parsed, and is not perfectly remembered. It often involve incremental learning. Realistic learners attempt to refine the hypothesis space to be closer to the truth of the universe.

Crucially, both types of learners must use realistic input in order for any claims to be made. Without realistic input, they have no inferential value.

Ground rule 3: Use realistic input

Debates often put a lot of focus on the type of learner:

Ideal learner:

An ideal learner has perfect knowledge of the input. That is, it knows all of the input simultaneously, and can manipulate properties of the input as necessary to achieve the desired target state. This entails that it has perfect memory, recall, parsing, etc. Ideal learners are like logical arguments: if an ideal learner can't learn a phenomenon (using certain mechanisms) then those mechanisms could not possibly be a solution to the learning problem.

Realistic learner:

A realistic learner attempts to better mimic the process of language learning in humans. This typically means input that is not perfectly parsed, and is not perfectly remembered. It often involve incremental learning. Realistic learners attempt to refine the hypothesis space to be closer to the truth of the universe.

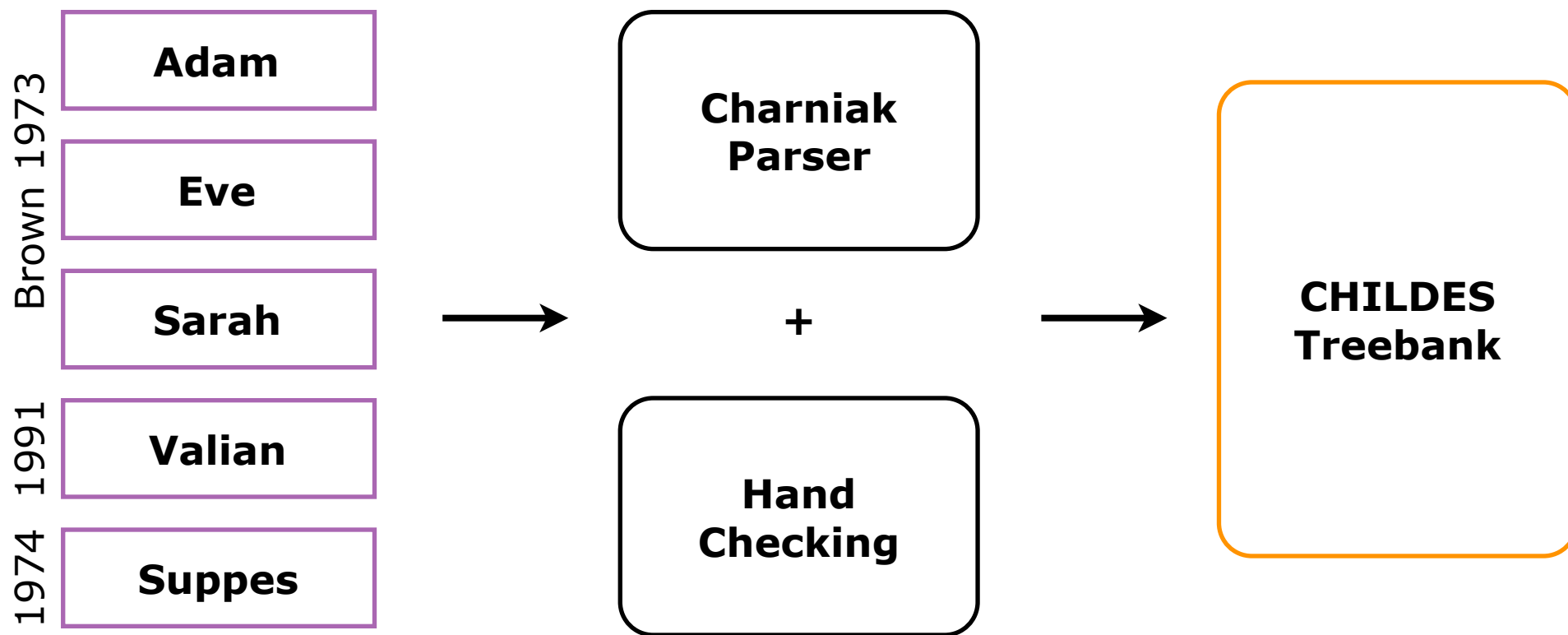
Crucially, both types of learners must use realistic input in order for any claims to be made. Without realistic input, they have no inferential value.

