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

Lisa Pearl
University of California, Irvine

Abstract

Generative approaches to language have long recognized the natural link between theories of knowledge representation and theories of knowledge acquisition. The basic idea is that the knowledge representations provided by Universal Grammar enable children to acquire language as reliably as they do because these representations highlight the relevant aspects of the available linguistic data. So, one reasonable evaluation of any theory of representation is how useful it is for acquisition. This means that when we have multiple theories for how knowledge is represented, we can try to evaluate these theoretical options by seeing how children might use them during acquisition. Computational models of the acquisition process are an effective tool for determining this, since they allow us to incorporate the assumptions of a representation into a cognitively plausible learning scenario and see what happens. We can then identify which representations work for acquisition, and what those representations need to work. This in turn allows us to refine both our theories of how knowledge is represented and how those representations are used by children during acquisition. I discuss two case studies of this approach for representations in metrical stress and syntax, and consider what we learn from this computational acquisition evaluation in each domain.

1 Introduction

A core premise of generative linguistic theorizing is that the knowledge representations provided by Universal Grammar (UG) are what makes acquisition happen so fast and so well. This is based on the natural dependence between theories of representation and theories of acquisition (Chomsky 1969; Pinker 1979; Chomsky 1986; Osherson et al. 1986; Chomsky and Lasnik 1995; Dresher 1999; Crain and Pietroski 2002; Heinz 2014). The basic idea is that children armed with the linguistic variables of the knowledge representation should have a huge advantage when it comes to learning their language-specific grammars. This is because UG defines the hypothesis space of possible grammars in terms of the *relevant* linguistic variables (e.g., parameters, constraints, or rules). So, children already know what part of the encoded input matters – it's the part that corresponds to those linguistic variables. This provides a convenient initial filter on children's input: their *acquisitional intake* (Lidz and Gagliardi 2015) is constrained by the linguistic variables in the knowledge representation. If some piece of the input doesn't relate to a linguistic variable in the representation, it can be filtered out. This highlights the acquisition implications for theories of representation: knowledge representations impact the way children view their linguistic data.

Given this, one reasonable way to evaluate different theories of knowledge representation is to see how useful they are for acquisition. Given a particular representation, is the child's hypothesis space helpfully constrained? Given that the linguistic variables of the hypothesis space impact a child's acquisitional intake, is that acquisitional intake sufficient for acquisition to happen successfully? The problem, of course, is that acquisition is complicated, so evaluating how useful a representation is for acquisition isn't easy. However, if we can find a reasonable way to do it, there's a lot to be gained for both theories of representation and theories of the acquisition processes that accompany those representations.

First, we would have a new metric for evaluating representational theories that leverages this link to acquisition. Second, because this evaluation requires us to be concrete about how acquisition works for a particular representation, we become aware of the learning assumptions each representation depends on. This lets us refine our theories of acquisition for a given representation, and can also make empirical predictions about how acquisition should unfold if one representation is being used vs. another.

Below I describe an implementation of this kind of evaluation via computational modeling. The goals are to (a) simulate a learner who incorporates the linguistic variables of a representation, (b) set that learner up in a cognitively plausible acquisition scenario, and (c) see if acquisition succeeds. I discuss two case studies of this approach, evaluating representations in metrical stress and syntax.

The metrical stress study, based on Pearl et al. (2014, 2015), compares representations on their ability to allow acquisition of the productive aspects of English metrical stress from the kind of data English children typically encounter. This computational acquisition evaluation identifies (i) learning assumptions that benefit children using different representations, and (ii) English-like grammars within these representations that are easier to learn from English child-directed speech than the current English grammar definitions. These English-like grammars are then empirically motivated alternative proposals for the grammar of English.

The syntax case study, based on Pearl and Sprouse (2013a,b, 2015), evaluates a theory of

dependency representation on its ability to allow acquisition of syntactic islands in English from the kind of data English children typically encounter. This evaluation identifies (i) the level of specificity when representing dependencies that enables a child to learn syntactic islands, and (ii) the learning assumptions required for success, only some of which are necessarily part of UG.

More generally, both case studies (i) validate knowledge representations by showing how acquisition can proceed using them, and (ii) provide computational modeling feedback about the nature of the English-specific representations.

2 Case study: Metrical stress

2.1 English as a non-trivial test case

For English metrical stress, the target knowledge for our purposes is the ability to account for word-level stress patterns (e.g., *octopus* pronounced *óctopus*, with the first syllable stressed and the remaining syllables unstressed). The observable data English children encounter are the stress contours associated with lexical items (e.g., if we indicate stressed syllables with 1 and unstressed syllables with 0, the stress contour of *óctopus* would be 100).

Theories of metrical stress representation specify the underlying structure that generates the observable stress contour for words of a language. For example, the three theories I'll discuss below all posit an underlying structure that involves grouping syllables of a word into larger units called metrical feet which then determine the placement of stress within the word. Generative theorizing is based on the idea that there is a language-specific grammar that compactly captures this underlying structure.

An important consideration for metrical stress grammars is that the complete grammar of a language captures both the productive patterns of the language (which generalize to novel words) as well as the non-productive patterns (which don't). For English, an example of a productive pattern is that syllable weight matters for stress placement; an example of a non-productive pattern is whether a given suffix alters primary stress (e.g., $-ity$ does while $-ness$ doesn't: *prodúctive* becomes *productívity* vs. *prodúctiveness*). The productive patterns are what the generative grammars in metrical stress representations are typically meant to capture. I'll abbreviate "***prod***uctive metrical stress ***k***nowledge ***r***epresentations" below as **prodKRs**.

So, learning metrical stress is already tricky because there are non-productive patterns that must be ignored by children trying to identify the productive grammar for the language. English, however, is especially tricky in this regard. It isn't necessarily obvious which patterns are the productive ones. Core patterns of monomorphemic English words include:

1. Stress must occur on at least one of the last three syllables (Hammond 1999).

2. Syllable weight impacts stress placement (Chomsky and Halle 1968; Halle and Kenstowicz 1991; Hammond 1999; Pater 2000), and heavy syllables (containing a tense vowel like /*i*/ or closed by at least one consonant like /*ɛn*/) often bear stress. This property is typically referred to as *quantity sensitivity*.

3. The stress pattern of nouns is different from that of verbs and adjectives (Chomsky and Halle 1968; Hayes 1982; Kelly 1988; Kelly and Bock 1988; Hammond 1999; Cassidy and Kelly 2001). For instance, there are examples like *cónduct/condúct* and *désert/desért*, where the syntactic category influences the stress pattern for the syllable sequence (i.e., the noun versions are stress-initial while the verb versions are stress-final).

For words containing more than one morpheme, Hammond (1999) notes that there is a class of affixes "outside the domain of stress assignment" (e.g., *-able, -ed, -ing, -s, -or, -er, -ly, -able, -ment, -ness*) that allow modifications to the patterns above. More generally, there are known interactions with inflectional and derivational morphology (Chomsky and Halle 1968; Kiparsky 1979; Hayes 1982, 1995). For example, in *prétty/préttier/préttiest* and *sensátion/sensátional/sensátionally*, adding inflectional and derivational morphology doesn't shift the primary stress, despite adding syllables to the word.

In terms of representation, the goal of prodKR theories has been to define an English productive grammar that compactly captures as much of this variation as possible, leaving non-productive aspects to be encoded some other way. For acquisition, this means that the child equipped with knowledge of the prodKR should be able to learn the productive aspects of English stress even with the non-productive aspects present in the input. This likely requires children to identify and filter out the non-productive aspects.

2.2 Empirical grounding: Trajectory and input

To empirically ground the acquisition evaluation, we need to know something about when children develop knowledge of the productive aspects of English metrical stress. Experimental data suggest that acquisition happens in stages. At age two, English children use a metrical template that operates over syllables (Echols 1993) and which has the leftmost syllable stressed (Gerken 1994, 1996). This is useful for capturing the stress pattern of words like *cáptain*, *húngry*, *fíftieth*, *láter*, *ópening*, *zébra*, *ángel*, *grátefully*, and *fábulous*, which are found in the child-directed speech corpus I describe below. By age three, children have recognized that the metrical system is quantity sensitive, though they don't recognize the full set of factors that determines how syllable weight impacts stress placement (Kehoe 1998). By age four or five, there is suggestive evidence that English children have identified the target English productive grammar: (i) Arciuli et al. (2010) find that children as young as five override orthographic cues to alternative stress patterns that violate the English productive grammar, and (ii) Pettinato and Verhoeven (2008) find that children as young as four are at ceiling for repeating nonsense words that obey the English productive grammar but not for words that violate it.

These experimental findings provide helpful guideposts for a computational acquisition evaluation. In particular, in terms of the data children are learning from, we probably want to restrict our analyses to child-directed speech encountered before the age of four or five. If we're also interested in the earliest stages of English stress acquisition, we may want to restrict our analyses to data encountered before the age of two.

Given this, the data used for the acquisition analyses below comes from the Brent corpus (Brent and Siskind 2001), which is part of the American English subsection of the CHILDES database (MacWhinney 2000). This corpus contains speech directed at children between the

ages of six and twelve months (4780 multisyllabic word types, 99968 multisyllabic word tokens). Because all the stress representations we consider operate over syllables, the 4780 multisyllabic word types were syllabified using CELEX 2 (Baayen et al. 1995), the CALLHOME English Lexicon (Kingsbury et al. 1997), and the MRC Psycholinguistic Database (Wilson 1988). Target stress patterns for these word types were estimated using the CALLHOME English Lexicon (Kingsbury et al. 1997).[1] For words not in these resources (typically nonsense words like *fussies* or child-register words like *fishies*), native speaker intuitions and comparisons to similar sounding words were used when possible.

