Three Equations:
A Formalist Perspective on Language Acquisition

Charles Yang[*]
charles.yang@ling.upenn.edu

February 15, 2018

Abstract

Language acquisition is a computational process by which language-specific experience is integrated into the learner's initial stage of knowledge. To understand language acquisition thus requires precise statements about these components and their interplay, stepping beyond the philosophical and methodological disputes in the field such as the generative vs. usage-based approaches. I review three simple equations that have guided the study of child language acquisition: How learners form and select linguistic hypotheses, and how researchers assess the progress of language acquisition with rigor and clarity. Furthermore, I suggest that these equations are equally applicable to second language acquisition, yielding potentially important insights on the continuities and differences between child and adult language.

1 A Formal Introduction

Don Knuth, perhaps the most renowned living computer scientist, infamously took *Syntactic Structures* on his honeymoon (2002). For many language scientists, and for many more outside of linguistics, generative grammar initiated a precise way of studying a quintessential aspect of human life. It was a refreshing change from the routine in traditional social and behavioral sciences: take measurements, fit a curve, repeat. The commitment to a causal and mechanical account of language helped establish the field of cognitive science; the mind may be studied with deductive methods as in the natural sciences.

This article provides a personal perspective on the formalist approach to language acquisition. I use the term "formalist" in the methodological sense: it reflects a commitment to the mechanistic study of language, rather than the conventionalized standin for "generative" and/or "nativist". In my opinion, methodological rigor should override the researcher's preference for the mode of explanation. Indeed, as someone trained in the generative tradition which emphasizes domain-specific knowledge, I have been invigorated by usage-based research which places

[*]I would like to thank Holger Hopp, Jason Rothman, and Neal Snape for comments on an earlier draft.

greater emphasis on the role of data-driven learning. More specifically, a formal theory of language acquisition is much more than a description: children know A at age X but B at age $X+Y$ (e.g., the transition from an item-based to a rule-based grammar), or that some variable P (e.g., input frequency) is correlated with some other variable Q (e.g., the correct rate of morphological marking). The theory must include a mechanistic account of how the A-to-B transition takes place and how P is causally responsible for Q. After all, it was such a mechanistic approach that got everyone excited about linguistics and helped build connections with other fields.

In the spirit of being formal, I have organized this paper along three simple equations. As I will review in Section 2, 3, and 4, they have been developed to address some long-standing puzzles and debates in first language acquisition. The main thread connecting these equations is the recognition that while the input data critically matters for language acquisition, the learner must go beyond and form abstract generalizations — the logical problem of language acquisition (Hornstein and Lightfoot 1981), which must be studied with rigor and precision if we are to make any progress. In Section 5, I submit, with some trepidation, that these equations may also prove useful for second language acquisition. This is because learning a second language, which is by definition a linguistic and cognitive process, must also go beyond the data to form generalizations – the logical problem of second language acquisition (White 1985, Bley-Vroman 1989). The equations provide very concrete predictions that can be easily confirmed or disconfirmed. To the extent they are confirmed, we may detect potential continuities between child and adult language acquisition. To the extent they are disconfirmed, we may be a step closer to understanding why children seem better at language learning than adults.

2 Selection: $p' = p + \gamma q$

There was a time when formal methods were integral to the empirical research on child language: I have in mind the influential work of Suppes (1974), Pinker (1979), Wexler and Culicover (1980), Berwick (1985), etc. Formal models make explicit statements about the learner's predisposition for language, the ecological condition of language acquisition (e.g., no negative evidence), and the learning algorithms likely within children's computational capacity (e.g., incremental learning). The commitment to a formal account of language learnability in fact inspired the modern study of machine learning and statistical inference (Solomonoff 1964, Gold 1967, Blum and Blum 1975, Angluin 1980).