Realistic input: Syntactically Annotated CHILDES

www.childes.psy.cmu.edu

www.cs.brown.edu/~ec

www.socsci.uci.edu/~lpearl/CoLaLab/TestingUG

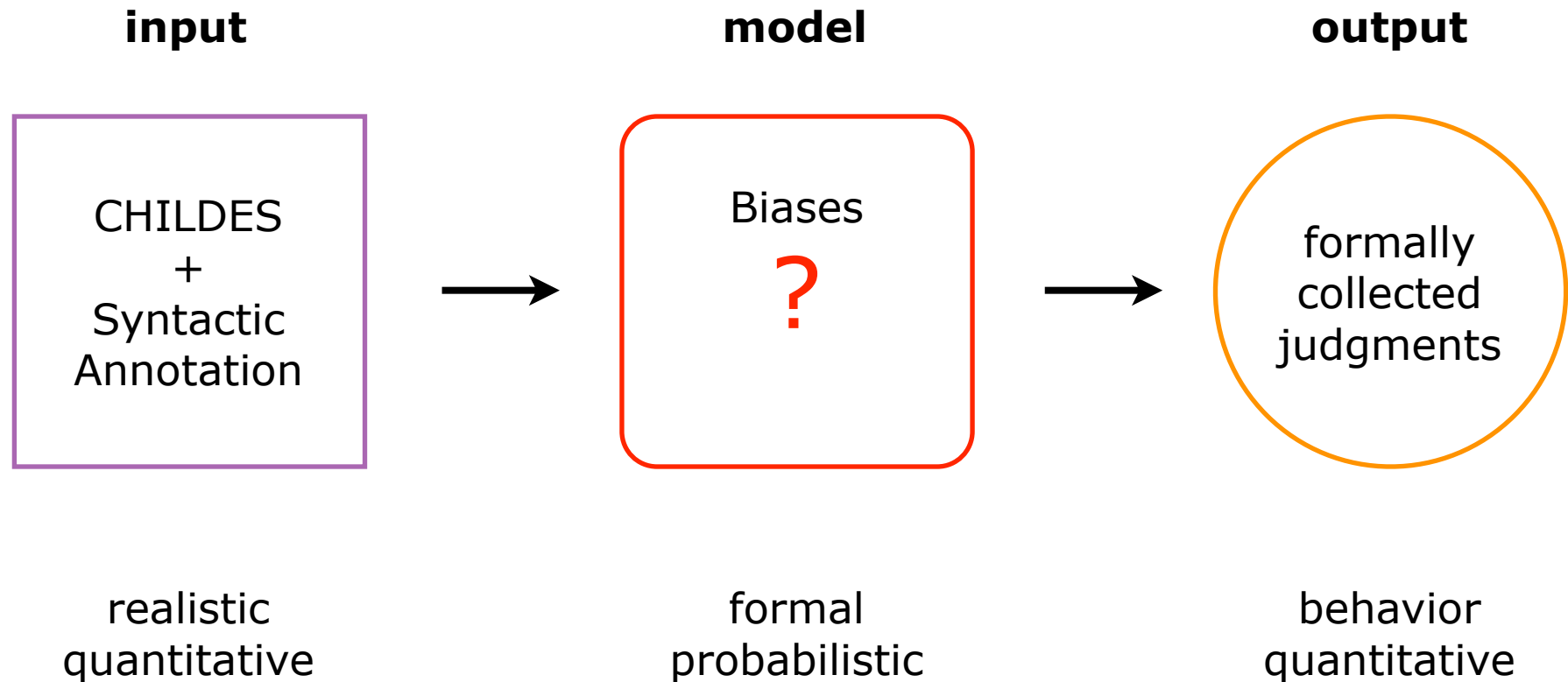


25 children
1:5 years
813,036 words

3 years and counting

Freely available!

Goals of the project



The explicit and implicit biases of the model will give us insight into the cognitive biases that are required by humans

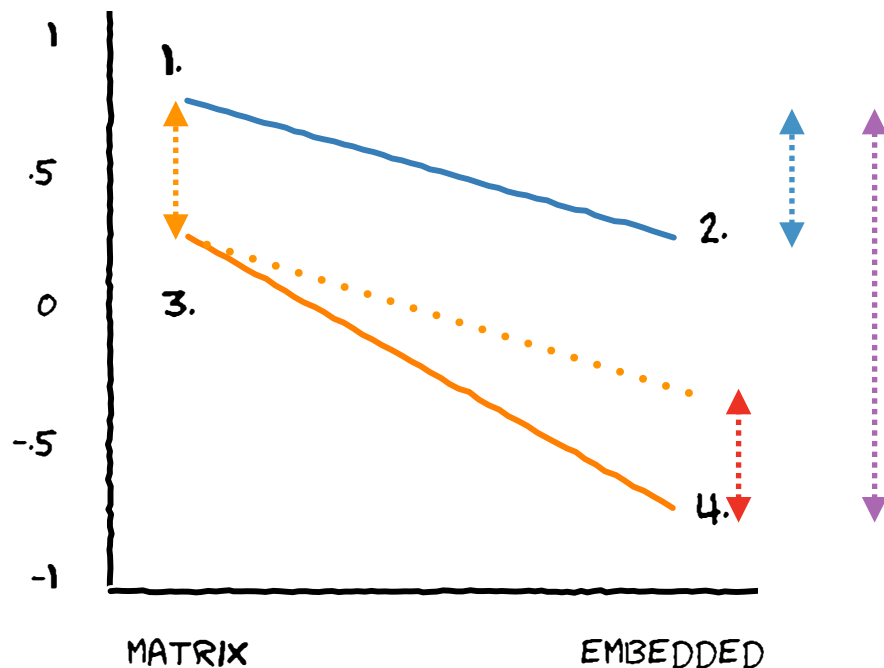
Remember this?

complex
phrase
effect

global effect

1. Who ___ thinks that Jack stole the necklace?
2. What do you think that Jack stole ___?
3. Who ___ wonders whether Jack stole the necklace?
4. *What do you wonder whether Jack stole ___?

dependency
effect

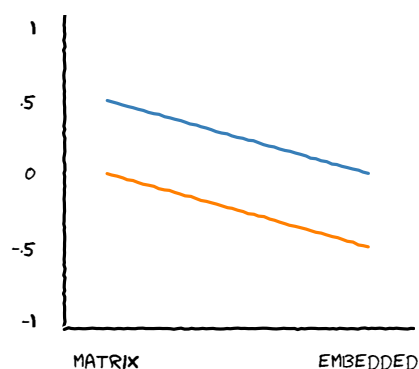


	dependency effect	(1-2)
	complexity effect	(1-3)
+	constraint	+
	global effect	+
		X
		(1-4)

If there is a **grammatical constraint**, the two independent effects won't be enough to explain the total global effect. We'll need to add the constraint's effect in.

