Title: When socioeconomic status differences don't affect input quality: Learning complex syntactic knowledge

Running Title: Learning complex syntactic knowledge from low-SES input

Authors: Alandi Bates and Lisa Pearl

Conflict of Interest Statement: We certify that there are no conflicts to report.

Acknowledgments: We are very grateful to the Quantitative Language Collective at UC Irvine for comments and discussion, as well as the audience at BUCLD 2018.

*Alandi Bates, University of California, Irvine, ajbates@uci.edu. Lisa Pearl, University of California, Irvine, lpearl@uci.edu.

When socioeconomic status differences don't affect input quality:
Learning complex syntactic knowledge

Research Highlights

- We investigate whether the quantity and quality of child-directed speech (**CDS**) necessary for learning complex syntactic knowledge differs across socioeconomic status (**SES**), the way it does for other linguistic knowledge. We focus on constraints on *wh*-dependencies, called *syntactic islands*.
- Descriptive corpora analyses and quantitative analyses suggest that the *wh*-dependency distributions across SES, from which children would learn syntactic island knowledge, are quantitatively similar.
- Cognitive modeling analyses suggest that the *wh*-dependency distributions are also qualitatively similar, leading to the same predicted learning outcomes for syntactic island knowledge.
- Interestingly, a crucial structural building block for learning syntactic island knowledge is present in both the low- and high-SES *wh*-dependency input distributions, but comes in the low-SES input from a dependency that is judged ungrammatical by high-SES adults.

Abstract

While there are known differences in the quantity and quality of child-directed speech (**CDS**) across socioeconomic status (**SES**) for some linguistic knowledge, it is unknown if these differences also extend to complex syntactic knowledge. We investigate *wh*-dependency constraints, known as *syntactic islands*, as a concrete case of complex syntactic knowledge where the quantity and quality of high-SES CDS has been previously assessed. Using quantitative analysis and cognitive modeling to assess low-SES CDS samples, we find that the *wh*-dependencies in low-SES children's complex syntactic input are quantitatively and qualitatively similar to those of high-SES children. In particular, the *wh*-dependencies (i) have similar distributions across SES, and (ii) would allow successful acquisition of syntactic island knowledge. Interestingly, at least one key building block for syntactic island knowledge comes from a different source in low-SES children's input, but is crucially still present. This highlights an important qualitative similarity across SES. Our results suggest that the linguistic evidence for more complex syntactic knowledge like syntactic islands, in contrast with more foundational linguistic knowledge, may not differ by SES. We discuss implications for linguistic development and adult syntactic knowledge variability across SES.

Key Words: socioeconomic status, linguistic development, child-directed speech, syntactic islands constraints, computational modeling, quantitative approaches, input quantity, input quality

1 Introduction: The impact of socioeconomic status

It's hotly contested if and how socioeconomic status (**SES**) affects children's linguistic input (Hart and Risley, 1995; Schwab and Lew-Williams, 2016; Sperry et al., 2018). For instance, with respect to input quantity, some studies have suggested differences as measured by words, with 4-year-old children from lower-SES backgrounds encountering 30 million fewer words than their higher-SES counterparts (Hart and Risley, 1995; Schwab and Lew-Williams, 2016). However, more recent studies considered other sources of caretaker speech besides maternal speech (Blum, 2015; Sperry et al., 2018), and found greater linguistic variation *within* SES than across SES for word quantity. For input quality, the most recent studies suggest that SES differences do impact child-directed speech (**CDS**) at the lexical and foundational syntactic levels (Huttenlocher et al., 2010; Rowe, 2012; Rowe et al., 2017); differences have been found in the relative frequency of word types, word tokens, rare words, the diversity of syntactic constructions, and the relative frequency of decontexualized utterances like explanations (*Oh, we can't put them in the bus because the bus is full of blocks*), pretend (*I'll save you from the wicked sister*), and narrations (*He is going to look in your nose and your throat and your ears*).

Importantly, these input quality differences are known to affect linguistic knowledge development. For instance, children from high-SES families generally have larger vocabularies and build these vocabularies earlier and faster than their low-SES counterparts (Hart and Risley, 1995; Hoff, 2003). There's also suggestive evidence that effective intervention is possible when needed (Huttenlocher et al., 2002; Rowe et al., 2017), which could potentially mediate the impact of SES-based input differences. For example, when teachers provided high-quality input to their students, low-SES students showed improvement in measures of comprehension on par with their high-SES peers by the end of the school year (Huttenlocher et al., 2002). Of course, input-based interventions like this are only effective if we know what high-quality input actually consists of. That is, if we identify significant differences in the input across SES, we then know more precisely what's missing and how to fix it.

Notably, far less is known about the development of complex syntactic knowledge across SES, the input differences that may exist at this level, and how those input differences impact knowledge development. For instance, Rowe et al. (2017) found that there was variation within low-SES input for *wh*-questions (e.g., *What kind of animal is that?*) asked by fathers; interestingly, the frequency of *wh*-questions was significantly and positively correlated with a measure of vocabulary at 24 months and a measure of verbal reasoning skills at 36 months. This may be because *wh*-questions elicit more syntactically complex responses than other types of questions (e.g., yes/no questions: *Is that a dog?*). In terms of *wh*-question knowledge itself, there's suggestive evidence that by age 4, children from diverse SES and linguistic backgrounds (i.e., both low- to high-SES children across dialects of American English and language-impaired populations) are capable of interpreting complex *wh*-questions similarly to adult speakers (de Villiers et al., 2008). Still, much remains unknown about input variation as it impacts the development of complex syntactic knowledge itself, such as how to form grammatical *wh*-questions.

Here we investigate input variation for complex syntactic knowledge that's intimately connected to *wh*-questions. More specifically, we look at the distributions of *wh*-dependencies (i.e., dependencies that involve so-called *wh*-words like *what* in *What did you see?*, where *what* is interpreted as the thing seen). These *wh*-dependencies reflect knowledge of *syntactic islands*, a form of complex syntactic knowledge that determines which *wh*-dependencies are allowed and which aren't. Knowledge of syntactic islands in turn allows children to know which *wh*-questions are well-formed and which aren't. We investigate the differences in both the quantity and quality of *wh*-dependency input in American English CDS between high-SES populations and low-SES populations. By assessing the *wh*-dependency distributions in the CDS of high- and low-SES children, we can determine whether the low-SES *wh*-dependency distribution supports the acquisition of syntactic island knowledge as well as the high-SES distribution has been shown to do (Pearl and Sprouse, 2013).

We first review the complex syntactic knowledge that syntactic islands require and its relationship to *wh*-dependencies, highlighting a strategy that children can use to learn about syntactic islands from the *wh*-dependencies in their input. Then, we review what low-SES CDS looks like for syntactic islands, providing both a descriptive corpus analysis and a quantitative analysis comparing high-SES to low-SES input quality. We then provide a cognitive modeling analysis of the input quality, where we predict the syntactic island knowledge that low-SES children would be able to attain on the basis of their *wh*-dependency input.

We find that the low-SES input, in terms of *wh*-dependency distribution and the syntactic building blocks needed for syntactic islands, is both quantitatively and qualitatively similar to the high-SES CDS distribution. This suggests low-SES input may support acquisition of syntactic islands as well as high-SES input does. Our cognitive modeling analysis accords with this prediction: children learning from the low-SES CDS can acquire the same syntactic islands knowledge that children learning from high-SES CDS would. Interestingly, a crucial syntactic building block involving complementizer *that* comes from a different *wh*-dependency type in low-SES CDS; this highlights that surface input quality differences may mask deeper input quality similarities. Taken together, our results suggest that the nature of the input for learning about syntactic islands doesn't fundamentally differ across SES; this notably contrasts with input differences found for more foundational lexical and syntactic knowledge. We discuss implications for linguistic development across SES and potential adult syntactic knowledge variation.

2 Syntactic islands and SES

2.1 *Wh*-dependencies & syntactic islands

A key component of human syntactic knowledge is the ability to have long-distance dependencies, where there's a relationship between two words that aren't adjacent to each other. Long-distance dependencies, such as the dependencies between the *wh*-word *what* and *eat* in (1), can be arbitrarily long (Chomsky, 1965; Ross, 1967; Chomsky, 1973). In (1), we can see that this dependency can stretch across one, two, three, or four clauses. In each case, *what* is understood as the thing Falkor ate, despite *what* not being adjacent to *eat*. This relationship is marked with $\_\_{what}$.