The scene had already changed when I started working on child language in the late 1990s. The field, at least the part of the field I was embedded in, had shifted primarily to developmental issues. A major debate at the time was whether the differences between child and adult language are due to competence or performance gaps (e.g., Pinker 1984, Borer and Wexler 1987, Demuth 1989, Bloom 1990, Valian 1991, Wang et al. 1992, Hyams and Wexler 1993). But both sides of the debate agreed that the child's grammar is already target-like. This was a natural position for the performance-based theorists but seemed paradoxical under the competence-based account: if children have exquisite knowledge of their grammar (Wexler 1998), how come they don't talk right? Conspicuously absent is a learnability account of *how* children's grammar becomes target-like: both sides in effect deny the role of input and experience, because language-specific data has no explanatory power on the acquisition of the grammar.[1]

As someone on the outside looking in, this state of affairs was fascinating but also puzzling. The quantitative work that began to emerge in the early 1990s, thanks in no small part to the CHILDES project (MacWhinney 2000), confirmed that many aspects of child language are indeed adult-like from very early on, as Roger Brown recognized long ago (1973, p. 156). A highlight was the near-perfect correlation between the position and inflection of the verb in the main clause (Pierce 1992, etc.), which poses significant challenges for adult second language learners (White 1990). At the same time, the problem of how children learn their specific grammars was left unaddressed: What was a computer scientist to do when the mechanism of learning from experience was sidelined?

On top of this, and perhaps because of it, there was a widely held belief that the commitment to Universal Grammar, shared broadly by both sides of the competence/performance divide, is inherently incompatible with input effects in language acquisition (Tomasello e.g., 2003, p. 97, Hoff 2014, p. 106). Any type of distributional learning from data was viewed as evidence against Universal Grammar; see, for instance, "learning rediscovered" (Bates and Elman 1996) after the discovery of statistical learning for word segmentation (Saffran et al. 1996). The same period also witnessed the so-called English past tense debate (Pinker and Ullman 2002, McClelland and Patterson 2002). Here the disagreement is on the treatment of the regular verbs — whether it is association-based or whether it is handled by a rule that adds "-ed". For both sides, irregular past tense is formed by associative memory thus sensitive to frequency effects, apparently incompatible with the symbolic treatment of irregulars throughout the history of linguistics (Bloch 1947, Chomsky and Halle 1968).

It was in this context that I proposed the variational learning model (Yang 2002). It was an acknowledgment that input effects matter for language acquisition but are completely consistent with the theory of Universal Grammar. It was also a return to the formalist tradition of language research. Rejecting the dominant view that the child language is characterized by a single grammar (e.g., the adult-like grammar, or a grammar in the space of possible grammars as in the Principles and Parameters framework and Optimality Theory), the variational model assumes that the grammars in the child's hypothesis space are associated with probabilities or weights. Learning takes place not by changing one grammar to another (e.g., Wexler and Culicover 1980, Berwick 1985, Gibson and Wexler 1994) but as changes in the probabilistic distribution of the grammars in response to input data. The simplest instantiation of the variational model is the Linear Reward Penalty scheme (Bush and Mosteller 1951), one of the oldest and best supported models from mathematical psychology.

For the purpose of illustration, consider a learner who has access to two grammars, the target A and a competitor B, which are currently associated with probabilities p and q. Upon encountering an input item s, the learner selects a grammar with its associated probability. Suppose A is chosen:

(1) a. If A can analyze s then $p' = p + \gamma q$ and $q' = (1-\gamma)q$
    b. If A cannot analyze s then $p' = (1-\gamma)p$ and $q' = q + \gamma p$

[1] Under some accounts, children's output is constrained by performance filters that are themselves subject to language variation (e.g., Gerken 1994, Demuth 1996), but these accounts also assert the correctness of child's competence grammar.