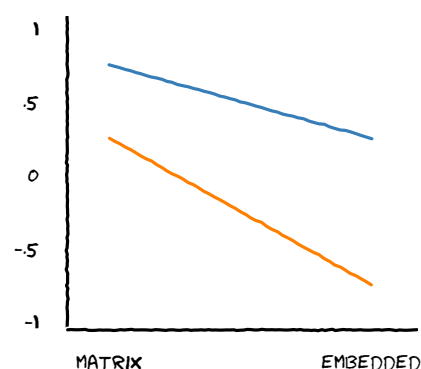
Evaluating the success of the model

linearly additive



Unsuccessful
Learning

super-additive



Successful Learning

The goal here is not to learn a specific grammatical constraint, but rather to learn a system that would give rise to the pattern of behavior that adults show.

This is a bit strange from a linguistic perspective, because we generally care about the specific constraint that is learned. But as we will see, there are interesting conclusions to be drawn from just learning the behavior!

Building a model

Goal: To build the **simplest** model that successfully learns the superadditive pattern of acceptability judgments

Simply tracking the sentence types won't work:

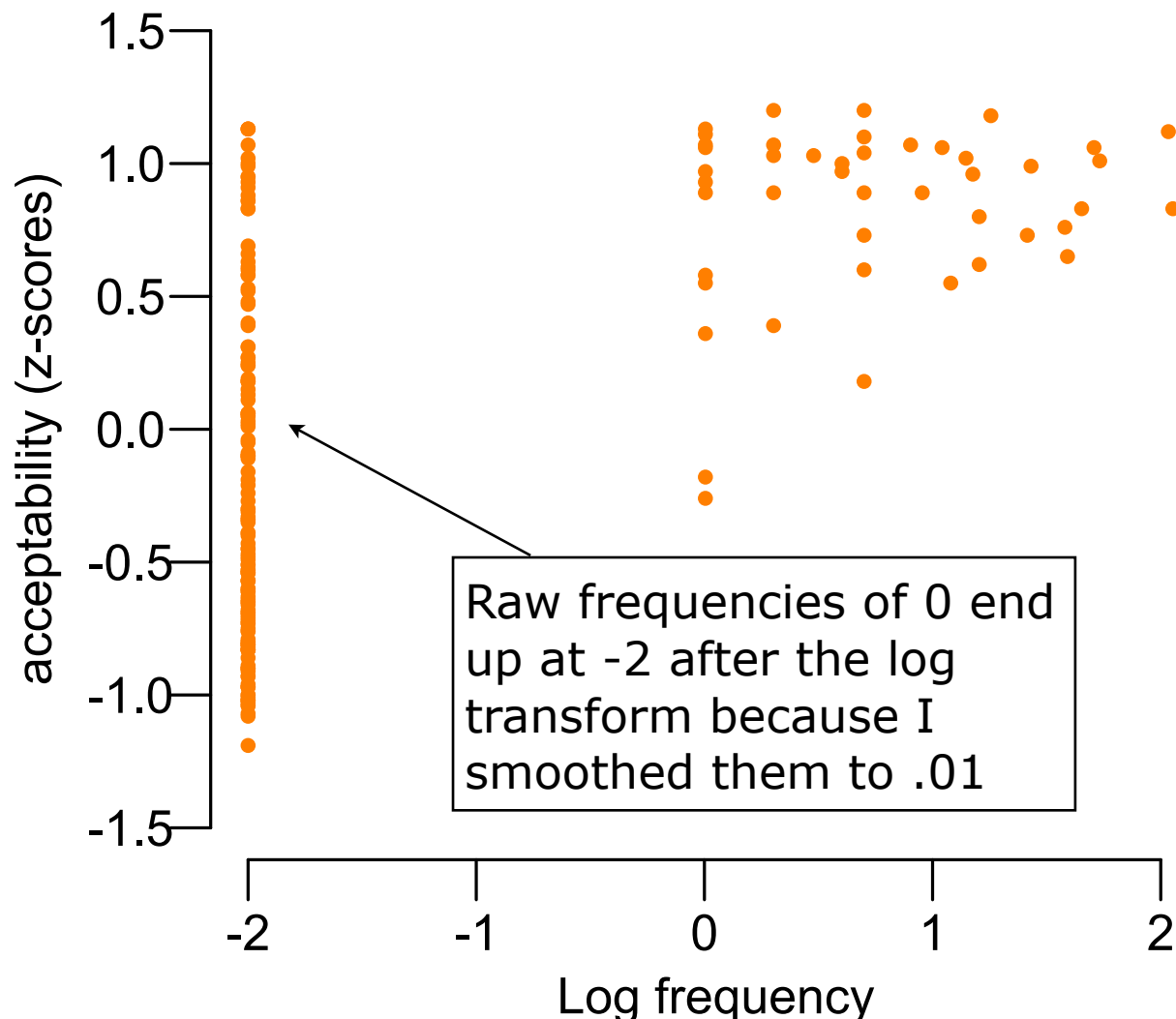
	MATRIX NON-ISLAND	EMBEDDED NON-ISLAND	MATRIX ISLAND	<i>EMBEDDED</i> <i>ISLAND</i>
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

If we just used the simple mapping that "0" = ungrammatical, we would predict that several of the grammatical sentences should be ungrammatical.

This is part of a broader problem with using frequency as a predictor

Linguists have long known that sentence frequency and grammaticality are not identical. This shows that the same is true of frequency and acceptability.