(1) a. What did Falkor eat $\_\_{what}$?
b. What did Atreyu see Falkor eat $\_\_{what}$?
c. What did the Childlike Empress say Atreyu saw Falkor eat $\_\_{what}$?
d. What did Bastian hear the Childlike Empress say Atreyu saw Falkor eat $\_\_{what}$?

However, there are specific syntactic structures that long-distance dependencies can't cross: syntactic islands (Chomsky, 1965; Ross, 1967; Chomsky, 1973). Four examples of syntactic islands are in (2), with * indicating ungrammaticality and [...] highlighting the proposed island structure that a *wh*-dependency can't cross in English.

(2) a. **Complex NP island**
*What did Falkor make [the claim [that Atreyu fought $\_\_{what}$]]?
b. **Subject island**
*What did Falkor think [[the joke about $\_\_{what}$] was hilarious]?
c. **Whether island**
*What did Falkor wonder [whether Atreyu bought $\_\_{what}$]?
d. **Adjunct island**
*What did Falkor worry [if Atreyu buys $\_\_{what}$]?

During language development, children must infer and internalize the constraints on long-distance dependencies (i.e., syntactic island constraints) that allow them to recognize that the questions in (2) are not allowed, while the questions in (1) are fine.

2.2 Learning syntactic islands

Pearl and Sprouse (2013) constructed a cognitive computational model for learning these syntactic island constraints. This model relies on the idea that children can characterize a long-distance dependency as a syntactic path from the head of the dependency (e.g., *What* in (3)) through a set of structures that contain the tail (e.g., $\_\_{what}$) of the dependency, as shown in (3a)-(3b). These structures correspond to phrase types such as Verb Phrases (VP), Inflectional Phrases (IP), and Complementizer Phrases (CP), among others. Under this view, children simply need to learn which long-distance dependencies have licit syntactic paths and which don't.

To model this learning process, Pearl and Sprouse (2013) implemented a probabilistic learning algorithm that tracks local pieces of these syntactic paths. It breaks the syntactic path into a collection of syntactic trigrams that can be combined to reproduce the original syntactic path, as shown in (3c). The learning model then tracks the frequencies of these syntactic trigrams in the input. It later uses them to calculate probabilities for all syntactic trigrams comprising a *wh*-dependency[1] and so generate the probability of any *wh*-dependency (as shown in (4)- (5)). More specifically, any *wh*-dependency's probability is the product of the individual trigram probabilities that comprise its syntactic path, as shown in (6). The generated probability corresponds to whether that dependency is allowed, with higher probabilities indicating grammatical dependencies and lower probabilities indicating ungrammatical dependencies.

(3) What did Falkor claim that Atreyu fought $\_\__{what}$?

a. Syntactic structures containing the *wh*-dependency:
What did [$_{IP}$ Falkor [$_{VP}$ claim [$_{CP}$ that [$_{IP}$ Atreyu [$_{VP}$ fought $\_\__{what}$]]]]]?

b. Syntactic path of *wh*-dependency:
start-IP-VP-CP$_{that}$-IP-VP-*end*

c. Syntactic trigrams $T \in$ syntactic path:
= *start*-IP-VP
IP-VP-CP$_{that}$
VP-CP$_{that}$-IP
CP$_{that}$-IP-VP
IP-VP-*end*

(4) Smoothed probabilities of trigrams:
p(*start*-IP-VP) $\approx \frac{count(start-IP-VP)}{total\ count\ of\ all\ trigrams}$
...
p(IP-VP-*end*) $\approx \frac{count(IP-VP-end)}{total\ count\ of\ all\ trigrams}$

(5) Probability of new *wh*-dependency: What did Engywook tell Atreyu $\_\__{what}$?
Syntactic structures = What did [$_{IP}$ Engywook [$_{VP}$ tell Atreyu $\_\__{what}$?]]
Syntactic path = *start*-IP-VP-*end*
trigrams = *start*-IP-VP, IP-VP-*end*
Probability = p(*start*-IP-VP-*end*) = p(*start*-IP-VP)*p(IP-VP-*end*)

(6) General formula for generating a *wh*-dependency's probability:
$\prod_{trigrams \in T} p(trigram)$

2.3 High-SES input quality for syntactic islands

To evaluate high-SES syntactic input quality, Pearl and Sprouse (2013) modeled a learner using this strategy and let the modeled learner use as input a realistic sample of high-SES American English CDS. These high-SES input data came from the structurally-annotated Brown-Adam (Brown, 1973), Brown-Eve (Brown, 1973), Valian (Valian, 1991), and Suppes (Suppes, 1974) corpora from the CHILDES Treebank (Pearl and Sprouse, 2013), comprising 102K utterances with 21K *wh*-dependencies. The modeled learner encountered a quantity of CDS equivalent to the quantity of data high-SES children typically encounter during the time when they're learning about syntactic island constraints (estimated to take three years), which was equivalent to ≈200K *wh*-dependencies. With this input, the model estimated syntactic trigram probabilities and could then generate probabilities for any desired *wh*-dependency.

The *wh*-dependencies that the model needed to generate probabilities were those that American English adults had given acceptability judgments for in Sprouse et al. (2012), corresponding to the four islands from (2); a sample set for each island type is shown in (7)-(10), where island

[1]It smooths these probabilities by adding 0.5 to all trigram counts. This allows the model to accept dependencies composed of trigrams it's never seen before, though it gives them a much lower probability than dependencies composed of trigrams it has in fact seen before. See Pearl and Sprouse (2013, 2015) for further discussion of this point.

structures are indicated with [...]. These stimuli were designed using a 2x2 factorial design, involving dependency length (matrix vs. embedded) and presence of an island structure in the utterance (non-island vs. island). Each island stimuli set therefore had four dependency types: matrix+non-island, embedded+non-island, matrix+island, and embedded+island; the embedded+island stimulus in each case involved a *wh*-dependency that crossed a syntactic island, and so was ungrammatical. These experimental stimuli can be characterized by the syntactic paths shown in Table 1. Note that many of the grammatical dependencies for each island type (e.g., matrix+non-island and matrix+island) are characterized by the same syntactic path (e.g., *start*-IP-*end*).

(7) Sample Complex NP island stimuli

a. matrix+non-island
Who $\_\__{who}$ claimed that Atreyu fought the goblin?

b. embedded+non-island
Who did Falkor claim that Atreyu fought $\_\__{who}$?

c. matrix+island:
Who $\_\__{who}$ made [the claim that Atreyu fought the goblin]?

d. embedded+island:
*Who did Falkor make [the claim that Atreyu fought $\_\__{who}$]?

(8) Sample Subject island stimuli

a. matrix+non-island:
Who $\_\__{who}$ thinks the joke is hilarious?

b. embedded+non-island:
What does Falkor think $\_\__{what}$ is hilarious?

c. matrix+island:
Who $\_\__{who}$ thinks the joke about Atreyu is hilarious?

d. embedded+island:
*Who did Falkor think [[the joke about $\_\__{who}$] was hilarious]?

(9) Sample Whether island stimuli

a. matrix+non-island:
Who $\_\__{who}$ thinks Atreyu bought the medallion?

b. embedded+non-island:
What does Falkor think Atreyu bought $\_\__{what}$?

c. matrix+island:
Who $\_\__{who}$ wonders if Atreyu bought the medallion?

d. embedded+island:
*What did Falkor wonder [whether Atreyu bought $\_\__{what}$]?

(10) Sample Adjunct island stimuli

a. matrix+non-island:
Who $\_\__{who}$ thinks Atreyu bought the medallion?

b. embedded+non-island:
What does Falkor think that Atreyu bought $\_\__{what}$?

c. matrix+island:
Who $\_\__{who}$ worries if Atreyu bought the medallion?

d. embedded+island:
*What did Falkor worry [if Atreyu buys $\_\__{what}$]?

This design allows syntactic island knowledge to surface as a superadditive interaction of acceptability judgments; this superadditivity appears as non-parallel lines in an interaction plot, such as those in Figure 1. In particular, if we consider the Complex NP plot in the top row, there are four acceptability judgments, one for each of the stimuli in (7). The matrix+non-island dependency of (7a) has a certain acceptability score – this is the top-lefthand point. There is a (slight) drop in acceptability when the matrix+island dependency of (7c) is judged in comparison to (7a) – this is the lower-lefthand point. We can interpret this as the unacceptability associated with simply having an island structure in the utterance. There's also a drop in acceptability when the

Table 1: Syntactic paths for experimental stimuli that acceptability judgments are generated for, in a 2x2 factorial design varying dependency length (*mat*rix vs. *emb*edded) and presence of an island structure (*non*-island vs. *island*). Ungrammatical island-spanning dependencies are indicated with *.