Several remarks are in order. First, the competition scheme in the variational model implies some notion of fitness, and it is the fitness differential of the grammars that drives learning. In the simplest case, the target grammar A by definition always succeeds: p will rise whenever A is selected. The competitor B, by definition, must fail on a certain proportion of the input: when that happens, q will decrease and p will thus increase. But importantly, B needn't fail all the time: there may be input items that are ambiguous between the grammars. Thus, the trajectory of learning in the long run is determined by the statistical composition of the input data. A grammar whose competitor is penalized more often will be learned faster. Second, the grammar-input compatibility, referred to as "analyze" in (1), can be flexibly defined as long as it is precise and independently motivated. The simplest case would be parsability, i.e., whether the grammar is compatible with an input string, but many other considerations are possible. For instance, if the child's parsing system has certain limitations (Trueswell et al. 1999), then even sentences compatible with the target grammar may fail to be analyzed. The fitness values may also socially conditioned: a stigmatized variant would put its competitors at an advantage.[2] Third, the variational model does not require that the hypotheses in competition are innately available. In fact the formalism is applicable to any finite set of hypotheses, including hypotheses that the learner constructs on the basis of specific language input. For instance, the model has been applied to word learning to represent the probabilistic association between the phonological form of a word and its meaning (Stevens et al. 2017), both of which are clearly learned from the environment. Finally, the variational model leaves space for individual variation. The statistical composition of the input may vary such that the target grammar may develop along different schedules for individual learners. It is also possible that some children are just slower at absorbing linguistic input than others; this is operationalized by the learning rate parameter γ in (1), which represents the magnitude of probability adjustment as the result of analysis, again a familiar notion from the mathematical psychology of learning. It has been suggested that individual variation in γ is a source for developmental delays in language (Legate and Yang 2007).

The variational learning model was originally applied to the problem of parameter setting: for A and B in (1), think of the opposite values of a parameter. The model provably converges on the target grammar in the limit (Straus 2008). In a complex domain of thirteen word-order parameters (Sakas and Fodor 2012), the variational model has been shown to converge on the target consistently and efficiently (Sakas et al. 2017). More important, the variational model resolves several major challenges associated with traditional approaches to parameter setting. Its probabilistic nature means that the target grammar will only gradually rise to dominance under the cumulative effect of unambiguous data in its favor. Two empirical consequences follow.

First, it is possible to establish the amount of unambiguous evidence for parameter values in child-directed input corpora to correlate with the developmental time course of the parameters. For instance, languages differ in the positioning of the main verb in the matrix clause: for languages like English, the verb follows adverbs (e.g., *John often drinks coffee*) whereas for languages like French, the verb precedes adverbs (e.g., *Jean boit souvent du café*). Only sentences that contain positional signposts such as the adverb *often/souvent* can unambiguously nudge the learner toward their language-specific option (White 1990); see Yang (2012) for a review of such

[2] All the same, it is important to recognize that the fitness value, e.g., the probability with which a grammar fails to analyze the input data, is *not* something the learner needs to calculate — no more than the mouse needs to tabulate the probabilities of receiving food pellets in conditioning experiments (Bush and Mosteller 1951).

input effects of parameters across languages. Recent work has devised intervention strategies (Hadley and Walsh 2014, Hadley et al. 2017) to boost individual learner's language development by amplifying the volume of informative data in the caretaker input as formulated by Legate and Yang (2007).

Second, under the variational model, children's systematic deviation from the target grammar may be attributed to non-target hypotheses before their eventual demise; see Crain et al. (2016) for a recent review. Naturally, this perspective is only as good as what we take to be the space of linguistically possible hypotheses available to child. While few would claim that human languages can vary arbitrarily, there is still considerable debate whether such constraints are specific to language or result from the constellation of other cognitive factors. At the same time, the theory of parameters is currently under review even within generative linguistics, especially in light of the severe constraints placed on the faculty of language by evolutionary considerations (Hauser et al. 2002, Chomsky 2005, Lardiere 2009, Yang et al. 2017).

In my opinion, a theory of language acquisition need not be overly bound to the latest theory of language structures. Child language can often be fruitfully studied as the level of empirical generalization which, if sufficiently robust, can not only withstand the changing theoretical perspectives but also actively constrain them. Here I review one specific line of evidence uncovered through the variational model. Consider the classic problem of null arguments in child English (Bloom 1970). English-learning children frequently omit subjects — up to 30% of the time — and they occasionally omit objects as well, quite contrary to the input data they hear. (2) provides some naturally occurring examples from the CHILDES database.