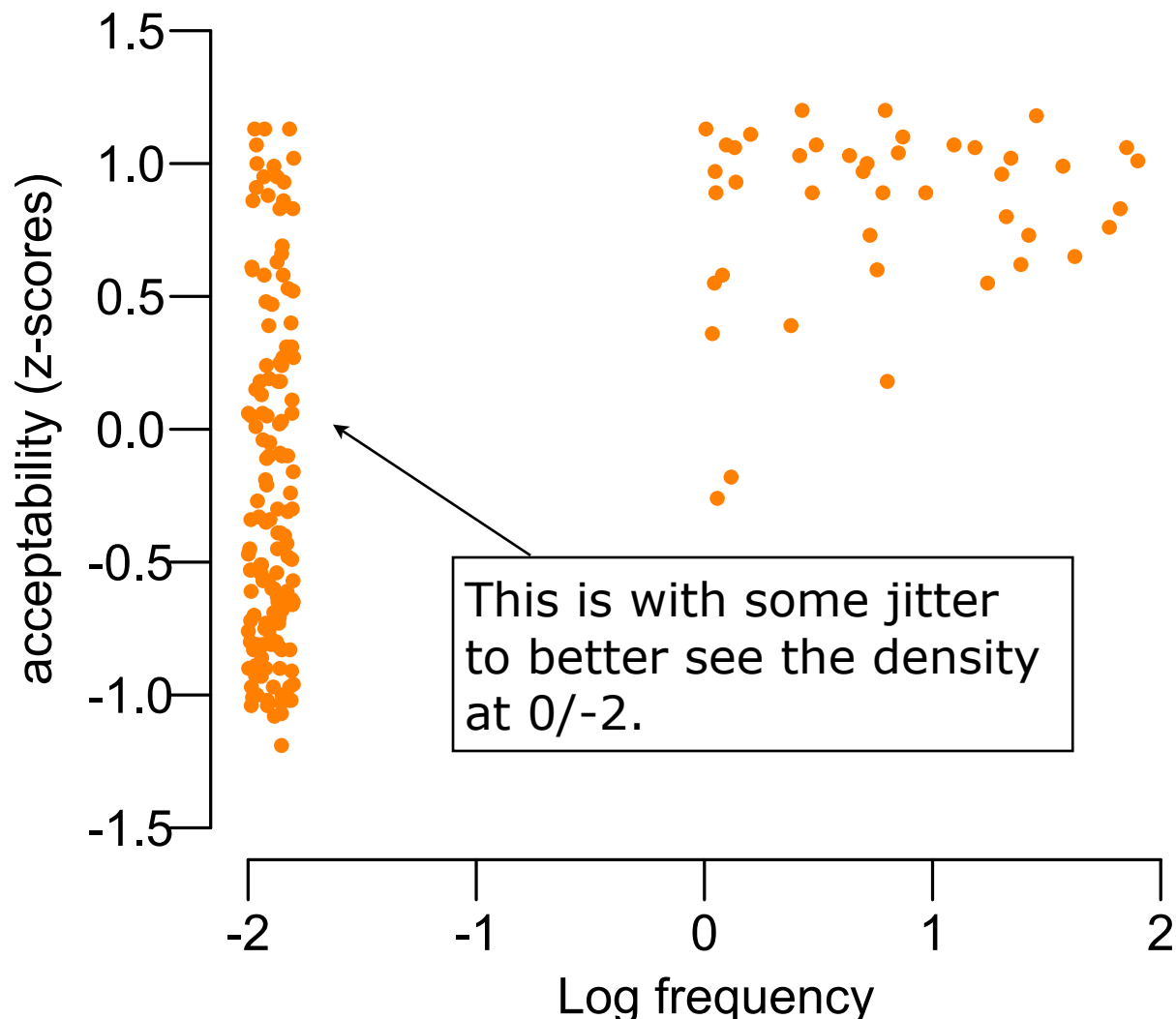
These are the judgments from Adger 2003 (textbook) as tested in Sprouse & Almeida 2012 versus frequency counts from Switchboard (spoken), Brown (written), and English Web (both).



This is part of a broader problem with using frequency as a predictor

Linguists have long known that sentence frequency and grammaticality are not identical. This shows that the same is true of frequency and acceptability.