		Complex NP islands	***Subject islands***
mat	non	*start*-IP-*end*	*start*-IP-*end*
emb	non	*start*-IP-VP-CP$_{that}$-IP-VP-*end*	*start*-IP-VP-CP$_{null}$-IP-*end*
mat	island	*start*-IP-*end*	*start*-IP-*end*
emb	island	**start*-IP-VP-NP-CP$_{that}$-IP-VP-*end*	**start*-IP-VP-CP$_{null}$-IP-NP-PP-*end*
		Whether islands	***Adjunct islands***
mat	non	*start*-IP-*end*	*start*-IP-*end*
emb	non	*start*-IP-VP-CP$_{that}$-IP-VP-*end*	*start*-IP-VP-CP$_{that}$-IP-VP-*end*
mat	island	*start*-IP-*end*	*start*-IP-*end*
emb	island	**start*-IP-VP-CP$_{whether}$-IP-VP-*end*	**start*-IP-VP-CP$_{if}$-IP-VP-*end*

embedded+non-island dependency of (7b) is judged in comparison to (7a) – this is the upper-righthand point. We can interpret this as the unacceptability associated with simply having an embedded *wh*-dependency. If the unacceptability of the embedded+island dependency of (7d) were simply the result of those two unacceptabilities (having an island structure in the utterance and having an embedded *wh*-dependency), the drop in unacceptability would be additive and the lower-righthand point would be just below the upper-righthand point (and so look just like the points on the lefthand side). But this isn't what we see – instead, the acceptability of (7d) is much lower than this. This is a superadditive effect for the embedded+island stimuli. So, the additional unacceptability of an island-crossing-dependency like (7d) – i.e., implicit knowledge of syntactic islands – appears as a superadditive interaction in these types of acceptability judgement plots.

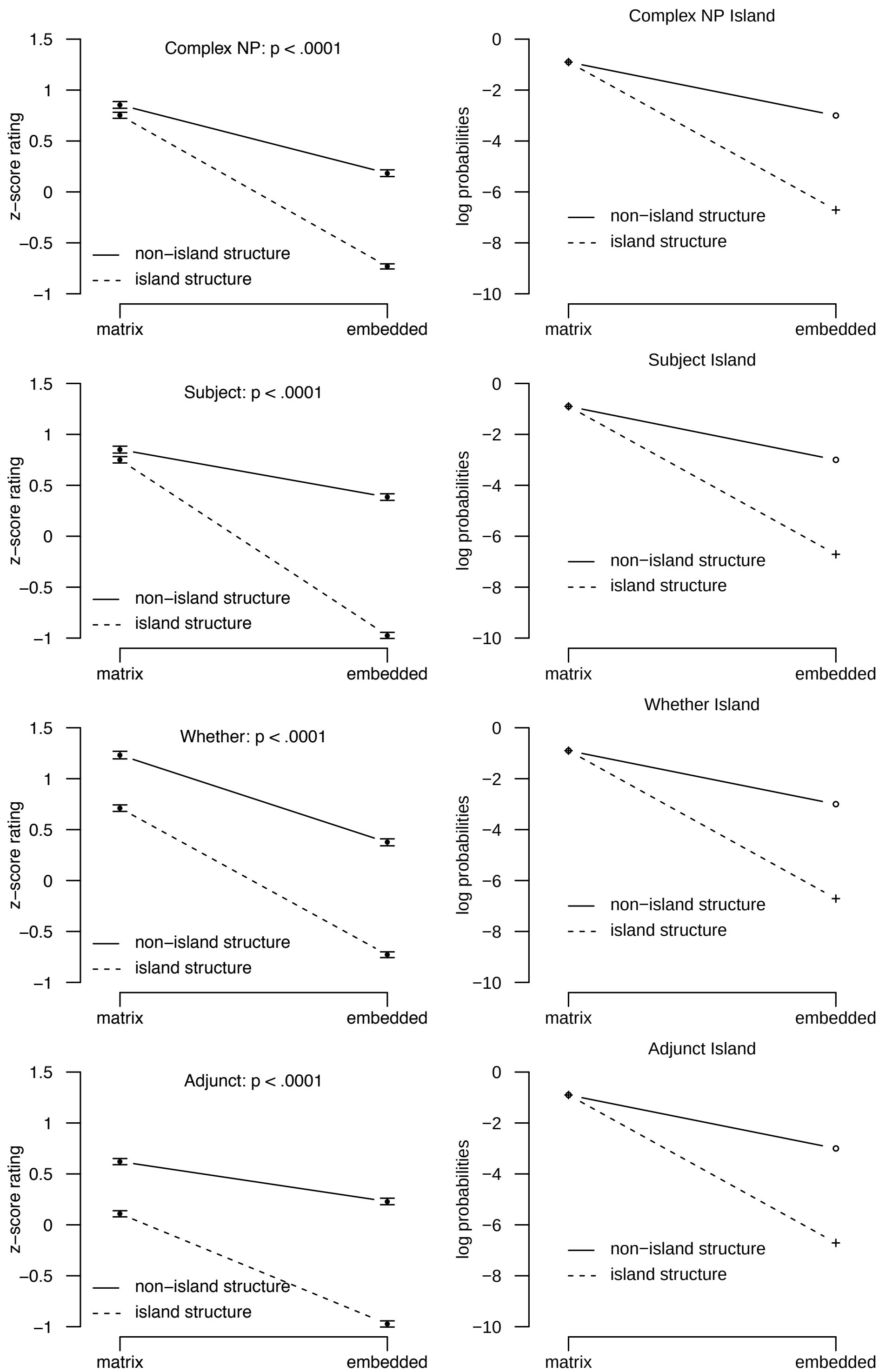
The left column of Figure 1 shows the results of collecting acceptability judgments from high-SES adult speakers using that design. The visible superadditive interactions demonstrate implicit knowledge of the four syntactic islands in (2) in English high-SES adults. The right column of Figure 1 shows the log probability for the same stimuli for each of the four islands, as predicted by the learning model in Pearl and Sprouse (2013). Log probabilities are reported for each dependency because the probabilities are very small numbers (due to the multiplication of syntactic trigram probabilities).[2] The visible superadditive interactions indicate that the high-SES input was sufficient to scaffold the development of these syntactic island constraints.

2.4 Low-SES input quality for syntactic islands

Here we assess low-SES input, focusing on the information necessary for the development of the implicit syntactic islands knowledge that was previously assessed by Pearl and Sprouse (2013) for high-SES input. We first want to identify if there are any quantitative differences between the high-SES and low-SES input samples we have in terms of the *wh*-dependencies and resulting syntactic trigrams available, as these dependencies and trigrams are the foundation of the development of syntactic island constraints. We will answer this question via quantitative analysis of the distribution of *wh*-dependencies and syntactic trigrams available. We then want to identify if there are any qualitative differences between the high-SES and low-SES input in terms of how well the *wh*-dependencies and syntactic trigrams available scaffold the development of syntactic island constraints. That is, whether any quantitative differences exist or not, does low-SES input differ from high-SES input in how it allows complex syntactic development to occur? We'll answer this

[2]For log probabilities, less negative numbers are equivalent to higher probabilities. For example, log(.001) = log(10^{-3}) = -3, while log(.000001) = log(10^{-6}) = -6.

Figure 1: Left column: High-SES adult judgments demonstrating implicit knowledge of four syntactic islands via a superadditive interaction. Right column: Modeled high-SES child judgments demonstrating the same implicit knowledge via a superadditive interaction.



question by applying the same computational learning model from Pearl and Sprouse (2013) that allows successful acquisition of this knowledge from high-SES input. If successful acquisition of island constraints occurs when learning from low-SES input, this would suggest low-SES input isn't qualitatively different from high-SES input in this respect. In contrast, if successful acquisition doesn't occur when learning from low-SES input, this would indicate a qualitative difference for complex syntactic acquisition between low-SES and high-SES input.