2.3 Representation overview

As I mentioned above, the three metrical stress representation theories have several points of agreement when it comes to deciding whether a given syllable in a word is stressed. All three assume a word is divided into syllables and those syllables are classified according to their syllable rimes only (e.g., *strong* (/stɹɑŋ/) is equivalent to /tɹɑŋ/, /ɹɑŋ/, and /ɑŋ/, which all have a rime consisting of a **V**owel followed by a **C**onsonant: VC). Grammars then form metrical feet made up of one or more syllables, indicated with parentheses in (1). Stress is determined by the metrical feet, with at most a single syllable within a metrical foot being stressed. A grammar defined by a prodKR will be associated with an underlying metrical structure, as shown in (1), whose observable form is the stress contour for the word.

(1) Sample metrical structure for *octopus* (/ɑktəpʊs/)

stress	1	0	0
metrical feet	(VC	V)	VC
syllable rimes	VC	V	VC
syllables	ɑk	tə	pʊs

Below I'll briefly highlight the relevant aspects of the three prodKR theories being compared, though interested readers should refer to Pearl et al. (2014, 2015) for a more complete description of each. These representational theories differ in several respects, including how they classify syllables, how metrical feet are constructed, and how stress interacts with metrical feet. Perhaps more fundamentally, they represent two different types of prodKR theories: a parametric type whose grammar variables are defined by parameters with values that must be set and a constraint-ranking type whose grammar variables are defined by violable constraints that must be ranked with respect to each other. Parametric representations generate the underlying structure from the grammar while constraint-ranking representations use the grammar to select an underlying structure from the available options.

2.3.1 The HV parametric representation

The first parametric prodKR is adapted from Halle and Vergnaud (1987) (**HV**), and its learnability has been previously investigated by Pearl (2007, 2009, 2011). The HV representation involves five main parameters with three sub-parameters, yielding 156 grammars in

[1] The syllabified and stressed annotations for the word types are available at *http://www.socsci.uci.edu/~lpearl/CoLaLab/uci-brent-syl-structure-Jul2014.xlsx*.

the hypothesis space. They affect how syllables are classified (*quantity sensitivity*), how metrical feet are constructed (*foot directionality*, *extrametricality*, *boundedness*), and how stress interacts with metrical feet (*foot headedness*). A grammar is then a set of these parameter and sub-parameter values.

The English productive grammar proposed for the HV representation differentiates syllables into Heavy and Light, treating syllables that end in VC (e.g., /*ɛn*/) as Heavy. The rightmost syllable of a word is extrametrical (indicated with ⟨...⟩ in (2)), and so not contained in a metrical foot. Metrical feet are built from the right edge of the word, a metrical foot spans two syllables when it can, and the leftmost syllable within a foot is stressed. Sample analyses using the English HV grammar are shown for *octopus* and *today* in (2). The generated stress contour matches the observed stress contour for *óctopus* (100=100) but not for *todáy* (10≠01).

(2) English HV grammar analyses for *octopus* (/ɑktəpʊs/) and *today* (/tədeː/):

stress	1	0	0	1	0
analysis	(Ĥ	L)	⟨H⟩	(Ĺ)	⟨H⟩
syllables	ɑk	tə	pʊs	tə	deː

2.3.2 The Hayes parametric representation

The second parametric system is adapted from Hayes (1995) (**Hayes**), and includes eight parameters that concern the basic distinction between stressed and unstressed syllables. These eight parameters yield 768 grammars in the hypothesis space. They affect how syllables are classified (*syllable weight*), how metrical feet are constructed (*extrametricality*, *foot directionality*, *foot inventory*, *parsing locality*, *stress analysis direction*), and how stress interacts with metrical feet (*degenerate feet*, *word layer end rule*).

All Hayes productive grammars differentiate syllables into Heavy and Light, and the proposed English grammar treats VC syllables as Heavy. It also views the rightmost consonant as extrametrical. Metrical feet are built from the right edge before word-level stress is assigned and as adjacently as possible. Each foot is two moras in size (Light syllable = one mora, Heavy syllable = two moras), and degenerate feet that deviate from the specified foot size are not allowed. Within a foot, the leftmost syllable is stressed. Sample analyses using the English grammar are shown for *octopus* and *today* in (3). The generated stress contour matches the observed stress contour for *todáy* (01=01) but not for *óctopus* (110≠100).

(3) English Hayes grammar analyses for *octopus* (/ɑktəpʊs/) and *today* (/tədeː/):

stress	1	1	0	0	1
analysis	(Ĥ)	(Ĺ	L)	L	(Ĥ)
syllables	ɑk	tə	pʊ⟨s⟩	tə	deː

2.3.3 The OT constraint-ranking representation

Optimality Theory (**OT**) (Tesar and Smolensky 2000; Prince and Smolensky 2002) characterizes linguistic knowledge as a universal set of constraints whose interaction determines the

form of observable linguistic data. A grammar is a ranking of these constraints. In general, violating higher-ranked constraints is worse than violating lower-ranked constraints.

Given n constraints, there are $n!$ possible rankings. In the instantiation of OT briefly reviewed below which is derived from Hammond (1999) and Pater (2000), there are nine metrical stress constraints, defining a hypothesis space of 9! = 362,880 grammars. These constraints affect how syllables are classified (*WeightToStress-VV*, *WeightToStress-VC*, *NoSonorantNucleus*), which metrical feet are preferred and where (*NonFinality*, *Parse-σ*, *FootBinarity*, *AlignLeft*, *AlignRight*), and how stress interacts with metrical feet (*Trochaic*). Additionally, there is one inviolable principle called *Rooting*, which requires all words to have some stress on them. Since a stressed syllable must be in a metrical foot, only candidate analyses that have at least one metrical foot are available to children.

Notably, the OT "grammar" for a language is often a partial ordering of constraints, and so corresponds to multiple grammars that are explicit rankings of all nine constraints.[2] In this vein, the English grammar derived from Hammond (1999) and Pater (2000) obeys ten constraint ranking relationships, which correspond to 26 grammars that explicitly rank all nine constraints. This partial ordering is shown in Figure 1, where each arrow represents a constraint ordering that is true of the proposed English grammar.

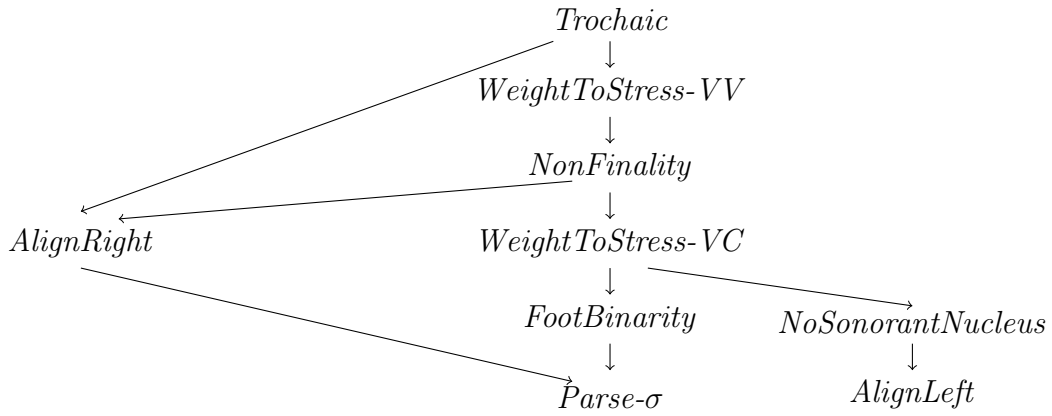

Figure 1: Partial ordering of constraints defining the OT English grammar.

The tableau in Figure 2 is an evaluation of *little* (/lɪɾɪl/) using a grammar satisfying the English constraint rankings. For this word form, the optimal candidate for the grammar has a stress contour that matches the observed stress contour of *little* (*líttle*=10). In contrast, the tableau in Figure 3 is an evaluation of *kitty* (/kɪɾi/), and shows that the same grammar selects a candidate whose stress contour does not match the observed stress contour (*kittý*=01≠10=*kítty*).

[2]In particular, OT theorists often reserve the term "grammar" for a mapping from underlying representations to surface representations (e.g., underlying /lɪ ɾɪl/ to surface /(lí) ɾl̩/). This can easily correspond to multiple explicit rankings of the constraints available in the prodKR, as it does here for English. This contrasts with how I use the term "grammar" here, where I mean a single combination of the variable values in the prodKR – in particular, a single explicit ranking of the constraints. I will continue with this latter use of grammar in order to be fair to the HV and Hayes prodKRs, whose grammars are also defined by the distinct combinations of the variable values available (which are parameter values for those prodKRs).