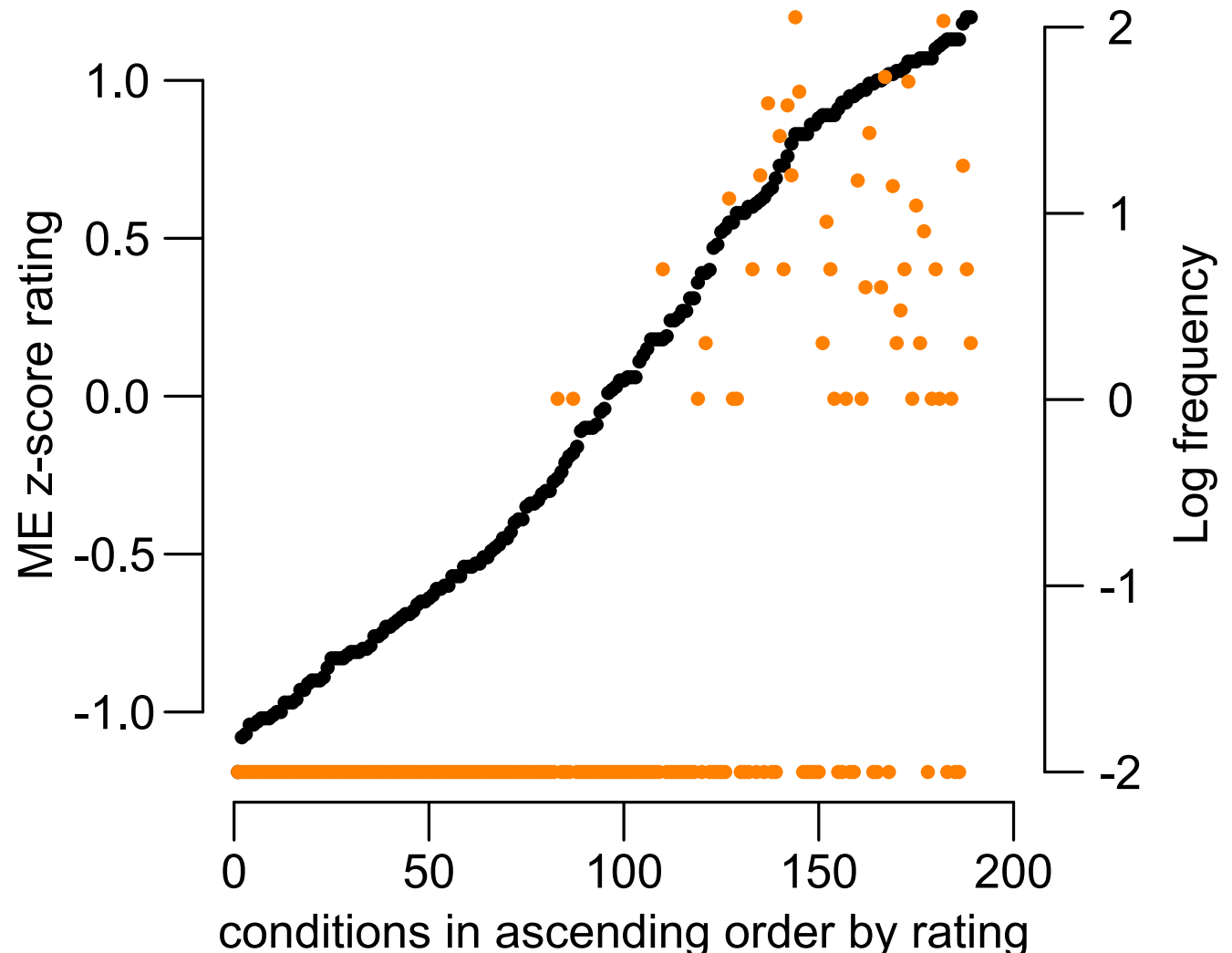
These are the judgments from Adger 2003 (textbook) as tested in Sprouse & Almeida 2012 versus frequency counts from Switchboard (spoken), Brown (written), and English Web (both).



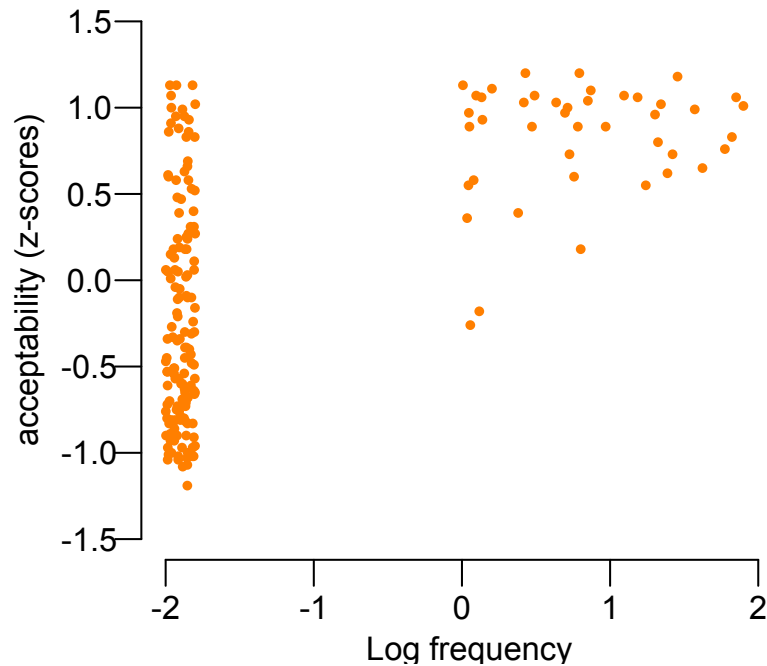
Another way to look at this graph

Another way to see that acceptability cannot easily be reduced to frequency is to plot the acceptability judgments in ascending order, and then plot the frequencies in ascending order.

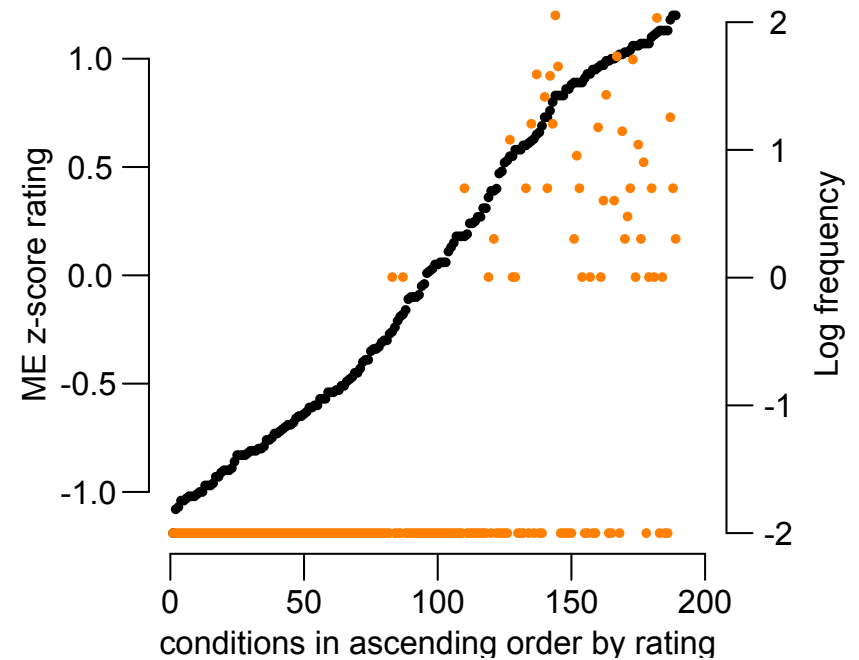
This shows the same information as before, but in a slightly different way.