2.5 Low-SES CDS input samples

Our low-SES CDS input samples come from a subpart of the HSLLD corpus (Dickinson and Tabors, 2001) in CHILDES (MacWhinney, 2000). SES was defined according to maternal education and annual income. Maternal education ranged from 6 years of schooling to some post-high school education. Annual income didn't have hard lower and upper bounds; instead, 70% of the families reported an annual income of $20,000 or less, while 21% of the families reported an income of over $25,000. The annual income of the remaining 9% was unreported. In this dataset, we focused on the Elicited Report, Mealtime, and Toy Play sections, which represent more naturalistic interactions. We also drew our samples from Home Visit 1, which recorded child language interactions involving children between the ages of three and five. Our sample contained 31,875 utterances and 3,904 *wh*-dependencies (12.2% of all utterances), directed at 78 children between the ages of 3 and 5.

We extracted and syntactically annotated all *wh*-dependencies following the format of the CHILDES Treebank (Pearl and Sprouse, 2013), which indicates the syntactic structure necessary to characterize the syntactic paths of *wh*-dependencies. We then coded the syntactic paths of the dependencies (as in (3b) and shown below with a different example in (11)). Following Pearl and Sprouse (2013), the *CP* phrase structure nodes were further subcategorized by the lexical item serving as complementizer, such as CP_{that}, $CP_{whether}$, CP_{if}, and CP_{null}. This allows the modeled learner of Pearl and Sprouse (2013) to distinguish dependencies judged by high-SES adults to be grammatical, like (11a), from those judged to be ungrammatical, like (11b). With these syntactic paths characterizing *wh*-dependencies, we can then assess the distribution of the *wh*-dependencies in the low-SES input sample.

(11) a. Who do you think $\_\_{who}$ read the book?
syntactic path: *start-IP-VP-*CP_{null}*-IP-end*
b. *Who do you think that $\_\_{who}$ read the book?
syntactic path: **start-IP-VP-*CP_{that}*-IP-end*

3 *Wh*-dependencies and syntactic trigrams across SES

Descriptive corpus analysis. For *wh*-dependencies, our corpus analysis revealed 16 *wh*-dependency types in the low-SES input, 12 of which also appeared in the high-SES corpus analysis of Pearl and Sprouse (2013).[3] Additionally, the low-SES input contained 3 *wh*-dependency types not in the high-SES input:

- *start-IP-VP-*CP_{null}*-IP-VP-NP-PP-end*
  (e.g., *What did he think it was a movie of* $\_\_{what}$*?*)
- *start-IP-VP-IP-VP-IP-VP-PP-IP-VP-end*
  (e.g., *What did you want to try to plan on doing* $\_\_{what}$*?*)
- *start-IP-VP-*CP_{that}*-IP-end*
  (e.g., *What do you think that* $\_\_{what}$ *happens*?)

[3] A more detailed description of the *wh*-dependency distribution across SES is available in Appendix A.

Interestingly, this last dependency type is an example of a "*that*-trace" violation and is judged ungrammatical by high-SES adults (Cowart, 1997). This represents a difference across SES, with respect to specific *wh*-dependencies. Additionally, when we compare the rate (and therefore quantity) of *wh*-dependencies across SES, we find another difference. The *wh*-dependency rate in the high-SES CDS sample of Pearl and Sprouse (2013) was 20.5%, while the *wh*-dependency rate in our low-SES CDS sample was 12.2% – this seems to be a far lower rate. However, there's a striking similarity when we look at the most frequent *wh*-dependencies types across SES: the two dependency types that account for the vast majority of the low-SES *wh*-dependency input (85.8%) are the same two that account for the vast majority of the high-SES input (89.5%), and they occur in about the same proportions (shown in (12)). This suggests a high-level qualitative similarity in the *wh*-dependency input across SES, despite the individual *wh*-dependency differences.

(12) Proportions of the two most frequent *wh*-dependency types across SES

a. 1^{st} most frequent: *start-IP-VP-end* (e.g., *What did Lily read* $\_\_{what}$*?*)
75.5% low-SES, 76.7% high-SES

b. 2^{nd} most frequent: *start-IP-end* (e.g., *What* $\_\_{what}$ *happened?*)
10.3% low-SES, 12.8% high-SES

For syntactic trigrams, which serve as the building blocks of *wh*-dependencies under the Pearl & Sprouse learning strategy, our corpus analysis revealed 21 trigram types in the low-SES input, 14 of which also appeared in the high-SES corpora analyses of Pearl and Sprouse (2013).[4] Additionally, the low-SES input contained 1 syntactic trigram not found in the high-SES input, which comes from one of the dependencies found only in the low-SES input:

- CP_{that}*-IP-end*
  (from *What do you think [that* $\_\_{what}$*)] happens*?)

Notably, just as with the *wh*-dependency analysis, the most frequent syntactic trigrams are very similar across SES. The three trigram types that account for the majority of the trigrams (85.0%) in the low-SES *wh*-dependency input are the same three that account for the majority of the trigrams (87.9%) in the high-SES *wh*-dependency input, and they occur in about the same proportions (shown in (13)). So, as with the *wh*-dependencies, this suggests a high-level qualitative similarity in the syntactic trigram input across SES, despite the individual syntactic trigram differences.

(13) Proportions of the three most frequent trigram types across SES

a. 1^{st} most frequent: *start-IP-VP*
41.4% low-SES, 41.8% high-SES

b. 2^{nd} most frequent: *IP-VP-end*
38.9% low-SES, 40.0% high-SES

c. 3^{rd} most frequent: *start-IP-end*
4.7% low-SES, 6.1% high-SES

Quantitative analysis. To more precisely quantify how similar the input distributions are for both *wh*-dependencies and syntactic trigrams across SES, we use the Jensen-Shannon divergence (**JSDiv**) (Endres and Schindelin, 2003). JSDiv values range from 0 to 1, with 0 indicating identical distributions. That is, higher JSDiv values indicate greater divergence in the distributions, while values closer to 0 indicate distributions that are more similar. In this way, JSDiv analysis provides a way to quantify similarity between distributions; this makes it useful as a comparative measure, where different distributions are assessed for their relative similarity to each other.

With this in mind, we additionally use JSDiv to assess child-directed speech in comparison to adult-directed speech and text, in order to provide a comparison baseline for the similarity across

[4] A more detailed description of the syntactic trigram distribution across SES is available in Appendix B.

input samples of both *wh*-dependencies and syntactic trigrams. In particular, we assess how similar the low-SES and high-SES CDS *wh*-dependency and trigram distributions are to those in high-SES adult-directed speech (**ADS**) and adult-directed text (**ADT**) samples from Pearl and Sprouse (2013). These adult-directed corpora are described in Table 2. This JSDiv analysis will reveal which factors impact *wh*-dependency and syntactic trigram distributions more: SES, whether the speech is directed at children or adults, or whether the input is speech-based vs. text-based.

Table 2: Corpora statistics for low-SES CDS (L-CDS), high-SES CDS (H-CDS), high-SES adult-directed speech (H-ADS), and high-SES adult-directed text (H-ADT) samples.

corpora	# utterances	# *wh*-dependencies	# children	ages
L-CDS	31,875	3,904	78	3 - 5
H-CDS	101,838	20,923	25	1 - 5
H-ADS	74,576	8,508	N/A	N/A
H-ADT	24,243	4,230	N/A	N/A

Table 3: The nine *wh*-dependencies shared across all four corpora that are used in the JSDiv analysis.

Shared dependencies	Example utterance	Corpora percentage
start-IP-*end*	*Who saw it?*	10.3% - 33.0%
start-IP-VP-*end*	*Who did she see?*	63.3% - 76.7%
start-IP-VP-CP_{null}-IP-*end*	*Who did he think stole it?*	0.1% - 0.6%
start-IP-VP-CP_{null}-IP-VP-*end*	*What did he think she stole?*	0.2% - 1.1%
start-IP-VP-CP_{null}-IP-VP-PP-*end*	*What did he think she wanted it for?*	<0.1% - 0.1%
start-IP-VP-IP-VP-*end*	*What did he want her to steal?*	1.3% - 7.5%
start-IP-VP-IP-VP-IP-VP-*end*	*What did he want her to pretend to steal?*	<0.1%
start-IP-VP-IP-VP-PP-*end*	*What did she want to get out from under?*	<0.1% - 0.8%
start-IP-VP-PP-*end*	*Who did she steal from?*	1.3% - 4.3%

***Wh*-dependencies.** Figure 2 shows the results of the JSDiv analysis for *wh*-dependencies, calculated over the distribution of the 9 *wh*-dependencies (shown in Table 3) that these four corpora had in common. We see that low-SES CDS and high-SES CDS are the most similar in *wh*-dependency distribution (JS: 0.00445), and appear to be twice as similar as the next closest comparison, which is high-SES CDS vs. high-SES ADS (JS: 0.00948). This affirms a quantitative similarity across SES in child *wh*-dependency input, in terms of *wh*-dependency distribution. Moreover, these results highlight that CDS *across* SES is more similar than CDS vs. ADS *within* SES. That is, whether the speech is directed at children or adults matters more than whether speech is coming from a high-SES or low-SES population. We also note that these JSDivs accord with intuitions that speech of any kind is more similar to other speech than it is to text: high-SES ADS diverges more from high-SES ADT (JS: 0.03156) than it does from either high-SES CDS (JS: 0.00948) or low-SES CDS (JS: 0.01576).