Input: /lɪ ɾɪl/	Tro	Wts-vv	NonFin	Wts-vc	Align-R	FtBin	Parse-σ	NoSonNuc	Align-L
(lí ɾɪl)			∗!	∗					
(lɪ ɾíl)	∗!		∗						
(lí) ɾɪl				∗!	∗	∗	∗		
lɪ (ɾíl)			∗!			∗	∗		∗
(lí)(ɾíl)			∗!		∗	∗∗			∗
(lí ɾl̩)			∗!					∗	
(lɪ ɾĺ̩)	∗!		∗					∗	
☞ (lí) ɾl̩					∗	∗	∗	∗	
lɪ (ɾĺ̩)			∗!			∗	∗	∗	∗
(lí)(ɾĺ̩)			∗!		∗	∗∗		∗	∗

Figure 2: Evaluation of *little* using a grammar that satisfies the English constraint rankings.

Input: /kɪ ɾi/	Tro	Wts-vv	NonFin	Wts-vc	Align-R	FtBin	Parse-σ	NoSonNuc	Align-L
(kí ɾi)		∗!	∗						
(kɪ ɾí)	∗!		∗						
(kí) ɾi		∗!			∗	∗	∗		
☞ kɪ (ɾí)			∗				∗		∗
(kí)(ɾí)			∗		∗!	∗			∗

Figure 3: Evaluation of *kitty* using a grammar that satisfies the English constraint rankings.

2.4 Computational acquisition evaluation: English

2.4.1 English non-productivity complications

Given this setup, the acquisition goal for English children is to identify the productive grammar from the prodKR that accounts for the productive word-level stress patterns in English data. However, remember that non-productive patterns abound in English and these can mask the productive grammar's expression.

One particular issue is that the non-productive patterns can lead to multiple stress contours for a single SYLLABIC WORD FORM: this is a word abstracted into syllable rimes comprised of vowels and consonants, such as *kitty* abstracted into a short vowel syllable followed by a long vowel syllable (V VV). Multiple word types may be instances of the same syllabic word form (e.g., *kitty*, *away*, and *uh oh* as instances of V VV). Notably, the issue all three representations have is at the syllabic word form level. This is because a productive grammar in these representations can only generate/select a single stress contour per syllabic word form. This means that the syllabic word form V VV is assigned a single stress contour per grammar, whether the word itself it *kítty*, *awáy*, or *úh óh*. Practically speaking, this

means there is no way for any single grammar to account for all the observed stress contours for English words. For example, a grammar the generates/selects the stress contour 10 for the syllabic word form V VV will match the contour for *kítty* but will, by necessity, mismatch the contours for *awáy* and *úh óh*.

So, how often does a syllabic word form have multiple stress contours associated with it in English child-directed speech? In the Brent corpus described in section 2.2, between 37% and 58% of the syllabic word forms (depending on the syllabic distinctions made by the prodKR[3]) have multiple stress contours associated with them. This highlights why filtering the input is likely helpful for identifying the productive grammar of English – some of these multiple stress contour instances can potentially be ignored.

	Total syllabic word forms	**Syllabic word forms with multiple stress contours**
HV	186	95 (51%)
Hayes	149	86 (58%)
OT	452	166 (37%)

Table 1: Syllabic word forms in the English child-directed speech sample with multiple stress contours for each prodKR.

2.4.2 Learnability metrics

The basic question is something like "How easily does this representation allow children to learn their specific language's productive grammar, when given the data that children from that language typically encounter?" Learnability analysis is one way of formalizing the answer to this question, and the approach I present below is similar to those taken by Pearl (2011) and Legate and Yang (2013). It assesses learnability (i) from child-directed speech input (the data we think children typically encounter), and (ii) at the computational level (in the sense of Marr 1982). A computational-level analysis focuses on the choices a rational learner would make, given the hypothesis space defined by the prodKR. One of the benefits of a computational-level analysis is that it can highlight if learnability issues exist simply because of the interplay between the hypothesis space and children's data, irrespective of any cognitive processing limitations children may have.

But how do we define a rational learner? A rational learner is one that chooses what it perceives to be the best grammar. So, what does "best" mean? Pearl et al. (2014, 2015) suggest it's the grammar that's able to account for the most data in children's acquisitional intake.[4] This relates to the utility of productive grammars: a productive grammar is useful

[3]The HV prodKR allows 3 syllabic distinctions, the Hayes prodKR allows 4, and the OT prodKR allows 8. See Pearl et al. (2014, 2015) for a more detailed discussion of these distinctions.

[4]I note that this is in the same spirit as the classical principle of Empirical Risk Minimization (ERM) in statistical learning theory (Vapnik 1992, 2013), where the learner picks a hypothesis that minimizes error on the training data. In this case, it would mean children should pick a hypothesis that best accounts for the data available, i.e., the data in their acquisitional intake. Interestingly, this can lead to a problem of overfitting, where the learner attempts to match the data too closely and loses generalization capability. This is very much a problem children face, as they want a grammar that correctly accounts for data they

because it allows the learner to compactly represent the productive aspects of the language data. This means that language data captured by the productive grammar do not need to be stored in detail. Instead, the productive aspects of these data can be generated by the compact representation provided by the grammar. So, the more data accounted for by the productive grammar, the more useful the grammar is because there are fewer data that must be dealt with separately (e.g., stored explicitly). Because of this, from a language use standpoint, the best productive grammar is naturally defined as the one that can account for the most data. Once we define the data in children's acquisitional intake, we can then simply examine which grammar in the hypothesis space defined by the prodKR is able to account for the most.

At an individual data point level, a grammar can either be compatible or incompatible with the data point. For example, a metrical stress grammar is compatible with a data point if it can generate/select the observed stress contour for that data point. The proportion of data points a grammar is compatible with is its RAW COMPATIBILITY with that data set (e.g., a grammar compatible with 70% of the data set has a raw compatibility of 0.70). When comparing productive grammars within a prodKR, a higher raw compatibility is better since this indicates the grammar is more useful at accounting for the available data.

However, for acquisition, what may matter more than raw compatibility is how a productive grammar compares to other grammars defined by the prodKR. This is captured by RELATIVE COMPATIBILITY, which is how a grammar's raw compatibility compares to the raw compatibilities of other grammars in the hypothesis space. This can be defined as the proportion of grammars in the hypothesis space that this grammar is better than, with respect to raw compatibility. The best grammar will be better than all other grammars, and so its relative compatibility approaches 1 as the number of grammars in the hypothesis space increases. For example, if there are 768 grammars, the best grammar is better than 767, which gives a relative compatibility of $767/768 = 0.999$. Importantly, no matter what the raw compatibility of the best grammar is, it's the one a rational learner would choose because it's the best of all the grammars defined by the prodKR.

While the previous two metrics focused on evaluating the learnability of specific grammars within a prodKR, we can also evaluate the prodKR itself. In particular, we can calculate a prodKR's LEARNABILITY POTENTIAL, which is simply the raw compatibility of the best grammar defined by the prodKR. For example, if the best grammar in a prodKR (with relative compatibility closest to 1) has a raw compatibility of 0.70, then we can say that prodKR has a learnability potential of 0.70. In effect, this metric indicates the utility of the prodKR, as implemented by the best grammar it defines. The basic idea is that the learnability potential measures how well the productive grammar variables defined by the prodKR account for the data in children's acquisitional intake. So, this is a more direct way to compare prodKRs, irrespective of the English-specific grammars within them.

However, we might expect that the proposed English grammar in each prodKR be the grammar that's learned most easily from English children's acquisitional intake. This can be empirically tested using the relative compatibility metric. More specifically, if the proposed English grammar is the one most easily learned within a prodKR, it should have the highest

have yet to encounter – i.e., that generalizes appropriately. Many thanks to an anonymous reviewer for bringing my attention to this.

raw compatibility with the English acquisitional intake, which will then lead to the relative compatibility closest to 1. At the prodKR level, the English grammar's raw compatibility should be equivalent to the prodKR's learnability potential because it would be the grammar within the prodKR that's best at accounting for the English acquisitional intake.

2.4.3 Evaluation for English

For each of the learnability analyses below, an algorithm was run that evaluated the compatibility of each grammar in a prodKR against the data in English children's acquisitional intake. From this, raw compatibility and relative compatibility scores could be calculated for each grammar, as well as the learnability potential for each prodKR.

I'll discuss two types of analyses. The first type represents the initial stages of acquisition, when children are not yet aware there are non-productive aspects in English stress. So, this kind of child (naively) assumes all the data are generated by the productive grammar. The second analysis type assumes children have a mechanism for filtering out data they recognize as non-productive.[5] So, this kind of (more sophisticated) child recognizes that not all data may be capturable by a productive grammar. Pearl et al. (2014, 2015) implemented this mechanism using a proposal from Yang (2005) and Legate and Yang (2013) called the Tolerance Principle that I'll describe below. Table 2 shows the results of both these analyses.

The basic idea of the Tolerance Principle is that it's a way to estimate how many exceptions a rule can tolerate before it's no longer useful to represent the rule at all. In essence, if there are too many exceptions, it's better to simply deal with the exceptions on an individual basis rather than bothering to learn a rule that's often violated. Using processing considerations and related mathematical formalizations, Yang (2005) proposed that, for N items, the total exceptions a rule can tolerate is equivalent to $\frac{N}{ln\ N}$. If there are more exceptions than this, then the rule is not productive.