(2) _ want cookies.
Where _ going?
How _ wash it?
Erica took _.
I put _ on.

These missing arguments generally do not impede language understanding as the intended meanings can generally be inferred from the context. Nevertheless, children do not start using subjects and objects consistently at adult level around the third birthday, in a direct contrast to the input data. Earlier attempts to equate the null argument stage to parameter missetting to the pro-drop or topic-drop option (Hyams 1986, 1991) were unsuccessful: during no stage of acquisition does the distribution of English-learning children's argument use resemble that of speakers or learners of pro-drop and topic-drop languages (Valian 1991, Wang et al. 1992), nor is there any evidence for sudden changes in the frequency of null subjects which would have supported the notion of parameter resetting (Bloom 1990, Legate and Yang 2007).

The variational model offers a new perspective on the null argument phenomenon. When an English-learning child probabilistically accesses the target grammar, no arguments would be omitted. But when the topic-drop grammar is accessed,[3] argument omission would be possible when the discourse conditions are met. The most telling evidence can be found in a striking distributional property in child English. It is easy to find hundreds of child English examples in the CHILDES corpus of the following type:

[3] See Yang (2002, p. 118f) on why the pro-drop grammar is eliminated much earlier.

(3) a. When _ bring the bag back? When _ rains?
b. Where _ get these? Where _ go? Why _ go slowly?
c. Why _ get scratched by the cat? Why _ laughing at me?
d. How _ fix my eye? How _ do open it?

These questions have target-like fronting of the Wh-question word but the subjects are missing. Notice that these are all *adjunct* questions with *when*, *where*, *why*, and *how*. By contrast, omitted subjects in *argument* Wh-questions are vanishingly rare:

(4) a. *Who_t _ kissing *t*?
b. *Who_t _ see *t*?
c. *What_t _ want to hit *t*?

That is, when the object (*who* and *what*) is fronted in a Wh-question, the subjects are almost never omitted. An exhausted search of Wh-questions produced by Adam, a prolific subject dropper, reveals a strong asymmetry (Yang 2002, p. 120):

(5) a. 95% (114/120) of the Wh-questions with an omitted subject are adjunct questions.
b. 97.2% (209/215) of the Wh-object questions contain subjects.

The null subject asymmetry in English-learning children's argument and adjunct questions is exactly mirrored in topic-drop languages. In Chinese, for instance, subject drop is possible under topicalization only if the fronted topic (in italic and marked with trace) is an adjunct (6a) but not an argument (6b):[4]

(6) a. *$\textit{Mingtian}_t$*, [_ juede [*t* hui xiayu]]. (_ = John).
Tomorrow, [_ believe [*t* will rain]].
'It is tomorrow that John believes will rain.'
b. **$\textit{Bill}_t$*, [_ juede [*t* shi laoshi]]. (_ = John).
Bill, [_ believe [*t* is teacher]].
'It is Bill that John believes is the teacher.'

Even more direct parallelism comes from Brazilian Portuguese, a language that has overt movement in Wh-questions (like English) but omits arguments in certain contexts (like Chinese).[5]

(7) a. $\text{Quando/Como/Onde}_t$ _ beijou *t*?
When/How/Where_t _ $\text{kissed}_{\text{2/3P}}$ *t*?
'When/How/Where_t did you/they kiss?'

[4]The null subject asymmetry in topic-drop grammars as shown in (6a) and (6b) seems to have been overlooked in the theoretical literature (e.g., Biberauer et al. 2010). So far as I know, its discovery was prompted by the need to explain the *English*-learning children's null subject usage — (3) vs. (4) —under the variational learning framework. For present purposes, it is not important to precisely formulate the theoretical underpinning of the observed asymmetry. One plausible idea is some discourse type of Relativized Minimality (Rizzi 1990): a new topic of the same type as the old topic blocks its accessibility. Thus, null subject via topic drop is possible if the new topic cannot cause interference (i.e., an adjunct not an argument.)