Syntactic trigrams. Figure 3 shows the results of the JSDiv analysis for syntactic trigrams, calculated over the distribution of the 14 trigrams shown in Table 4 (see Table 7 in Appendix B for the full list of trigrams) that these four corpora had in common across all *wh*-dependencies. These trigrams accounted for 99.5 %-99.8% of the total trigrams in these corpora. As with the analysis of the dependencies, we see the same pattern emerge: (i) low-SES CDS is more similar to high-SES CDS (JSDiv: 0.00850) than any other input type, and (ii) all speech is more similar to other types of speech than to text (speech vs. speech: JSDiv=0.00850-0.02836; speech vs. text: JSDiv=0.07183-0.16279).

Figure 2: JSDiv analyses for low-SES CDS (L-CDS), high-SES CDS (H-CDS), high-SES adult-directed speech (H-ADS), and high-SES adult-directed text (H-ADT). Line thickness corresponds to similarity, with thicker lines indicating more similar distributions.

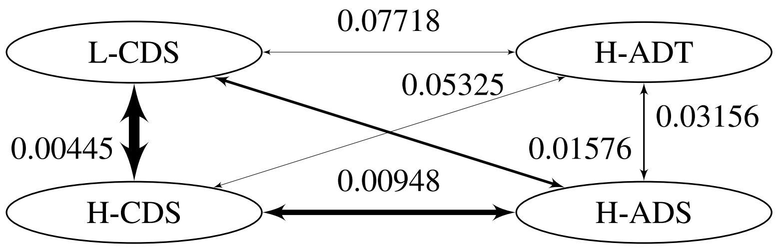
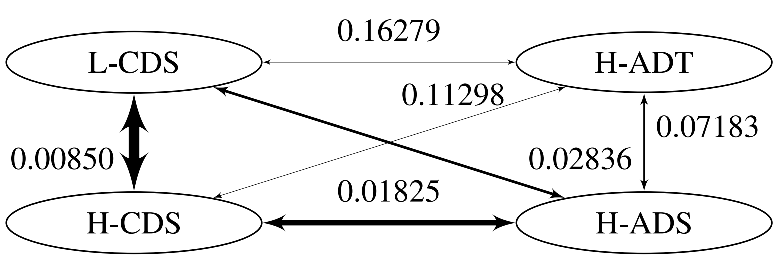


Table 4: Distribution of the 14 syntactic trigrams across child-directed Low-SES (L-CDS) and High-SES (H-CDS), as well as High-SES adult-directed speech (H-ADS) and text (H-ADT).

Syntactic trigrams	Syntactic trigram percentage
CP_{null}-*IP-VP*	0.1% - 0.7%
CP_{null}-*IP-end*	<0.1% - 0.3%
IP-VP-CP_{null}	0.3% - 0.7%
IP-VP-CP_{that}	<0.1%
IP-VP-IP	0.9% - 4.0%
IP-VP-NP	<0.1% - 0.1%
IP-VP-PP	0.8% - 2.5%
IP-VP-end	38.5% - 39.9%
VP-CP_{null}*-IP*	0.3 - 0.7%
VP-CP_{that}*-IP*	<0.1%
VP-IP-VP	0.9% - 4.0%
VP-PP-end	0.8% - 2.3%
start-IP-VP	38.6% - 41.7%
start-IP-end	4.7% - 19.0%

Figure 3: JSDiv analyses for low-SES CDS (L-CDS) trigrams, high-SES CDS (H-CDS) trigrams, high-SES adult-directed speech (H-ADS) trigrams, and high-SES adult-directed text (H-ADT) trigrams. Line thickness corresponds to similarity, with thicker lines indicating more similar distributions.



Quantitative analysis summary. Our quantitative analyses suggest that the input children encounter for learning about syntactic islands is very similar across SES. In particular, both the *wh*-distributions and the syntactic trigram distributions appear quite similar, despite some individual *wh*-dependency and trigram differences. However, it's unclear if even these comparatively small differences may lead to different qualitative outcomes. That is, using the low-SES distributions, do we predict children's acquisition of syntactic island knowledge will be the same as what high-SES children are predicted to learn? To assess the qualitative similarity of high-SES and low-SES CDS input with respect to predicted learning outcomes, we need to use the low-SES input distribution to predict what low-SES children could learn about syntactic islands.

4 Predicting what low-SES children could learn about syntactic islands

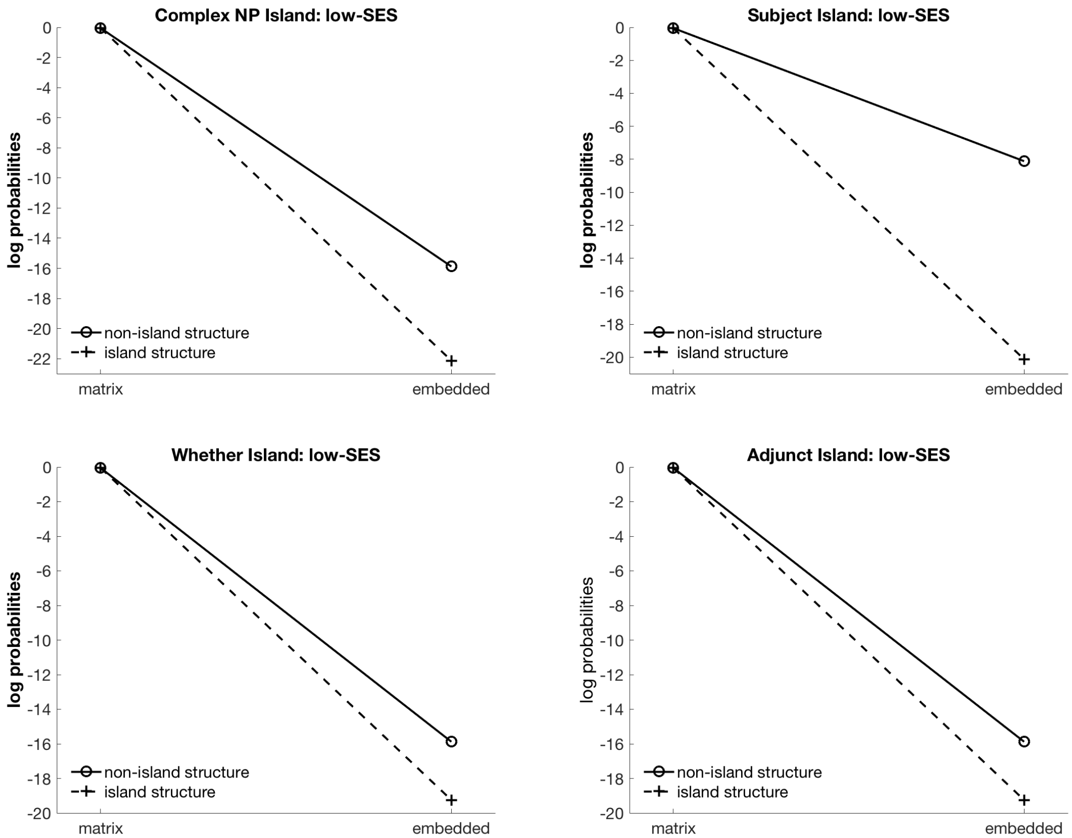
We use the same cognitive learning model developed by Pearl and Sprouse (2013); the modeled learner learns from the *wh*-dependency distribution in low-SES CDS input and generates probabilities for the four sets of experimental stimuli of Sprouse et al. (2012), which correspond to Complex NP, Subject, Whether, and Adjunct islands. Recall that these experimental stimuli can be characterized by the syntactic paths shown in Table 1, where many of the grammatical dependencies are characterized by the same syntactic path (e.g., *start*-IP-*end* for both matrix+non-island and matrix+island); this is why Table 5, which shows the modeled learner's generated log probabilities of the relevant *wh*-dependencies, has only three grammatical dependency syntactic paths listed. Figure 4 shows the low-SES CDS log probabilities plotted on interaction plots for each of the four island types. To aid comparison of predicted learning outcomes across SES, Table 5 also shows the log probabilities generated by learners learning from the high-SES CDS, as well as the high-SES ADS and ADT reported in Pearl and Sprouse (2013).