One way children might use this principle when learning metrical stress is to deal with syllabic words forms that have multiple stress contours. When a syllabic word form like this is encountered (e.g., V VV: *kítty*, *awáy*, *úh óh*), one stress contour may be the productive stress contour while the others can be viewed as exceptions (which can be safely ignored for purposes of learning the productive grammar). At any point during acquisition, there are two possible outcomes. One option is that one contour is the productive contour according to the Tolerance Principle, and so the learner would attempt to account for only the data with that stress contour (e.g., *kítty*), ignoring the other data for that syllabic word form (e.g., *awáy*, *úh óh*) when trying to learn the productive grammar. The other option is that no contour is productive according to the Tolerance Principle, and so all the data for that syllabic word form are ignored when trying to learn the grammar. The upshot is that a productive subset of the available input is now the child's acquisitional intake. In particular, Table 2 shows what happens when children's intake is limited to the productive subset (+prod filter), as defined by the Tolerance Principle, and each of the learnability metrics is calculated over that productive subset.

[5]This can be helpful for preventing overfitting, which happens when the learner attempts to account for data that actually aren't relevant for generalization. That is, these data are in the input, but they're anomalies in the sense that they aren't generated by the productive system for the language. So, these data can lead the child astray when identifying the correct productive system for the language.

Table 2: Learnability analyses for the three prodKRs: HV, Hayes, and OT. The three metrics shown are learnability potential of the prodKR (prodKR:Pot), raw compatibility of the (best) English grammar (Eng:Raw), and relative compatibility of the (best) English grammar (Eng:Rel), which are computed over word types in English child-directed speech. Results are shown for learners not using a productive filter (-prod filter) and learners using a productive filter implemented via the Tolerance Principle (+prod filter).

		prodKR:Pot	**Eng:Raw**	**Eng:Rel**
-prod filter	**HV**	0.67	0.59	0.67
	Hayes	0.68	0.49	0.68
	OT	0.65	0.57	0.82
+prod filter	**HV**	0.95	0.87	0.62
	Hayes	0.93	0.70	0.68
	OT	0.84	0.63	0.80

One of the first things we can observe from Table 2 is that this significantly impacts the learnability potential of the three prodKRs, as might be expected when variation is removed from the intake.[6] For a learner without a productive filter, all three prodKRs have a grammar that can account for around $\frac{2}{3}$ of the acquisitional intake (0.65-0.68). While this is pretty useful data coverage, there's a big advantage to recognizing that some data are unproductive. In particular, for a learner using a productive data filter like the Tolerance Principle, the prodKRs have a grammar that can account for 84% (OT), 93% (Hayes), or 95% (HV) of the acquisitional intake. So, perhaps the parametric prodKRs (HV, Hayes) have an advantage in terms of productive data coverage in English children's input, though all three prodKRs are doing a very good job of capturing these data.

However, something interesting is going on when we look at the compatibility of the proposed English grammars in each prodKR. In every single case, the compatibility of the proposed English grammar (or the best one, in the case of OT) is *below* the learnability potential of the prodKR (i.e., Eng:Raw is always less than prodKR:Pot in Table 2). This means the proposed English grammars are *not* the grammars within the prodKRs able to account for the most data in the acquisitional intake, whether filtered down to the productive subset or not. How much are they not? That's what the relative compatibility scores indicate. For naive learners not using a productive data filter, $\frac{1}{3}$ of the grammars in the HV and Hayes prodKRs can account for more data (HV: 1- 0.67=0.33; Hayes: 1-0.68=0.32), which works out to tens (HV: 51) or hundreds (Hayes: 249) of grammars in the hypothesis space. In the OT prodKR, about $\frac{1}{5}$ of the hypothesis space is better (1-0.82 = 0.18), which works out to tens of thousands of grammars (OT: 66,407) being able to account for more data than the best proposed English grammar.

This doesn't change for children who use a productive data filter. While the overall

[6]Interestingly, just because variation is removed doesn't mean the productive data are completely captured by any single grammar. This can be seen by the learnability potential still being less than 1 in Table 2. This is because "productive" according to the Tolerance Principle is simply about relative frequency and doesn't necessarily accord with the variables a prodKR uses to define its grammars. See Pearl et al. (2015) for an explicit demonstration of this point.

data coverage is higher for the proposed English grammars, the (best) English grammar coverage is still less than the learnability potential of the prodKR. How bad is it? About the same as before: tens of grammars (HV: 59), hundreds of grammars (Hayes: 246), or tens of thousands of grammars (OT: 73,302) in the prodKR hypothesis space are able to account for more English acquisitional intake data.

How do we interpret this? One idea is that there are additional filters English children employ to winnow down the acquisitional intake still further, and the proposed English grammars are best able to account for that further-filtered acquisitional intake. Pearl et al. (2014, 2015) explored a few cognitively plausible filtering options of this kind, though they were unable to find any that caused the proposed English grammars to improve their relative compatibility.

Another idea is that we should consider alternatives for what the productive English grammar actually is that English children acquire from their input. We can look to the grammars within each prodKR that are more compatible than the currently proposed English grammars and see how these high compatibility grammars differ. What values do the high compatibility parametric grammars use? What constraint rankings do the high compatibility constraint-ranking grammars use? Results are summarized in Table 3. Let's consider each prodKR in turn.

Table 3: Learnability analyses for the three prodKRs: HV, Hayes, and OT. The three metrics shown are learnability potential of the prodKR (prodKR:Pot), raw compatibility of the (best) alternative English grammar (Eng':Raw), and relative compatibility of the (best) alternative English grammar (Eng':Rel), which are computed over word types in English child-directed speech. Results are shown for learners not using a productive filter (-prod filter) and learners using a productive filter implemented via the Tolerance Principle (+prod filter).

		prodKR:Pot	**Eng':Raw**	**Eng':Rel**
-prod filter	**HV**	0.67	0.64	0.94
	Hayes	0.68	0.64	0.91
	OT	0.65	0.65	0.99
+prod filter	**HV**	0.95	0.88	0.71
	Hayes	0.93	0.93	0.96
	OT	0.84	0.84	0.99

For the HV grammar, it turns out that many high compatibility grammars use a different quantity sensitivity value (preferring syllables not to be differentiated by weight). This allows these grammars to handle words like like *béllybútton* and *sátisfied*, which have unstressed heavy syllables that aren't at the edge of the word (*ly* in *bellybutton*, *tis* in *satisfied*). If we simply swap this value in the current English grammar definition, we get a very similar grammar – let's call it English'. For a naive learner not using a productive filter, English' is able to account for nearly as much data as the best grammar (Potential: 0.67, English': 0.64), and is better than 94% of the productive grammars in the HV hypothesis space, as indicated by the relative compatibility score (0.94). This is a significant improvement, and

makes it much easier for a child to select this grammar given the English child-directed speech data – that is, this English' grammar is more easily learnable from these data, given the hypothesis space of grammars. However, a more sophisticated learner using a productive data filter doesn't fare so well with English' – the data coverage difference is larger (Potential: 0.95, English': 0.88), but more importantly, 29% of the productive grammars available can account for more data (Relative: 1-0.71 = 0.29 = 45 grammars in the HV hypothesis space). This suggests that the English' alternative is only more easily learnable than the original proposed English grammar in the earlier stages of acquisition when children think all data are productive.

For the Hayes prodKR, it turns out that many high compatibility grammars use a different metrical foot value than the current definition of the English grammar: they use syllabic trochees rather than moraic trochees. This allows these grammars to account for bisyllabic words with an unstressed heavy syllable at the edge, such as *báby* and *kítty*, as well as trisyllabic compound words with unstressed syllables in the middle and heavy syllables at the edge, such as *sléepyhéad*. If we alter the current English grammar to use syllabic trochees, this English' alternative grammar is much easier for a naive learner to learn. It accounts for nearly as much data as the best grammar available (Potential: 0.68, English': 0.64) and is better than 91% of the grammars in the Hayes hypothesis space. A learner using a productive filter fares even better when trying to learn the English' grammar: this grammar accounts for effectively the same amount of data as the best grammar (Potential: 0.93, English': 0.93) and is better than 96% of the grammars in the Hayes hypothesis space. This suggests that the English' alternative is more easily learnable than the original proposed English grammar in both earlier and later stages of English metrical stress acquisition.

For the OT prodKR, it turns out that all the highest compatibility grammars use a constraint ranking that the current English grammar definition doesn't: ranking NonFinality higher than WeightToStress-VV. This means it's more important to exclude the rightmost foot from getting stress (NonFinality) than it is to stress long vowel syllables (WeightToStress-VV). This allows these grammars to account for words like *báby* and *kítty*, which have an unstressed long vowel syllable as the rightmost syllable. If we swap this ranking in the current English grammar definition to create English', we find that learning becomes much easier for both naive and more sophisticated learners. For both stages of acquisition, English' can account for effectively the same amount of data in the acquisitional intake as the best grammar in the OT prodKR (without a productive filter: Potential = 0.65, English' = 0.65; with a productive filter: Potential = 0.84, English' = 0.84). This leads to English' being better than 99% of the grammars in the OT hypothesis space – much more easily learnable! So, this update seems to be incredibly helpful from an acquisition perspective, no matter what stage of acquisition the English child is in.

2.5 Metrical stress representations: Summary

So what have we learned? First, we have a concrete demonstration that all three prodKRs are useful for English acquisition, in the sense that they can define grammars that account for a large portion of English child-directed speech data. This is particularly true if children are using a productive data filter – and in that case, we may even be able to rank some prodKRs higher than others (HV and Hayes over OT) based on the amount of data that can be

captured by the best grammar in the prodKR. More generally, for theories of metrical stress acquisition, this highlights the usefulness of a mechanism for filtering out non-productive English data.