Putting the two graphs next to each other



This one really highlights that there are acceptable sentences that are infrequent.



This one really highlights that frequency is a bad predictor of acceptability.

We can try even fancier measures of frequency...

For example, instead of looking at the raw frequency of the entire sentence, we could break the sentence down into **bigrams of categories**:

John bought a car

John bought		DP V	=	$p(\text{DP V}) = .025$
bought a	=	V D	=	$p(\text{V D}) = .04$
a car		D NP	=	$p(\text{D NP}) = .01$

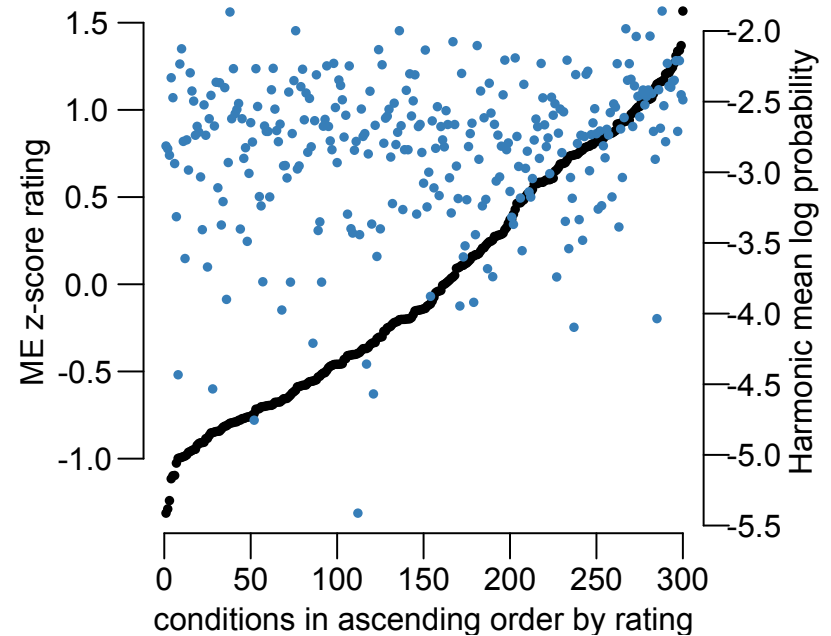
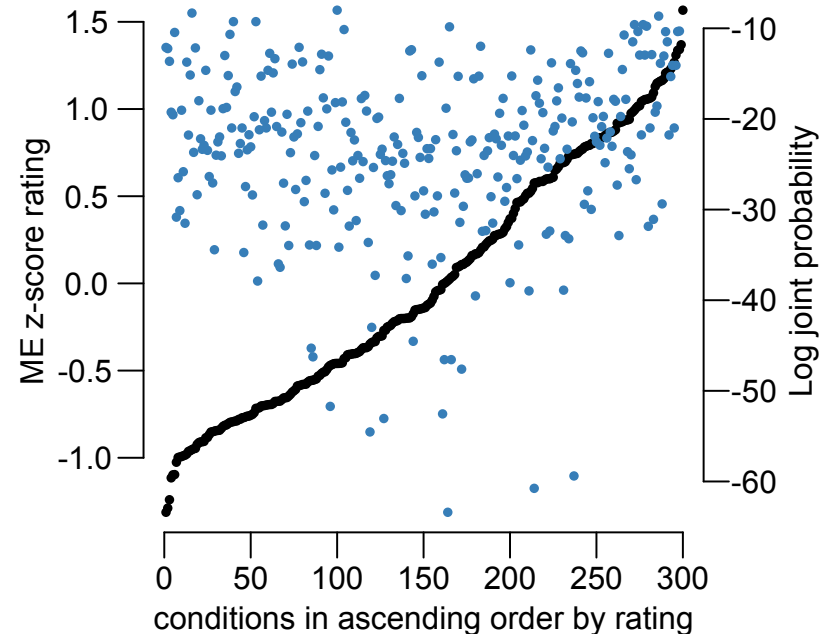
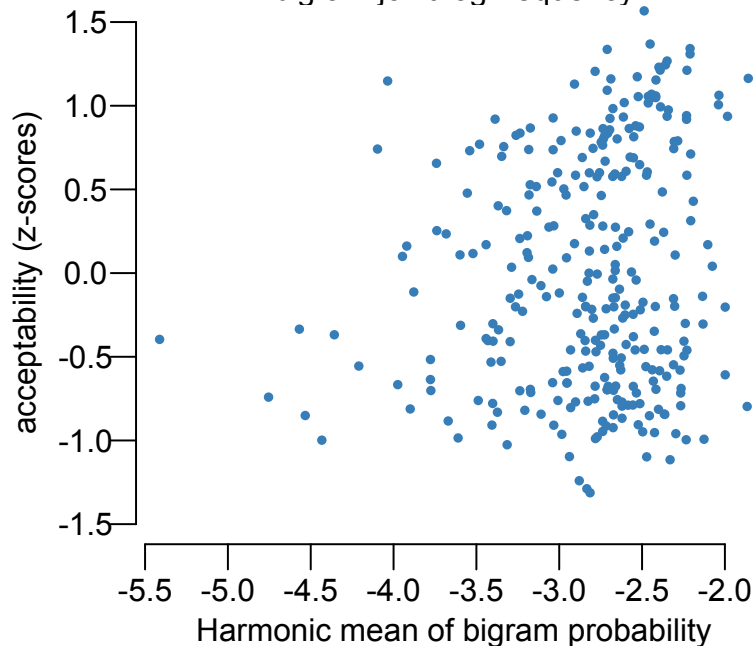
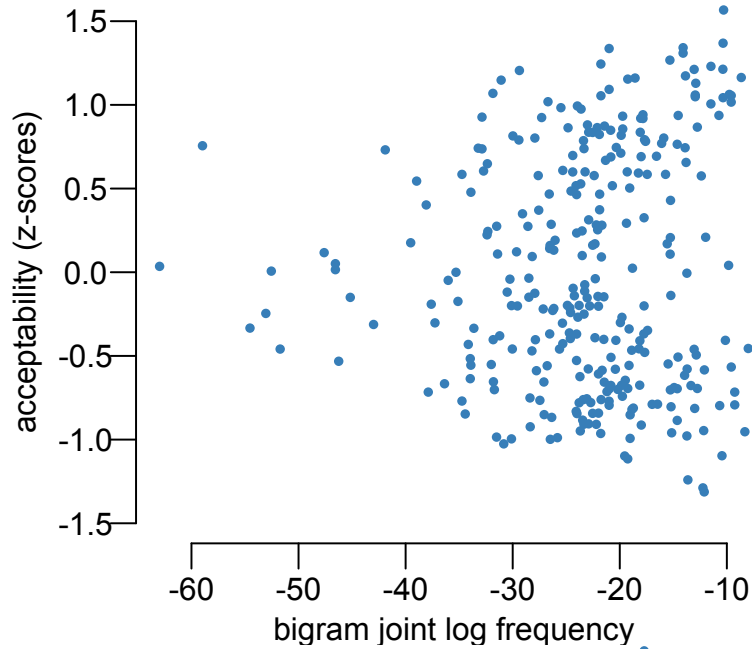
Then we can look at statistics derived from these categories, such as the joint probability (the probability of having all of these bigrams together):