Table 5: Log probabilities of different *wh*-dependencies, representing acceptability judgments, for modeled learners learning from low-SES child-directed speech (L-CDS), as well as prior results from Pearl & Sprouse (2013) of modeled learners learning from high-SES child-directed speech (H-CDS) and high-SES adult-directed speech and text (H-ADS+H-ADT).

	L-CDS	H-CDS	H-ADS + H-ADT
Grammatical dependencies			
start-IP-*end*	-0.48	-1.21	-0.93
start-IP-VP-CP$_{null}$-IP-*end*	-8.11	-7.89	-7.67
start-IP-VP-CP$_{that}$-IP-VP-*end*	-15.88	-13.84	-11.00
Island-spanning dependencies			
start-IP-VP-NP-CP$_{that}$-IP-VP-*end*	-22.13	-19.81	-18.93
start-IP-VP-CP$_{null}$-IP-NP-PP-*end*	-20.12	-20.17	-20.36
start-IP-VP-CP$_{whether}$-IP-VP-*end*	-19.25	-18.54	-18.46
start-IP-VP-CP$_{if}$-IP-VP-*end*	-19.25	-18.54	-18.46

We can see that a core pattern emerges when learning from low-SES CDS: all grammatical dependencies have higher probabilities (equivalent to less negative log probabilities) than the island-spanning dependencies. In particular, grammatical dependencies have log probabilities ranging from -0.48 to -15.88, while island-spanning dependencies range from -19.25 to -22.13. So, even the least acceptable grammatical dependency (with log probability -15.88) is predicted to be over 2000 times more acceptable than the most acceptable ungrammatical dependency (with log probability -19.25), because $\frac{10^{-15.88}}{10^{-19.25}} \approx 2344$. This is the same pattern which was found when learning from either high-SES child-directed or adult-directed input (high-SES grammatical: -0.93 to -13.84; high-SES island-spanning: -18.46 to -20.36). Importantly, in Figure 4, we see the superadditivity that indicates implicit knowledge of syntactic island constraints. That is, just as with the log probabilities generated from the high-SES data and the acceptability judgments from high-SES

Figure 4: Judgments derived from a modeled learner using low-SES CDS, demonstrating implicit knowledge of syntactic islands as indicated by superadditivity (which appears as non-parallel lines in these interaction plots).



adults, island-spanning dependencies are more unacceptable than would be predicted, given that they're embedded dependencies and they have an island structure in the utterance. This affirms what the JSDiv analysis between the low-SES and high-SES CDS *wh*-dependencies suggested: the input quality is the same across SES, with respect to the development of the complex syntactic knowledge of syntactic island constraints.

Additionally, although quantitative differences exist across ADS/ADT and CDS, these quantitative differences also don't impact predicted qualitative learning outcomes. For example, the low-SES CDS *wh*-dependency distribution is 17 times more similar to the high-SES CDS *wh*-dependency distribution than it is to high-SES ADT *wh*-dependency distribution, based on the JSDiv. Yet, all four *wh*-dependency distributions (low- and high-SES CDS, high-SES ADS, and high-SES ADT) contain the necessary information for the cognitive learning model to learn syntactic island constraints. This indicates that even larger JSDiv differences lead to the same predicted learning outcomes. This in turn suggests that learning syntactic islands knowledge may be fairly robust to input variation.

5 Discussion

Our results suggest that the *wh*-dependency input, and in turn the syntactic trigram input, that low-SES children receive is quantitatively and qualitatively similar to the input of high-SES children. This similarity allows a modeled child to learn implicit knowledge of syntactic islands from low-SES input just as easily as from high-SES input, as demonstrated by the modeled judgment behavior. So, these cognitive modeling results serve as predictions of children's learning behavior for syntactic islands, and predict no learning outcome differences due to input differences across SES.

Interestingly, there's a striking difference in the exact *wh*-dependency distribution across SES that turns out to be crucial for acquisition success for two of the syntactic island types. This difference involves a particular structural building block, which comes from dependencies that are characterized with CP_{that}.

As noted in (11), the only distinction between certain dependencies judged grammatical and certain dependencies judged ungrammatical by high-SES adults is the complementizer. Example (11) showed this for a grammatical dependency with the *null* complementizer and an ungrammatical dependency with complementizer *that*. Another key example is the difference between grammatical dependencies with complementizer *that* (14a) and ungrammatical dependencies with complementizers like *whether* (whether islands) or *if* (adjunct islands) (14b). Again, the only difference in the syntactic path of these dependencies is the CP building block, which is CP_{that} for the dependency judged grammatical and $\text{CP}_{whether}$ or CP_{if} for the dependencies judged ungrammatical.

(14) a. What do you think that Jack read $\_\__{what}$?
syntactic path: *start-IP-VP-*$\textit{CP}_{that}$*-IP-VP-end*
b. *What do you wonder whether/if Jack read $\_\__{what}$?
syntactic path: **start-IP-VP-*$\textit{CP}_{whether/if}$*-IP-VP-end*

So, it's important that the child encounter *wh*-dependencies in her input that involve complementizer *that* (and not ones that involve complementizers *whether* or *if*). When this happens, the probabilistic learning strategy we used here can leverage the CP_{that} building block to predict that (14a) should be judged as better than (14b). However, dependencies involving CP_{that} are actually fairly rare in naturalistic usage. Pearl and Sprouse (2013) only found 2 of 20,923 (0.0096%) in high-SES CDS, 7 of 8,508 (0.082%) in high-SES ADS, and 2 of 4,230 (0.048%) in high-SES ADT.

In the high-SES CDS sample, both dependencies involving CP_{that} are of the same type: *start-IP-VP-*$\textit{CP}_{that}$*-IP-VP-end* instances like (14a). However, in our low-SES CDS sample, there are 2 of 3,094 (0.051%) dependencies involving CP_{that}, and they are both of a different type, which happens to be judged ungrammatical by high-SES adults: *start-IP-VP-*$\textit{CP}_{that}$*-IP-end* instances like (15).

(15) What do you think that $\_\__{what}$ happens?
What do $[_{IP}$ you $[_{VP}$ think $[_{CP_{that}}$ that $[_{IP}$ $\_\__{what}$ $[_{VP}$ happens]]]]]?
syntactic path: *start-IP-VP-*$\textit{CP}_{that}$*-IP*

So, the presence of this *wh*-dependency type, which is ungrammatical in the high-SES dialect, provides the crucial CP_{that} building block necessary for the acquisition of whether and adjunct islands. That is, the key linguistic experience that would allow a child learning from low-SES CDS to acquire the same syntactic knowledge as a high-SES child actually comes from data that's ungrammatical for a high-SES child. This underscores the power of learning strategies that generate linguistic knowledge of larger structures from smaller building blocks; a child relying on smaller building blocks may be able to find evidence for those building blocks in unexpected places.

More generally, our results indicate that the input for the development of complex syntactic knowledge may not differ in impactful ways across SES, the way that it does for lexical or more foundational syntactic knowledge. That is, there may not be a "complex syntax gap" across SES. For instance, a difference in input quantity, as indicated by the relative rate of *wh*-dependencies doesn't appear to be a meaningful one, as shown by our quantitative analyses using JSDiv. Moreover, there doesn't appear to be any difference with respect to input quality, based on our cognitive modeling results. So, while there may be some surface-level quantitative differences in input across SES, there don't appear to be qualitative differences. Interestingly, this is the mirror image of input differences that seem to exist for foundational syntactic and lexical knowledge; in particular, there seem to be qualitative input differences for foundational syntactic and lexical knowledge

(Huttenlocher et al., 2007; Rowe, 2012), but not quantitative differences.

In the specific case of learning about syntactic islands, we would predict that once low-SES children are able to leverage the *wh*-dependency information in their input, they should learn about these syntactic islands as well as high-SES children do. That is, the target knowledge low-SES children eventually achieve for syntactic islands is predicted to be the same as that of high-SES children (and adults), even if the low-SES target knowledge may differ for other syntactic knowledge like *that*-trace violations. A first step for evaluating this prediction is to collect judgment data from low-SES adults for syntactic islands in general and for *that*-trace violations specifically. We would expect low-SES adult judgments to be the same as high-SES adult judgments, except for the *that*-trace violation, which low-SES adults should find grammatical.