Second, we also know something about the current English grammar definitions in each prodKR – *none* of them are the grammars most easily learnable from the data English children typically encounter. Instead, alternative grammars differing by a single parameter value or constraint ranking are usually far easier to learn. I think we can consider these empirically motivated alternative proposals for the English grammar in each prodKR that should be given more theoretical, experimental, and computational consideration. For example, do we see evidence of these alternative English grammars in children? Do we see evidence in adults? Are these grammars more compatible with the data an English adult typically encounters? More generally, for theories of representation, we can use computational learnability results like the ones we have here to identify promising alternative language-specific representations.

So, to sum up, we've seen how to use this approach to evaluate metrical stress representations by using them for acquisition. This allows us to refine our theories of both the metrical stress knowledge representation and the acquisition process that accompanies that representation.

3 Case study: Syntactic islands

3.1 About syntactic islands

Let's turn now to syntactic islands. Pearl and Sprouse (2013a,b) chose to focus on them because they're central to acquisition theories that rely on Universal Grammar (UG). In particular, knowledge of syntactic islands is often difficult to characterize without referring to relatively abstract syntactic structure, which has led many syntacticians to propose relatively complex, abstract constraints to capture island effects in adult grammars (e.g., Chomsky 1986). This has led many acquisition researchers to hypothesize domain-specific, innate knowledge (i.e., UG knowledge) to explain how children learn about syntactic islands. So, this is a non-trivial acquisition problem that has implications for long-standing debates about the role of domain-specific, innate knowledge in language acquisition (e.g., Ambridge et al. 2014; Pearl 2014).

Now, what *are* syntactic islands? They have to do with long-distance dependencies, like those in (4) between the *wh*-word and where it's understood (indicated by ___). One defining characteristic of long-distance *wh*-dependencies is that they appear to be unconstrained by length (Chomsky 1965; Ross 1967), as shown in (4b–4c).

(4) a. What does Jack think ___?
    b. What does Jack think that Lily said ___?
    c. What does Jack think that Lily said that Sarah heard ___?

However, we know that if the gap position of a *wh*-dependency appears within certain syntactic structures, such as those in square brackets in (5), the resulting utterance is unacceptable (Chomsky 1965; Ross 1967; Chomsky 1973; Huang 1982, among others). These

structures have been called *syntactic islands* (Ross 1967). So, the unacceptability of these utterances can be explained by the long-distance dependency in the utterance crossing a syntactic island, which isn't allowed.

(5) Some examples of island-crossing dependencies, with island structures in brackets
   a. * What did you make [the claim that Jack bought ___]?
   b. * What do you think [the joke about ___] was hilarious?
   c. * What do you wonder [whether Jack bought ___]?
   d. * What do you worry [if Jack buys ___]?

3.2 The acquisition target

How do we tell if someone has knowledge of syntactic islands? One way is to see the effect of that knowledge, via judgments about how acceptable different utterances are. Sprouse et al. (2012) did this by collecting formal acceptability judgments about four different islands (like the ones above in (5)): Complex NP, Subject, Whether, and Adjunct islands. For acquisition modeling purposes, these acceptability judgments provide a concrete set of behaviors that a modeled learner should aim to reproduce. That is, a successful learner will generate similar acceptability judgments.

Sprouse et al. (2012)'s design used a factorial definition that controlled for two salient properties of island-crossing dependencies: (1) the length of the dependency (matrix (MAT) vs. embedded (EMB)), and (2) the presence of an island structure, whether or not the dependency actually crosses it (non-island (NON) vs. island (ISL)). This led to stimuli like those in (6)-(9), which have island structures bracketed.

(6) Complex NP islands
   a. Who ___ claimed that Lily forgot the necklace? MAT | NON
   b. What did the teacher claim that Lily forgot ___? EMB | NON
   c. Who ___ made [the claim that Lily forgot the necklace]? MAT | ISL
   d. * What did the teacher make [the claim that Lily forgot ___]? EMB | ISL

(7) Subject islands
   a. Who ___ thinks the necklace is expensive? MAT | NON
   b. What does Jack think ___ is expensive? EMB | NON
   c. Who ___ thinks [the necklace for Lily] is expensive? MAT | ISL
   d. * Who does Jack think [the necklace for ___] is expensive? EMB | ISL

(8) Whether islands
   a. Who ___ thinks that Jack stole the necklace? MAT | NON
   b. What does the teacher think that Jack stole ___? EMB | NON
   c. Who ___ wonders [whether Jack stole the necklace]? MAT | ISL
   d. * What does the teacher wonder [whether Jack stole ___]? EMB | ISL

(9) Adjunct islands

a.	Who ___ thinks that Lily forgot the necklace?	MAT | NON
b.	What does the teacher think that Lily forgot ___?	EMB | NON
c.	Who ___ worries [if Lily forgot the necklace]?	MAT | ISL
d.	* What does the teacher worry [if Lily forgot ___] ?	EMB | ISL

A syntactic island is apparent when there's a superadditive effect of the two factors. In particular, it's the additional unacceptability that arises beyond the dependency being an embedded clause dependency and beyond the island structure being present in the utterance. The presence of a syntactic island effect then becomes visually salient: If the acceptability of the four stimuli types for each island (as indicated by their z-scores) is plotted on an interaction plot, the presence of a syntactic island appears as two non-parallel lines. This effect is also statistically significant.

In contrast, the lack of a syntactic island appears as two parallel lines and results in no significant statistical interaction, because the unacceptability of the utterance is completely explainable by the summed effects of it being an embedded clause dependency and having an island structure in it. Sprouse et al. (2012) found superadditivity (i.e., non-parallel lines and a statistically significant interaction) for all four islands investigated, as shown in Figure 4. So, this suggests that the knowledge that dependencies can't cross these island structures is part of adult knowledge of syntactic islands. This is then one kind of target behavior that a successful learner should produce: a superadditive interaction when given the same stimuli to judge.

3.3 Representations

3.3.1 Subjacency

In the Government and Binding framework of the 1980s, syntacticians proposed a constraint called the Subjacency Condition. This basically says that dependencies can't cross two or more *bounding nodes* (Chomsky 1973; Huang 1982; Lasnik and Saito 1984, among others). If a dependency crosses two or more bounding nodes, a syntactic island effect occurs. What counts as a bounding node varies cross-linguistically, though bounding nodes are always drawn from the set {NP, IP, CP}. So, when using this representation, children need to learn which of these are bounding nodes in their language, though they already know (via UG) about the restriction the Subjacency Condition imposes and the set of possible bounding nodes.

3.3.2 Subjacency-ish

Pearl and Sprouse (2013a,b, 2015) investigated a representation that shares the intuition with Subjacency that there's a local structural anomaly when syntactic islands occur. However, instead of characterizing this anomaly with bounding nodes, Pearl & Sprouse suggested that it could be described as a low probability region with respect to the phrase structure nodes that contain the dependency (which they called *container nodes*).

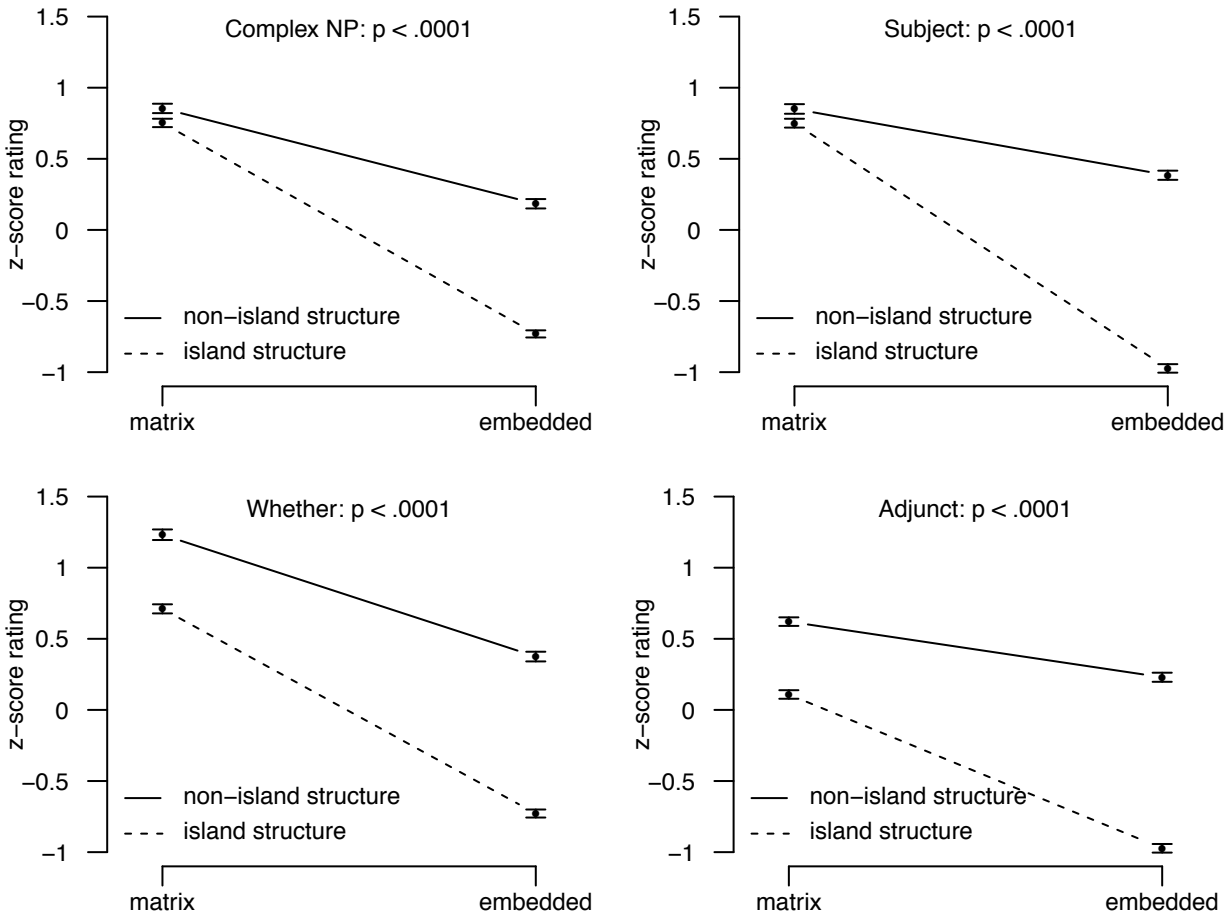

Figure 4: Experimentally derived acceptability judgments for the four island types from Sprouse et al. (2012).