[5]I think Pablo Faria and Guilherme Garcia for the data reported here.

b. *Quem$_t$ _ beijou t?
Whom$_t$ _ you/they kiss t?
'Whom did you/they kiss?'

The verbal morphology of Brazilian Portuguese has become too impoverished to support the agreement-based pro-drop option in its European cousin, as can be seen in the inflectional form of *beijou* in (7). Topic drop á la Chinese is the only option, and we see the exact asymmetry between adjunct and argument Wh-questions.

Some may object to calling these grammatical options *parameters* but what's at stake is not terminological. The main generalization is that English-learning children spontaneously exercise a grammatical option never attested in their environment but used by speakers thousands of miles away. This option has to be suppressed by language-specific data: i.e., the use of non-referential (expletive) subjects such as *There is a car coming* and *It seems that the kids are tired*, which is not required in languages such as Chinese (Huang 1984). For instance, where the grammatical subject *it* must be present for the English expression *It is going to rain*, the position can be empty in Chinese ('_ yao xiayu' or "will rain"; Wang et al. 1992). Because expletive subject sentences are infrequent, making up about 1% of child-directed input, the rise of the obligatory subject grammar is gradual, according to the variational model. And the topic-drop grammar will be exercised during the process, resulting in null subjects in child English as well as occasionally null objects, which the object happens to be the discourse topic.

Naturally, the same model ought to account for the acquisition of topic- and pro-drop languages (Valian 1991, Wang et al. 1992, Kim 2000, Grinstead 2000) which, in contrast to the considerable delay in English, show very early adult-like command of subject use. Yang (2002) provides a detailed quantitative account of these findings which will be summarized in Section 5.1 when I discuss the role of UG in adult language acquisition.

3 Rules: $p = 1/(rH_n)$

At the 2005 LSA summer institute, I organized a workshop called "Nuts and Core",[6] borrowing Culicover's (1999) term for linguistic idiosyncracies that cannot all be plausibly attributed to an innate grammatical core. The questions posed to the participants, all prominent scholars in the generative vs. constructivist debate, were as follows:

(8) a. If the core is dispensed with, how does the learner go from specific constructions to general regularities in syntax (Tomasello 2003)? What kind of constraints are needed for learning to be efficient and successful?

b. If the core is to be maintained, how might one construe a principled theory that keeps the core and the nuts separate (Fodor 2001)? How does the setting of a parameter value tolerate exceptions?

The workshop was lively but ended in a state of impasse. The main point of contention was whether child language is abstract and productive or item-based and lexically conservative. In many ways, the debate resembled the earlier competence-performance dispute: both sides offer

[6] https://linguistlist.org/issues/16/16-2050.html

useful but only partial explanations of child language. Later in this paper I will turn to my own proposal of how children discover productive rules — and potentially a reconciliation of the two approaches — but first, an assessment of the empirical and theoretical claims is in order.

The first point to make is that an early stage of lexically specific language is neither a novel observation nor a feature unique to the usage/construction-based approach (despite being its "central tenet"; Diessel 2013). Again, let's turn to the study of English past tense. A well-known pattern is the U-shaped curve of development. Children's verbal inflection is initially conservative: very few regular verbs are consistently marked in past tense, and the irregular verbs, when marked, are marked correctly. This stage is followed by the emergence of overregularization errors, which we can observe in longitudinal records. For instance, Adam's transcripts started at 2;3; all irregular verbs were marked correctly until 2;11, when he produced the utterance "What dat feeled like" (Marcus et al. 1992, Pinker 1995). Since *feeled* cannot be attributed to the input, the error marks the elevation of "-ed" to the status of a productive suffix. Thus English past tense is a classic case of initial conservatism followed by productive generalization. The phenomenon was central to the past tense debate and especially the dual-route model developed by Clahsen, Marcus, Pinker, and Prince: all avowed nativists.