We note that the ability to leverage the *wh*-dependency and syntactic trigram information isn't trivial – there are known delays in language processing in low-SES children compared to their high-SES counterparts (Fernald et al., 2013; Weisleder and Fernald, 2013). These delays could lead to low-SES children being less able to harness the complex syntactic information available in their input, even if the information is in fact there. However, our results here suggest that once the developmental milestones are met which allow successful processing of the available *wh*-dependency information in low-SES children's input, no other gap remains in low-SES children's input.

More concretely, the syntactic islands learning strategy applied here to the low-SES CDS data requires several foundational knowledge components and processing abilities to be "good enough" – that is, what the child must both know and be able to do in real time. First, the child must know about syntactic phrase structure; she must be able to use that phrase structure knowledge to extract the syntactic path of a *wh*-dependency in real time. Second, the child must know to break syntactic paths into smaller trigram building blocks that can be used to generate a probability for any *wh*-dependency; she must be able to identify these syntactic trigrams in real time. Third, the child must know to track the relative frequency of the syntactic trigrams; she must be able to track these frequencies in real time. Fourth, the child must know to combine these syntactic trigrams to generate the probability for a new *wh*-dependency; she must be able to do so in real time. Any or all of these components could be affected by processing deficits that arise from input quantity and quality differences in low-SES CDS, and it remains an open question which ones are in fact adversely affected by low-SES children's prior linguistic experience. Still, our current work has demonstrated that once low-SES children can use the *wh*-dependency information available to them, their input wouldn't cause them to lag behind their high-SES counterparts when it comes to learning about complex syntactic knowledge like syntactic islands.

6 Conclusion

We have aimed to broaden the body of research on linguistic input quantity and quality across SES to include the nature of the input for more complex syntactic knowledge. We focused on the development of syntactic islands knowledge, which depends on the *wh*-dependencies in children's input. To our knowledge, this is the first comparison across SES for this type of complex syntactic knowledge. More specifically, we examined the distribution of *wh*-dependencies in American English CDS to children from low- and high-SES populations. While we found some quantitative differences – perhaps most notably the rate at which *wh*-dependencies occur in the input – these differences aren't predicted to lead to different learning outcomes. So, despite any quantitative differences, there are no predicted qualitative differences across SES for learning these syntactic islands. That is, surface quantitative differences mask underlying qualitative similarities across SES in terms of what the input is used for. This is turn suggests that if we do see developmental differences in syntactic island knowledge across SES, it's not because of the information available in the input. Instead, children's ability to harness that information may differ. In short, the information is there for children to use, no matter their SES – a key developmental step is for them to figure out how to use it.

References

SD Blum. 2015. Wordism: Is there a teacher in the house. *Journal of Linguistic Anthropology* 25(1):74–75.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Noam Chomsky. 1973. Conditions on transformations. In S. Anderson and P. Kiparsky, editors, *A Festschrift for Morris Halle*, Holt, Rinehart, and Winston, New York, pages 237–286.

Wayne Cowart. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgements*. Thousand Oaks, CA: Sage.

Jill de Villiers, Thomas Roeper, Linda Bland-Stewart, and Barbara Pearson. 2008. Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics* 29(1):67–103.

David K Dickinson and Patton O Tabors. 2001. *Beginning literacy with language: Young children learning at home and school.*. Paul H Brookes Publishing.

Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory* .

Anne Fernald, Virginia A Marchman, and Adriana Weisleder. 2013. SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science* 16(2):234–248.

Betty Hart and Todd Risley. 1995. *Meaningful differences in the everyday experience of young American children*. P.H. Brookes, Baltimore, MD.

Erika Hoff. 2003. The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development* 74(5):1368–1378.

Janellen Huttenlocher, Marina Vasilyeva, Elina Cymerman, and Susan Levine. 2002. Language input and child syntax. *Cognitive Psychology* 45(3):337–374.

Janellen Huttenlocher, Marina Vasilyeva, Heidi R Waterfall, Jack L Vevea, and Larry V Hedges. 2007. The varieties of speech to young children. *Developmental Psychology* 43(5):1062.

Janellen Huttenlocher, Heidi Waterfall, Marina Vasilyeva, Jack Vevea, and Larry V Hedges. 2010. Sources of variability in children's language growth. *Cognitive Psychology* 61(4):343–365.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Lisa Pearl and Jon Sprouse. 2013. Computational Models of Acquisition for Islands. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Islands Effects*, Cambridge University Press, Cambridge, pages 109–131.

Lisa Pearl and Jon Sprouse. 2015. Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research* 58:740–753.

John Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT, Cambridge, MA.

Meredith L Rowe. 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development* 83(5):1762–1774.

Meredith L Rowe, Kathryn A Leech, and Natasha Cabrera. 2017. Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive Science* 41:162–179.

Jessica F Schwab and Casey Lew-Williams. 2016. Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science* 7:264–275.

Douglas E Sperry, Linda L Sperry, and Peggy J Miller. 2018. Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child development* .

Jon Sprouse, Matt Wagers, and Colin Phillips. 2012. A test of the relation between working memory capacity and syntactic island effects. *Language* 88(1):82–124.

Patrick Suppes. 1974. The semantics of children's language. *American Psychologist* 29:103–114.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40(1):21–81.

Adriana Weisleder and Anne Fernald. 2013. Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science* 24(11):2143–2152.

Appendices

A *Wh*-dependency distribution across SES

Table 6 shows the distribution of *wh*-dependencies across the different corpora, including the low-SES and high-SES child-directed speech, as well as high-SES adult-directed speech and adult-directed text.

Table 6: Distribution of *wh*-dependencies in child-directed Low-SES (L-CDS) and High-SES (H-CDS), as well as High-SES adult-directed speech (H-ADS) and text (H-ADT). Percentages are shown for syntactic paths, based on the total *wh*-dependencies in each corpus, with the quantity observed in the corpus on the line below. An example of each syntactic path is given below the path. Dependencies used in the Jensen-Shannon divergence (JSDiv) analysis are in teal. The dependency in the Low-SES dialect that's judged to be ungrammatical in the High-SES dialect is in pink.

Distribution of *wh*-dependencies in the input				
Syntactic path and example utterance	L-CDS	H-CDS	H-ADS	H-ADT
IP Who saw it?	10.3% 402	12.8% 2680	17.2% 1464	33.0% 1396
IP-VP What did she see?	75.5% 2949	76.7% 16039	73.0% 6215	63.3% 2677
IP-VP-AdjP-IP-VP What are you willing to see?	0.0% 0	0.0% 0	<0.1% 1	0.1% 5
IP-VP-AdjP-IP-VP-PP What are you willing to go to?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-AdjP-PP What are they good for?	0.0% 0	0.0% 0	<0.1% 1	<0.1% 1
IP-VP-CP$_{for}$-IP-VP-PP What did she put on for you to dance to?	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP$_{null}$-IP Who did he think stole it?	0.1% 5	0.1% 24	0.6% 52	0.3% 12
IP-VP-CP$_{null}$-IP-VP What did he think she stole?	0.9% 39	1.1% 236	0.4% 30	0.2% 8
IP-VP-CP$_{null}$-IP-VP-IP-VP What did he think she wanted to steal?	<0.1% 3	0.1% 28	<0.1% 3	0.0% 0
IP-VP-CP$_{null}$-IP-VP-IP-VP-IP-VP What did he think she wanted to pretend to steal?	0.0% 0	<0.1% 2	0.0% 0	0.0% 0
IP-VP-CP$_{null}$-IP-VP-IP-VP-IP-VP-PP Who did he think she wanted to pretend to steal from?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-CP$_{null}$-IP-VP-IP-VP-PP Who did he think she wanted to steal from?	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP$_{null}$-IP-VP-NP What did he think she said about it?	0.0% 0	<0.1% 1	<0.1% 5	<0.1% 1
IP-VP-CP$_{null}$-IP-VP-NP-PP What did he think it was a movie of?	<0.1% 3	0.0% 0	0.0% 0	0.0% 0
IP-VP-CP$_{null}$-IP-VP-PP What did he think she wanted it for?	0.1% 4	0.1% 28	<0.1% 5	<0.1% 1
IP-VP-CP$_{null}$-IP-VP-PP-PP	0.0%	<0.1%	0.0%	0.0%