A phrase structure node contains the dependency if the path from the gap to the *wh*-word must pass through the phrase structure node. As an example, consider the utterance *Who did Jack think that the story about penguins amused?*. Starting at the gap, the path must move up through the embedded VP, the embedded IP, the CP, the main VP, and the main IP before it finally reaches a phrase structure node that has the *wh*-word *who* as its child. These are the container nodes for this dependency, shown in (10).

(10) Who did [$_{IP}$ Jack [$_{VP}$ think [$_{CP}$ that [$_{IP}$ the story about penguins [$_{VP}$ amused ___]]]]]?

Children using the Subjacency-ish representation have to learn which are the low probability sequences of container nodes for their language, though they already know (possibly through UG) how to recognize container nodes for their language.

3.3.3 Subjacency vs. Subjacency-ish

What separates these representations is the amount of language-specific knowledge built in just for islands, as shown in Table 4. Subjacency requires children to know that dependen-

cies are defined over bounding nodes, which are drawn from a circumscribed set of phrase structure nodes. Children also have to know that dependencies crossing too many of these bounding nodes are bad. As far as I know, none of these knowledge components were intended to be useful beyond learning syntactic islands, so I think it's reasonable to classify them as "islands-only" knowledge.

This contrasts with the Subjacency-ish representation, which requires children to know that dependencies are defined over container nodes, which are drawn from the set of phrase structure nodes. While container nodes are going to be very helpful for learning syntactic islands (as I'll show below), there's evidence that they're also used when processing dependencies more generally (Crain and Fodor 1985; Frazier and Flores D'Arcais 1989; Phillips 2006). So, container nodes aren't obviously "islands-only" in quite the same way. The last piece the Subjacency-ish representation requires is that low probability sequences be dispreferred, which is something useful for learning all kinds of things, linguistic or otherwise. Given this, Pearl & Sprouse decided to test the Subjacency-ish representation to see if it could accomplish the same acquisition task that the Subjacency representation was meant to.

Table 4: Comparison of representations with respect to the amount of required prior knowledge that's specifically for learning syntactic islands.

Representation	**Required knowledge**	**Islands only?**
Subjacency	Dependencies defined over bounding nodes (BNs)	✓
	BNs ∈ {NP, IP, CP}	✓
	Crossing 2+ BNs is bad	✓
Subjacency-ish	Dependencies defined over container nodes (CNs)	?
	CNs ∈ phrase structure nodes	?
	Low probability CN sequences are dispreferred	

3.4 Subjacency-ish evaluation

To evaluate the Subjacency-ish representation using acquisition modeling, we first need to define the acquisition task. This means we need to know what the hypothesis space is, what children's acquisitional intake is, how the learning process works, how long the learning period is, and what the target knowledge and behavior are. I've discussed the target knowledge and its behavioral signature above in section 3.2, so now I'll turn to the other key pieces.

3.4.1 Hypothesis space

The hypothesis space for a child using this representation can be defined over dependencies, as characterized by the sequence of container nodes in those dependencies. Given this, the child's goal is to identify the set of container node sequences that are grammatical for the language's dependencies.

So how exactly are container nodes defined? A default hypothesis might be that container nodes correspond to "basic-level" phrase structure nodes like VP and CP. Before going any further, it's useful to check if this definition will allow children to distinguish between the grammatical and ungrammatical dependencies we have from Sprouse et al. (2012). Table 5 shows these stimuli characterized in terms of this kind of container node.

We can make a few observations. First, multiple stimuli are actually characterized by the same container node sequence. For example, all the matrix, non-island (MAT | NON) stimuli like *Who ___ claimed that Lily forgot the necklace?*, *Who ___ thinks that the necklace is expensive?*, *Who ___ thinks that Jack stole the necklace?*, and *Who ___ thinks that Lily forgot the necklace?* are matrix subject dependencies characterized by the container node sequence IP.

Second, if we use only the basic-level phrase structure nodes as container nodes, there are problems for the Whether and Adjunct stimuli. In particular, grammatical stimuli like *What does the teacher think that Jack stole ___?* are characterized by the same sequence as ungrammatical stimuli like *What does the teacher wonder whether Jack stole ___?*: IP-VP-CP-IP-VP. This means that a child using this definition of container nodes could not possibly generate different judgments for these stimuli. So, to reach the target knowledge – and target behavior – where there *is* a difference between these dependencies, a child has to have a different definition of container nodes. This highlights how considerations about the target of acquisition can cause us to refine our proposals about a knowledge representation.

Pearl & Sprouse proposed that a child could make a minor adjustment to the container node definition: for the CP phrase structure nodes only, container nodes would include subcategorization information about the lexical head. All other container nodes would remain unsubcategorized. This allows the child to distinguish between the dependencies mentioned before: *What does the teacher think that Jack stole ___?* = IP-VP-CP_{that}-IP-VP, while *What does the teacher wonder whether Jack stole ___?* = IP-VP-$CP_{whether}$-IP-VP. Table 5 shows that this updated definition causes all grammatical dependencies to have different container node representations than the ungrammatical dependencies. So, in principle, it is now possible to have different judgments about them using this representation.

3.4.2 Acquisitional intake

Children's acquisitional intake is based on the representation they're using. Here, that means that any *wh*-dependency in the input can be characterized by a container node sequence and so that *wh*-dependency becomes relevant information about the *wh*-dependencies allowed in the language. So, what *wh*-dependencies do we expect to find in English children's input?

Pearl & Sprouse estimated this from American English child-directed speech from the CHILDES database (MacWhinney 2000). In particular, they used a sample of 101,838 child-directed utterances aggregated from several commonly used American English corpora: the Adam and Eve corpora by Brown (1973), the Valian (1991) corpus, and the Suppes (1974) corpus. Collectively, these utterances were directed at 24 children between the ages of 1;6 and 5;2.

It turns out that most (89.5%) of the *wh*-dependencies are of two types: matrix-object dependencies like *What did she see?* (container node sequence: IP-VP = 76.7%) and matrix-subject dependencies like *Who saw it?* (container node sequence: IP = 12.8%). The re-

Table 5: Stimuli from Sprouse et al. (2012) characterized using different definitions of container nodes (basic-level vs. subcategorized CP). Grammatical and ungrammatical dependencies for each stimuli type are shown. Grammatical and ungrammatical stimuli characterized by the same container node sequence are ***italicized and bolded***.

	Grammatical	Ungrammatical
	Basic-level	
Complex NP	IP IP-VP-CP-IP-VP	IP-VP-NP-CP-IP-VP
Subject	IP IP-VP-CP-IP	IP-VP-CP-IP-NP-PP
Whether Adjunct	IP ***IP-VP-CP-IP-VP***	***IP-VP-CP-IP-VP***
	Subcategorized CP	
Complex NP	IP IP-VP-CP$_{that}$-IP-VP	IP-VP-NP-CP$_{that}$-IP-VP
Subject	IP IP-VP-CP$_{null}$-IP	IP-VP-CP$_{null}$-IP-NP-PP
Whether	IP IP-VP-CP$_{whether}$-IP-VP	IP-VP-CP$_{that}$-IP-VP
Adjunct	IP IP-VP-CP$_{if}$-IP-VP	IP-VP-CP$_{that}$-IP-VP

maining 10.5% of the *wh*-dependencies included 24 different dependency types. So, there are quite a variety of *wh*-dependencies available in the acquisitional intake, though some appear far more than others.

3.4.3 Learning process

When we talk about the learning process, what we really need for acquisition evaluation purposes is a step-by-step procedure that children could use to update their internal knowledge representations. Pearl & Sprouse proposed one that extracts local information about every *wh*-dependency encountered, by breaking each dependency into a set of container node trigrams (11c).[7] So, each trigram represents a local chunk of the dependency.