The second, and more important, point is empirical: the evidence for an initial item-based stage of child language has been overstated and uncritically accepted. For example, high frequency combinations such as "give me" (sometimes "gimme") have been interpreted as "unanalyzed" collocations and presented as evidence for the lexically specific stage of language development (Lieven et al. 1992, Tomasello 1992, 2003). These expressions are indeed statistically dominant but to conclude that they are item based requires more work. At a minimum, one needs to show that their frequencies are conspicuously higher than expected had the words been combined statistically independently. In fact there seems nothing remarkable at all about "give me". Consider the transcripts of the Harvard children Adam, Eve, and Sarah (Brown 1973), datasets that have been in the public domain for decades. A simple search reveals the frequencies of *give me*, *give him* and *give her* to be 95, 15, and 12, for the ratio of 7.75:1.23:1. It is thus quite likely that when working with a relatively small child corpus, *give* is only paired with *me*, which was perhaps the reason behind "give me" as a paradigm case of item-specific learning. But another simple search shows that the frequencies of *me*, *him*, and *her* in these children's production data are 2,949, 484, and 375, or 7.86:1.29:1. That is, the statistics of "give me" actually support the productive and independent combination of the verb and the object!

A quantitative assessment of a (coherently formulated) null hypothesis appears totally lacking in the usage-based literature. Consider three key case studies highlighted in Tomasello's influential paper *Do young children have adult syntactic competence* (2000) and other publications:

(9) a. The Verb Island Hypothesis (Tomasello 1992). Most of the verbs and predicates in early child language are used with one or very few possible frames.

b. Limited morphological inflection (Pizzuto and Caselli 1994). Almost half of the verbs in child Italian were used in one person-number agreement form (out of six possibilities), and only 13% of all verbs appeared in four or more forms.

c. Determiner imbalance. Pine and Lieven 1997 find that only 20-40% of nouns that have been used the determiner *a* or *the* are used with both, despite the general interchangeability of the determiners (e.g., *a/the dog*, *a/the chair*, etc.).

So far as I can tell, these claims have been presented, and accepted, with evaluating an alternative hypothesis. For instance, it would have been worthwhile to subject *adult* language to item-based claims: alarm bells would have sounded. With the respect to determiner use (9c), quantitative analysis of English print materials such as the Brown Corpus (Kučera and Francis 1967) and child-directed speech reveals comparable, and comparably low, combinatorial diversity as children (Valian et al. 2009), yet adults' grammatical ability is not in question.[7]

To develop a principled quantitative interpretation of language must take Zipf's Law (1949) into account. For reasons no one quite understands (Miller 1957, Chomsky 1958, Mandelbrot 1953; see Yang (2013b) for an exposition), word frequency is inversely proportionally related to rank. Specifically, let there be N unique words in a corpus. For the r-th ranked word, its frequency f is C/r where C is some constant. Thus, its probability of use p can be expressed as:

(10)

$$\begin{aligned} p &= \frac{C/r}{C/1 + C/2 + \ldots + C/N} \\ &= \frac{1}{rH_N} \text{ where } H_N = \sum_{i=1}^{N} \frac{1}{i} \text{ is the } N\text{th Harmonic number} \end{aligned}$$

The most obvious feature of Zipf's Law is the characteristic long tail: most linguistic units such as words, and by extension, combinations of words, are rarely used even in very large corpora (Jelinek 1998). This suggests that the sparsity of syntactic combinations in children's early language, or indeed any linguistic sample, is inherent: It does not automatically support lexically specific learning, and it may even support a rule-based grammar.