$$\text{joint probability} = p(\text{DP V}) * p(\text{V D}) * p(\text{D NP}) = .025 * .04 * .01 = .00001$$

Or the harmonic mean (the reciprocal of the mean of the reciprocals):

$$\text{h.mean} = \frac{3}{\frac{1}{.025} + \frac{1}{.04} + \frac{1}{.01}} = .0181818\dots$$

... and they still don't predict acceptability



So we need to build a more sophisticated model to capture islands

- 1. CONTAINER NODE PATH:** **What** do [you [wonder [**whether** [Jack [**stole** ___]]]]]

(SUBCATEGORIZE CP) start IP VP CP_{whether} IP VP end

- 2. TRACK FREQUENCY OF TRIGRAMS OF NODES:**

start IP VP

 IP VP CP_{whether}

 VP CP_{whether} IP

 CP_{whether} IP VP

 IP VP end

- 3. CALCULATE THE PROBABILITY OF A SEQUENCE OF TRIGRAMS:**

$p(\text{dependency}) = p(\text{trigram1}) * p(\text{trigram2}) * p(\text{trigram3})$

- 4. USE P(DEPENDENCY) AS A MAJOR COMPONENT OF ACCEPTABILITY:**

$p(\text{dependency}) \sim \text{Acceptability}$

How did we evaluate the results?

THE LEARNING PERIOD: About 1M utterances between 0:3 (Hart & Risley 1995)

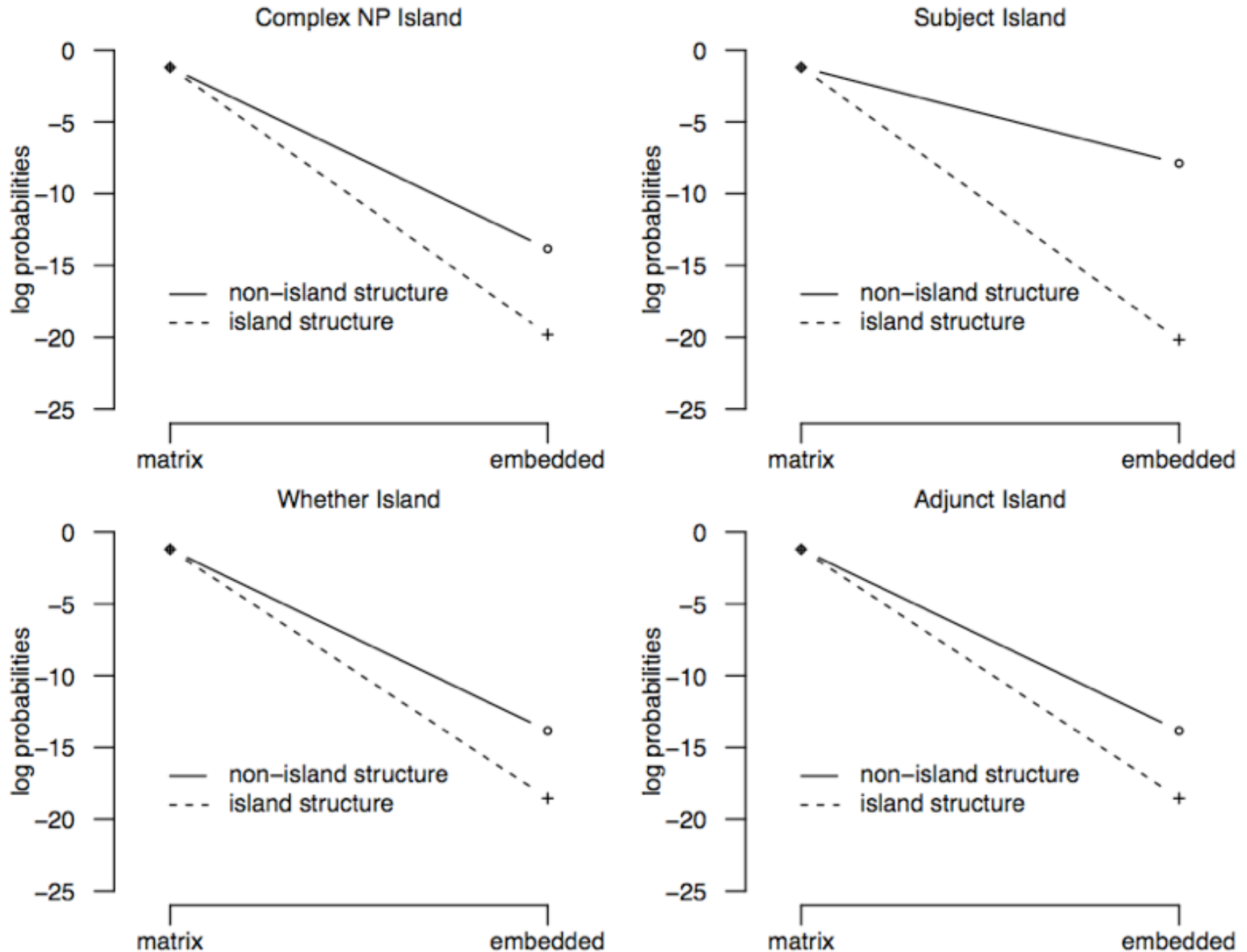
About 20% wh-utterances in the CHILDES Treebank