(11) Who did Jack think that the story about penguins amused ___?

a. Phrase structure nodes containing the *wh*-dependency:
Who did [$_{IP}$ Jack [$_{VP}$ think [$_{CP}$ that [$_{IP}$ the story about penguins [$_{VP}$ amused ___]]]]?

b. Container node characterization of *wh*-dependency with CP subcategorization:
IP-VP-CP$_{that}$-IP-VP

[7]Note that trigram encodings typically represent the beginning and ending of sequences with special symbols (*start* and *end* here), since this is relevant information.

c. Trigrams of container nodes $\in Trigrams_{IP-VP-CP_{that}-IP-VP}$:
= *start*-IP-VP
IP-VP-CP$_{that}$
VP-CP$_{that}$-IP
CP$_{that}$-IP-VP
IP-VP-*end*

During learning, children track the frequency of the container node trigrams. This means that a single dependency can provide information about more than one trigram – for example, the dependency in (11) provides one instance of five different trigrams. After encountering many *wh*-dependencies, the child has a collection of frequencies for each of the container node trigrams observed. These can be normalized so the child has a sense, for any trigram, how relatively frequent it is. Conveniently, that's all the child needs to learn.

After this relative frequency information is internalized, children can then assign any *wh*-dependency (even one they've never seen before) a probability, based on the probabilities of the trigrams that comprise that dependency. If we allow probability to stand in for grammaticality, this means a child can have a judgment about the grammaticality of any *wh*-dependency, based on its probability.

To generate the probability of a *wh*-dependency from its container node trigrams, Pearl & Sprouse proposed to simply use the smoothed product of its trigrams, as in (12). Smoothing the trigram probabilities means assigning a very small amount of probability to trigrams that have never been observed, just in case they're actually okay but haven't appeared in the child's intake for whatever reason. That is, the child doesn't automatically rule out a dependency containing a trigram she's never encountered before – she just doesn't like it very much.

(12) p(Who did Jack think that the story about penguins amused ___?)
$= \prod_{trigram \in Trigrams_{IP-VP-CP_{that}-IP-VP}} p(trigram)$

= p(*start*-IP-VP)*p(IP-VP-CP$_{that}$)*p(VP-CP$_{that}$-IP)*p(CP$_{that}$-IP-VP)*p(IP-VP-*end*)

So, to generate behavior we can compare against the acceptability judgments from Sprouse et al. (2012), we have the modeled learner generate a probability for each of the stimuli, based on the trigram probabilities that the learner's internalized during the learning process. If the modeled learner implicitly has the same knowledge about syntactic islands as adults do, it should demonstrate the same superadditivity in its judgments.

3.4.4 Learning period

How long does our modeled learner get to learn? One way to think about this concretely for acquisition modeling is how much data the learner encounters before the learning period is over. Hart and Risley (1995) determined that American children in their samples were exposed to approximately one million utterances between birth and three years old, and so

Pearl & Sprouse leveraged that information. By positing that three years would be a reasonable learning period for syntactic islands (say, between the ages of two and five), Pearl & Sprouse assumed the modeled learner would encounter a million utterances during those three years. Because *wh*-dependencies made up approximately 20% of the child-directed input utterances in the CHILDES-based sample, this translates to the modeled learner encountering 200,000 *wh*-dependencies during the learning period that are distributed similarly to the CHILDES-based sample (i.e. mostly IP-VP and IP, with others appearing infrequently).

3.4.5 Results

Acquisition success is measured by the ability of the modeled learner to judge the *wh*-dependency stimuli from Sprouse et al. (2012) the same way adults did. So, after learning from the container node trigrams in the acquisitional intake, the modeled learner can generate a probability for each dependency. We can plot the log probability on the y-axis of an interaction plot[8] to indicate how grammatical the dependency is perceived to be, just as z-scores were used in the adult judgment data. The main signature of syntactic islands is the qualitative pattern of superadditivity.[9] Specifically, if we see superadditivity in the modeled learner's generated judgments, the learner has demonstrated implicit knowledge of these four syntactic islands. Figure 5 shows the modeled learner's generated judgments for each of the four island types after learning from child-directed speech data, with the log probability on the y-axis.

As we can see from the interaction plots, the modeled learner does indeed display the qualitative target behavior indicating implicit knowledge of these four syntactic islands. So, a child using the Subjacency-ish representation would be able to learn about these islands.

3.5 Subjacency-ish representation: Summary

So what did we discover? First, the Subjacency-ish representation coupled with a learning strategy that relies on container node (CN) trigram frequencies is useful for acquisition. This provides validation for this representation: if dependencies are represented as CN sequences like these, acquisition works well for these four islands. So, children could leverage CN trigrams to implicitly internalize a representation of syntactic islands. Moreover, by considering the acquisition implications, we were able to refine the definition of CNs along the way to include subcategorized CP nodes.

Second, these results also have something to contribute to the UG debate. Remember from Table 4 that the Subjacency-ish representation required no components that were obviously dedicated solely to learning islands the way that some of the Subjacency components

[8]Note that all log probabilities are negative because raw probabilities are between 0 and 1, and so the log probability is between negative infinity and 0 (e.g., log(0.000001) = -6 while log(1) = 0). This means the numbers closer to zero are more positive and appear higher on the y axis – these represent structures judged by the modeled learner as "more acceptable". Numbers further from zero are more negative and appear lower on the y axis – these represent structures judged "less acceptable".

[9]Because we currently don't have a precise theory for translating probabilities into acceptability judgments, it doesn't make as much sense to look for a quantitative match. This is because actual acceptability judgments are based on many factors that are not included in this model, such as lexical item choice, semantic probability, and processing difficulty (Schütze 1996; Cowart 1997; Keller 2000; Sprouse 2009).

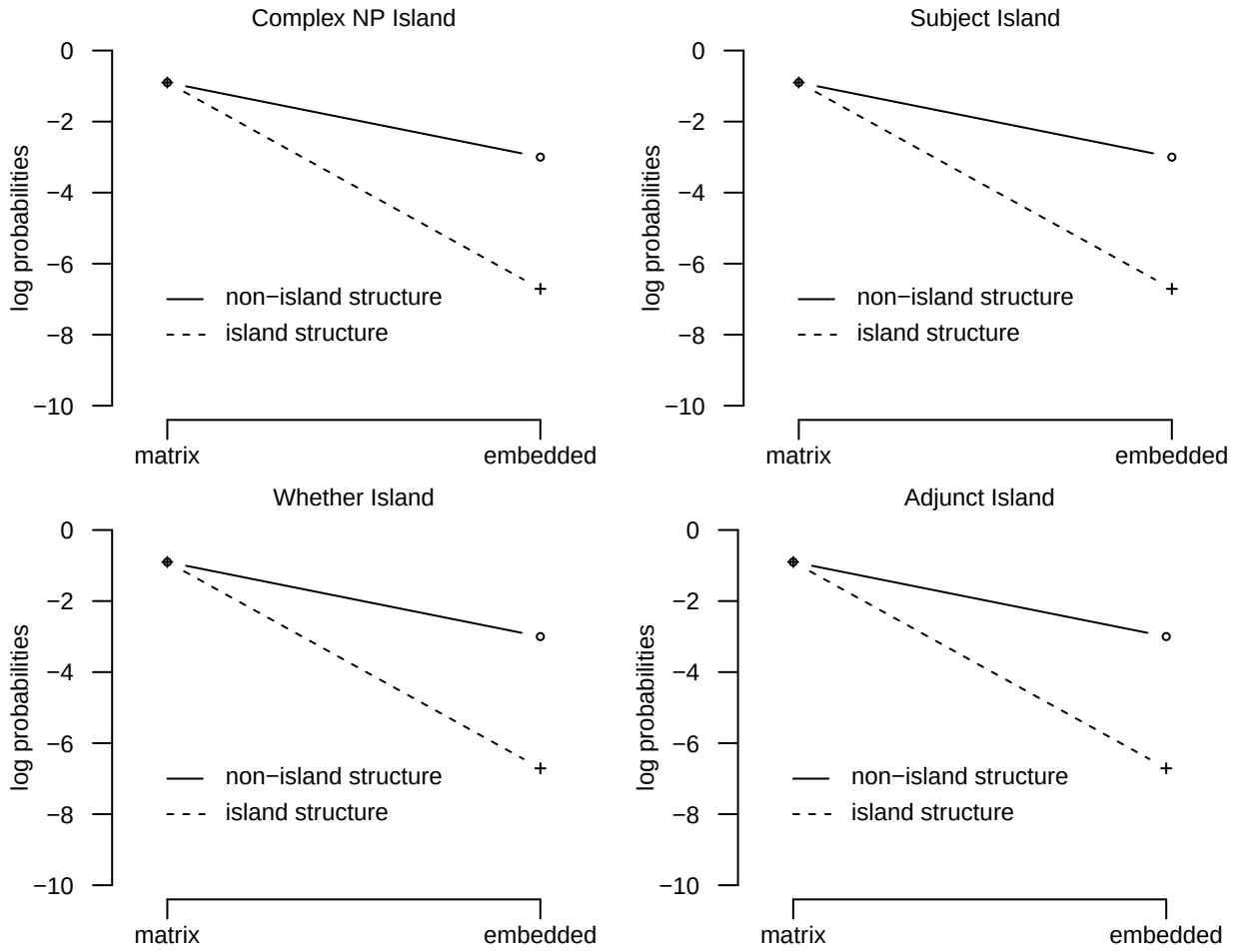

Figure 5: Modeled learner results, after learning from child-directed speech data.

were. In the same way, Subjacency-ish also requires fewer components that are necessarily UG. More specifically, a UG component is both innate and language-specific. While defining *wh*-dependencies with container nodes is clearly language-specific, it's an open question whether container nodes are necessarily innate (though some part of them might well turn out to be). Moreover, tracking trigram frequencies and dispreferring low-probability sequences is probably innate, but clearly not language-specific. So, the upshot is that the Subjacency-ish representation offers an alternative for representing syntactic island knowledge that has fewer necessarily UG components, and the potentially UG components are likely more general-purpose rather than being useful only for syntactic islands.