Distribution of *wh*-dependencies in the input				
Syntactic path and example utterance	L-CDS	H-CDS	H-ADS	H-ADT
What did he think she wanted out of?	0	1	0	0
IP-VP-CP$_{that}$-IP What do you think that happens?	<0.1% 2	0.0% 0	0.0% 0	0.0% 0
IP-VP-CP$_{that}$-IP-VP What did he think that she stole?	0.0% 0	<0.1% 2	<0.1% 5	<0.1% 2
IP-VP-CP$_{that}$-IP-VP-IP-VP What did he think that she wanted to steal?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-CP$_{that}$-IP-VP-PP Who did he think that she wanted to steal from?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-IP Who did he want to steal the necklace?	0.0% 0	<0.1% 9	<0.1% 2	0.0% 0
IP-VP-IP-VP What did he want her to steal?	7.5% 296	5.6% 1167	3.4% 287	1.3% 57
IP-VP-IP-VP-IP-VP What did he want her to pretend to steal?	<0.1% 2	<0.1% 11	<0.1% 6	<0.1% 1
IP-VP-IP-VP-IP-VP-PP Who did he want her to pretend to steal from?	0.0% 0	0.2% 43	<0.1% 6	0.0% 0
IP-VP-IP-VP-IP-VP-PP-IP-VP What did you want to try to plan on doing?	<0.1% 1	0.0% 0	0.0% 0	0.0% 0
IP-VP-IP-VP-NP What did he want to say about it?	0.0% 0	<0.1% 6	0.0% 0	0.0% 0
IP-VP-IP-VP-NP-IP-VP What did he have to give her the opportunity to steal?	0.0% 0	0.0% 0	0.0% 0	<0.1% 1
IP-VP-IP-VP-NP-PP What did she want to steal more of?	0.0% 0	<0.1% 1	<0.1% 1	0.0% 0
IP-VP-IP-VP-PP What did she want to steal from?	0.8% 35	0.4% 74	0.4% 33	<0.1% 4
IP-VP-IP-VP-PP-PP What did she want to get out from under?	0.0% 0	0.0% 0	0.0% 0	<0.1% 1
IP-VP-NP What did she say about the necklace?	0.0% 0	0.2% 52	0.1% 10	0.1% 5
IP-VP-NP-IP-VP What did he give her the opportunity to steal?	0.0% 0	0.0% 0	<0.1% 1	<0.1% 2
IP-VP-NP-PP What was she a member of?	<0.1% 1	<0.1% 7	<0.1% 6	0.0% 0
IP-VP-PP Who did she steal from?	4.0% 159	2.5% 524	4.3% 369	1.3% 57
IP-VP-PP-CP$_{null}$-IP What did she feel like was a very good place?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-PP-CP$_{null}$-IP-VP What did she feel like he saw?	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
IP-VP-PP-IP-VP What did she think about buying?	<0.1% 2	0.0% 0	<0.1% 3	0.0% 0
IP-VP-PP-NP Where was she at in the building?	0.0% 0	0.0% 0	<0.1% 2	0.0% 0

Distribution of *wh*-dependencies in the input				
Syntactic path and example utterance	L-CDS	H-CDS	H-ADS	H-ADT
IP-VP-PP-NP-PP What do you put it on top of?	0.0% 0	<0.1% 2	0.0% 0	0.0% 0
IP-VP-PP-NP-PP-IP-VP What is she in the habit of doing?	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
IP-VP-PP-PP What does he eat out of?	0.5% 1	0.1% 22	0.0% 0	0.0% 0
IP-VP-PP-IP-VP What did he think about stealing?	0.0% 0	<0.1% 1	0.0% 0	0.0% 0

B Syntactic trigram distribution across SES

Table 7 shows the distribution of the syntactic trigrams across the different corpora, including the low-SES and high-SES child-directed speech, as well as high-SES adult-directed speech and adult-directed text. The shared syntactic trigrams were used when calculating the Jensen-Shannon divergence (JSDiv) analyses.

Table 7: Distribution of the syntactic trigrams across child-directed Low-SES (L-CDS) and High-SES (H-CDS), as well as High-SES adult-directed speech (H-ADS) and text (H-ADT). The 14 shared trigrams used in the JSDiv analysis are in teal.

Distribution of trigrams in the input				
Trigrams	L-CDS	H-CDS	H-ADS	H-ADT
AdjP-IP-VP	0.0% 0	0.0% 0	<0.1% 2	<0.1% 5
AdjP-PP-end	0.0% 0	0.0% 0	<0.1% 1	<0.1% 1
CP_{for}-IP-VP	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
CP_{null}-IP-VP	0.6% 49	0.7% 298	0.2% 44	0.1% 10
CP_{null}-IP-end	<0.1% 5	<0.1% 24	0.3% 53	0.2% 12
CP_{that}-IP-VP	0.0% 0	<0.1% 2	<0.1% 7	<0.1% 2
CP_{that}-IP-end	<0.1% 2	0.0% 0	0.0% 0	0.0% 0
IP-VP-AdjP	0.0% 0	0.0% 0	<0.1% 3	<0.1% 6
IP-VP-CP_{for}	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP_{null}	0.6% 54	0.7% 321	0.6% 96	0.3% 22
IP-VP-CP_{that}	<0.1% 2	<0.1% 2	<0.1% 7	<0.1% 2
IP-VP-IP	4.0% 340	3.2% 1398	2.1% 353	0.9% 65

Distribution of trigrams in the input				
Trigrams	L-CDS	H-CDS	H-ADS	H-ADT
IP-VP-NP	<0.1% 4	0.1% 67	0.1% 23	0.1% 9
IP-VP-PP	2.4% 202	1.6% 698	2.5% 423	0.8% 63
IP-VP-end	38.9% 3292	39.9% 17487	38.5% 6553	37.4% 2753
NP-IP-VP	0.0% 0	0.0% 0	<0.1% 1	<0.1% 3
NP-PP-IP	0.0% 0	0.0% 0	<0.1% 1	0.0% 0
NP-PP-end	<0.1% 4	<0.1% 10	<0.1% 7	0.0% 0
PP-CP_{null}-IP	0.0% 0	<0.1% 1	<0.1% 1	0.0% 0
PP-IP-VP	<0.1% 3	<0.1% 1	<0.1% 4	0.0% 0
PP-NP-PP	0.0% 0	<0.1% 2	<0.1% 1	0.0% 0
PP-NP-end	0.0% 0	0.0% 0	<0.1% 2	0.0% 0
PP-PP-end	<0.1% 1	<0.1% 23	0.0% 0	<0.1% 1
VP-AdjP-IP	0.0% 0	0.0% 0	<0.1% 2	<0.1% 5
VP-AdjP-PP	0.0% 0	0.0% 0	<0.1% 1	<0.1% 1
VP-CP_{for}-IP	0.0% 0	<0.1% 1	0.0% 0	0.0% 0
VP-CP_{null}-IP	0.6% 54	0.7% 321	0.6% 96	0.3% 22
VP-CP_{that}-IP	<0.1% 2	<0.1 2	<0.1% 7	<0.1 2
VP-IP-VP	4.0% 340	3.2% 1389	2.1% 351	0.9% 65
VP-IP-end	0.0% 0	<0.1% 9	<0.1% 2	0.0% 0
VP-NP-IP	0.0% 0	0.0% 0	<0.1% 1	<0.1% 3
VP-NP-PP	<0.1 4	<0.1% 8	<0.1 7	0.0% 0
VP-NP-end	0.0% 0	0.1% 59	<0.1% 15	<0.1% 6
VP-PP-CP_{null}	0.0% 0	<0.1% 1	<0.1% 1	0.0% 0

Distribution of trigrams in the input				
Trigrams	L-CDS	H-CDS	H-ADS	H-ADT
VP-PP-IP	<0.1 3	<0.1% 1	<0.1 3	0.0% 0
VP-PP-NP	0.0% 0	<0.1% 2	<0.1% 3	0.0% 0
VP-PP-PP	<0.1 1	<0.1% 23	0.0% 0	<0.1 1
VP-PP-end	2.3% 198	1.5% 671	2.4% 416	0.8% 62
start-IP-VP	41.4% 3502	41.7% 18283	41.5% 7049	38.6% 2835
start-IP-end	4.7% 402	6.1% 2680	8.6% 1464	19.0% 1396