So assuming children learn islands within a three year period (e.g., 2:5), they would encounter approximately 200,000 wh-utterances

THE RESULTS: We can calculate the log-probability of the four sentence types for each island type after the learning period is over.

We can then plot the log-probability of the four sentence types for each island type to see if it shows the characteristic superadditive shape.

Perhaps surprisingly, the model works!



So what is going on?

The goal of this project is to create a model that learns a certain behavior, without committing to any specific grammatical model. Then we can ask which of the biases that are necessary are potentially innate and domain-specific.

	domain-specific	domain-general	Potential UG biases:
innate	Universal Grammar	tracking frequencies calculating probability chunking n-grams	tracking wh-input paying attention to container nodes paying attention to subcategorized CPs tracking 3-grams
derived	phrase structure identifying container nodes subcategorizing CPs		

But others are potentially in the red square

Putting the potential UG biases into words:

1. The first question is why the system even tries to **track information about wh-dependencies**. Some have suggested it may have to do with parsing efficiency: Fodor 1978, Berwick and Weinberg 1984, Hawkins 1999, etc.
2. The next question is why the system **pays attention to container nodes**. Information about container nodes is clearly available (as part of the parsing of the dependency), but so is other information: number of nouns, number of arguments, etc. Why are container nodes special?
3. Another question is why the system **pays attention to subcategorized CPs**. The fact that CPs are subcategorized is straightforward: differences in CPs lead to differences in sentential semantics. But so do other finer-grained XPs. Why does this system treat these as distinct but not others? In other words, why is it this specific level of analysis and not one of the other possible levels?
4. And the final question is why are the container nodes tracked as trigrams? Why not bigrams or tetragrams? (The answer is that trigrams work and the others don't, but how does the learner know that?)

Problems with the model: parasitic gaps

Under normal circumstances, a gap inside of a **clausal adjunct** leads to unacceptability, suggesting that one or more of the trigrams are low frequency.

* **Which book** did [you [laugh] [**before** [[**reading** ___]]]]

start IP VP CP_{adj} IP VP end

But placing a second gap in the main clause leads to an acceptable sentence. These are called parasitic gap constructions because the adjunct gap can only exist as a parasite on the extra gap:

Which book did [you [judge ___ [**before** [[**reading** ___]]]]

start IP VP CP_{adj} IP VP end

These are the same sequence, so this model can't predict the difference

Problems with the model: Cross-linguistic variation

		2x2 Definition			
Language	Type	WH	NP	SUB	ADJ
English	wh-move	Red	Red	Red	Red
	rc-move	Red	Red	Red	Green
	in-situ	Green	Green	Green	Green
	d-linking	Red	Red	Red	Red
Italian	wh-move	Red	Red	Red	Red
	rc-move	Red	Red	Green	Red
Swedish	wh-move	Red	Red	Red	Red
Norwegian	wh-move	Red	Red	Red	Red
Arabic	wh-move	Green	Green	Grey	Orange
Japanese	in-situ	Green	Green	Green	Green

Whether we use informal definitions of islands, or the 2x2 design, there is cross-linguistic variation in island effects.

The current model does not impose any constraints on the pattern of island effects that can be learned. It simply learns whatever is in the input.

So if there are constraints on variation, like the old IP/CP parameter for bounding nodes, then this model will fail to capture that fact.

The best this model can ever do is call the pattern **accidental**.

Evaluating the components of the model

We can look at each of the components (explicit and implicit) of the model, and ask (i) where they appear in the space of learning biases, and (ii) what they tell us about the necessary components of a solution to island effects.

	domain-specific	domain-general	Potential UG biases:
innate	Universal Grammar	tracking frequencies calculating probability chunking n-grams	tracking wh-input paying attention to container nodes paying attention to subcategorized CPs tracking 3-grams
derived	phrase structure identifying container nodes subcategorizing CPs		parasitic gaps cross-linguistic variation

One final problem: Integration with other theories of acquisition

This model is highly specific - it only works for **constraints on wh-questions**.

A complete model of syntactic acquisition will require additional models, raising the difficult question of how all of these different models integrate with each other (or not).

Case

Phrase
Structure

Agreement

Wh-questions

Complementation

Coreference

THANK YOU!
and thank you to Lisa Pearl!

Lisa Pearl
UC Irvine