Similar to our case study with English metrical stress, this acquisition modeling approach has allowed us to evaluate a representation (here, of *wh*-dependencies) by using it for acquisition. We're then able to refine our theories about (1) what part of the representation must be in UG, and (2) the acquisition strategy that accompanies the representation for syntactic islands.

4 Closing thoughts

What I've hoped to show you here is how we can use acquisition modeling to get at the larger goal of informing theories of representation and theories of acquisition at the same time, given the natural link between them. In the metrical stress case study, I identified learning assumptions about using productive data filters that benefit children learning English and which can differentially benefit proposed stress knowledge representations. I also identified alternative English grammars within these representations that are similar to the current definitions in many respects but which are likely easier to learn from the English data children typically encounter. These serve as empirically motivated alternative proposals for what the English grammar actually is for each representation. For the syntactic islands case study, I provided empirical validation for a proposed syntactic island representation, which then yielded alternative proposals for the contents of Universal Grammar. I also provided a concrete demonstration of a learning strategy that could use that representation and succeed when given cognitively plausible input data. So, I believe this acquisition modeling approach can be a really useful tool for linking theories of representation with theories of acquisition. I hope that we'll keep using it to inspire, test, and adapt both kinds of theories.

Acknowledgements

I am especially grateful to Jon Sprouse, Tim Ho, and Zephyr Detrano for being delightful collaborators on the projects discussed here, and to Joanna Lee for her help analyzing the metrical stress data. I'm also indebted to Jeff Lidz and Jeff Heinz for being wonderfully supportive of this work, and to Rachel Dudley for organizing the workshop where these ideas were presented. These ideas have additionally benefited from discussion with Pranav Anand, Misha Becker, Bob Berwick, Adrian Brasoveanu, Alex Clark, Sandy Chung, Bob Frank, Norbert Hornstein, Jim McCloskey, Armin Mester, Colin Phillips, William Sakas, Virgina Valian, Matt Wagers, Charles Yang, and several anonymous reviewers, as well as the audiences at ISA 2012, UMaryland Mayfest 2012, NYU Linguistics 2012, JHU Cognitive Sciences 2013, UC Irvine IMBS 2013, UC Irvine LPS 2013, UC Santa Cruz Linguistics 2014, and BLS 2014. This work has additionally been supported by NSF grants BCS-0843896 and NSF BCS-1347028.

References

Ben Ambridge, Julian Pine, and Elena Lieven. Child language acquisition: Why Universal Grammar doesn't help. *Language*, 90(3):e53–e90, 2014.

Joanne Arciuli, Padraic Monaghan, and Nada Seva. Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language*, 63(2):180–196, 2010.

R Baayen, R Piepenbrock, and L Gulikers. CELEX2 LDC96L14. `https://catalog.ldc.upenn.edu/LDC96L14`, 1995.

Michael R Brent and Jeffrey Mark Siskind. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44, 2001.

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

Kimberly Cassidy and Michael Kelly. Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin & Review*, 8(3):519–523, 2001.

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

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 1969.

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

Noam Chomsky. *Barriers*, volume 13. MIT press, 1986.

Noam Chomsky and Morris Halle. *The Sound Pattern of English*. Harper and Row, New York, NY, 1968.

Noam Chomsky and Howard Lasnik. The Theory of Principles and Parameters. In Noam Chomsky, editor, *The Miminalist Program*, pages 13–128. MIT Press, Cambridge, MA, 1995.

Wayne Cowart. *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications, 1997.

Stephen Crain and Janet D Fodor. How can grammars help parsers? In David Dowty, Lauri Kartunnen, and Arnold Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 94–128. Cambridge University Press, Cambridge, UK, 1985.

Stephen Crain and Paul Pietroski. Why language acquisition is a snap. *The Linguistic Review*, 19:163–183, 2002.

B. Elan Dresher. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30(1):27–67, 1999.

Catharine Echols. A perceptually-based model of children's earliest productions. *Cognition*, 46(3):245–296, 1993.

Lyn Frazier and Giovanni B Flores D'Arcais. Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language*, 28(3):331–344, 1989.

LouAnn Gerken. Young children's representation of prosodic phonology: Evidence from english-speakers' weak syllable productions. *Journal of Memory and Language*, 33(1): 19–38, 1994.

LouAnn Gerken. Prosodic structure in young children's language production. *Language*, 72 (4):683–712, 1996.

Morris Halle and Michael Kenstowicz. The Free Element Condition and cyclic versus non-cyclic stress. *Linguistic Inquiry*, 22:457–501, 1991.

Morris Halle and Jean-Roger Vergnaud. *An essay on stress*. Mit Press, Cambridge, MA, 1987.

Michael Hammond. *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. Oxford University Press, Oxford, UK, 1999.

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

Bruce Hayes. Extrametricality and English stress. *Linguistic Inquiry*, 13:215–225, 1982.

Bruce Hayes. *Metrical stress theory: Principles and case studies*. University of Chicago Press, Chicago, IL, 1995.

Jeffrey Heinz. Computational Theories of Learning and Developmental Psycholinguistics. In Jeffrey Lidz, William Synder, and Joe Pater, editors, *The Cambridge Handbook of Developmental Linguistics*. Cambridge University Press, 2014. To appear.

C.-T. James Huang. *Logical relations in Chinese and the theory of grammar*. PhD thesis, MIT, Cambridge, MA, 1982.

Margaret Kehoe. Support for metrical stress theory in stress acquisition. *Clinical linguistics & phonetics*, 12(1):1–23, 1998.

Frank Keller. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh, Edinburgh, UK, 2000.

Michael Kelly. Rhythmic alternation and lexical stress differences in English. *Cognition*, 30: 107–137, 1988.

Michael Kelly and Kathryn Bock. Stress in time. *Journal of Experimental Psychology*, 14: 389–403, 1988.

Paul Kingsbury, Stephanie Strassel, Cynthia McLemore, and Robert MacIntyre. CALLHOME American English Lexicon (PRONLEX). `https://catalog.ldc.upenn.edu/LDC97L20`, 1997.

Paul Kiparsky. Metrical structure assignment is cyclical. *Linguistic Inquiry*, 10(4):421–441, 1979.

Howard Lasnik and Mamuro Saito. On the nature of proper government. *Linguistic Inquiry*, 15:235–289, 1984.

Julie Legate and Charles Yang. Assessing Child and Adult Grammar. In Robert Berwick and Massimo Piatelli-Palmarini, editors, *Rich Languages from Poor Inputs*, pages 168–182. Oxford University Press, Oxford, UK, 2013.

Jeffrey Lidz and Annie Gagliardi. How Nature Meets Nurture: Universal Grammar and Statistical Learning. *Annual Review of Linguistics*, 1(1):333–352, 2015.

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

David Marr. *Vision*. W.H. Freeman, San Francisco, CA, 1982.

Daniel N. Osherson, Michael Stob, and Scott Weinstein. *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press, Cambridge, MA, 1986.

Joe Pater. Non-uniformity in English secondary stress: The role of ranked and lexically specific constraints. *Phonology*, 17(2):237–274, 2000.

Lisa Pearl. *Necessary Bias in Natural Language Learning*. PhD thesis, University of Maryland, College Park, College Park, MD, 2007.

Lisa Pearl. Learning English Metrical Phonology: When Probability Distributions Are Not Enough. In Jean Crawford, Koichi Otaki, and Masahiko Takahashi, editors, *Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition, North America (GALANA 2008)*, pages 200–211. Cascadilla Press, Somerville, MA, 2009.

Lisa Pearl. When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition*, 18(2):87–120, 2011.

Lisa Pearl. Evaluating learning strategy components: Being fair. *Language*, 90(3):e107–e114, 2014.

Lisa Pearl and Jon Sprouse. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20:19–64, 2013a.

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

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

Lisa Pearl, Timothy Ho, and Zephyr Detrano. More learnable than thou? Testing metrical phonology representations with child-directed speech. In Herman Leung, Zachary O'Hagan, Sarah Bakst, Auburn Lutzross, Jonathan Manker, and Nicholas Rolle, editors, *Proceedings of the Berkeley Linguistics Society*, pages 398–422. Berkeley Linguistics Society, 2014.

Lisa Pearl, Timothy Ho, and Zephyr Detrano. An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. Manuscript, 2015.

Michèle Pettinato and Jo Verhoeven. Production and perception of word stress in children and adolescents with Down syndrome. *Down Syndrome Research & Practice*, 13:48–61, 2008.

Colin Phillips. The real-time status of island phenomena. *Language*, 82:795–823, 2006.

Steven Pinker. Formal models of language learning. *Cognition*, 7:217–283, 1979.

Alan Prince and Paul Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. ROA, New Brunswick, NJ, 2002.

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

Carson T Schütze. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press, 1996.

Jon Sprouse. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, 40(2):329–341, 2009.

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

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

Bruce Tesar and Paul Smolensky. *Learnability and optimality theory*. The MIT Press, Boston, MA, 2000.

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

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

Michael Wilson. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.

Charles Yang. On productivity. *Yearbook of Language Variation*, 5:333–370, 2005.