I have developed a statistically rigorous benchmark for assessing grammatical productivity (Yang 2013a). The test incorporates Zipf's Law to approximate word probabilities and their combinations. Let's consider its application to the sparsity of determiner-noun combinations in child and adult language. Suppose we have a corpus of N (singular) noun types that appear in S pairs of *a/the*-noun combinations. The expected probability for the r-th ranked noun having been paired with both *a* and *the* is given below:

(11)

$$E_r = 1 - (1-p)^s - \sum_{i=1}^{2} \left[(\mathbf{f_i p} + 1 - p)^S - (1-p)^S \right] \text{ where } p = \frac{1}{rH_N}$$

Here f_1 and f_2 are the probabilities of the two determiners. In general, nouns heavily favor one of the two determiners, again characteristic of Zipf's Law. An example can be seen in the noun *bathroom*: although both *a bathroom* and *the bathroom* are grammatical, the latter is a great deal more frequent than the former. And for the noun *bath*, the opposite is true. The imbalance of determiner-noun combinations greatly contributes to the lower diversity in child English samples — and indeed all language samples. But the most important point of the equation

[7] Similar observations hold for verb islands (9a) and inflectional morphology (9b); see Kowalski and Yang (2012) and Yang (2016, chapter 2). Again, no alternative hypotheses were rigorously tested.

Table 1: Empirical and expected combinatorial diversity in L1 English (adapted from Yang 2013a)

Subject	Sample size (S)	Types (N)	Empirical	Expected
Naomi (1;1-5;1)	884	349	19.8%	21.8%
Eve (1;6-2;3)	831	283	21.6%	25.4%
Sarah (2;3-5;1)	2453	640	29.2%	28.8%
Adam (2;3-4;10)	3729	780	32.3%	33.7%
Peter (1;4-2;10)	2873	480	40.4%	42.2%
Nina (1;11-3;11)	4542	660	46.7%	45.1%
Brown corpus	20650	4664	25.2%	26.5%

in (11) is highlighted in boldface, where we multiply the probability of the noun (p) with those of the determiners (f_1 and f_2). This assumes that their combinations are statistically independent, i.e., *not* lexically specific. If a sample of determiner-noun combinations has been generated by an abstract and productive rule, then the average diversity value calculated from (11) should closely match the empirical value, i.e., the percentage of the N nouns used with both determiners. As shown in Table 1, although the combinatorial diversity is quite low across both child and adult languages, it is statistically indistinguishable from the expected diversity under a rule where the combinations are fully productive.[8]

The method developed in (11), which has been independently replicated by other groups (e.g., Silvey and Christodoulopoulos 2016), has broader applicability, benefiting from the accuracy of Zipf's Law (10), or $r = 1/rH_N$. We can calculate the expected combinatorial diversity based only on the sample size (S) and types (N) appearing in the sample, without even knowing the identities of the words. In recent work (Goldin-Meadow and Yang 2017), the method has been applied to home signs, the gestural systems created by deaf children with properties akin to grammatical categories, morphology, sentence structures, and semantic relations found in spoken and sign languages (Goldin-Meadow and Mylander 1998). Quantitative analysis of predicate-argument constructions suggests that, despite the absence of an input model, home signs show the expected degree of combinatorial productivity. By contrast, the test has also been used to provide rigorous supporting evidence that Nim Chimpsky, the chimpanzee raised in an American Sign Language environment, never mastered the productive combination of signs (Terrace et al. 1979, Terrace 1987): Nim's combinatorial diversity such as "'give Nim" and "give me" falls far below the level expected of a productive rule (Yang 2013a).

I must be clear about what the determiner productivity study does and does not show. It demonstrates that, at least for one aspect of child language, combinatorial productivity is on full display from the earliest testable stage. Thus, the usage-based claim for a lexically specific grammar is not supported. Furthermore, it provides a methodological example of how to develop

[8] As Table 1 also makes clear, the actual value of diversity does not tell us anything about the underlying productivity of the grammar: unlike the claims in the usage-based literature, a higher diversity value (such as Nina) does not mean a "more" productive rule than a lower diversity value (such as the Brown corpus); see Pine et al. (2013) for a recent example of this fallacy. The formal analysis of the variation and how the empirical data are not predicted by usage-based learning models can be found in Yang (2013a).

statistically rigorous assessments of language data, a point to which I return in Section 5. But I do not suggest that all aspects of child language are productive from the get-go: see the remarks about English past tense earlier. Usage-based theories may be right about some specific cases of child language but a sweeping claim for a general, lexically specific, stage devoid of abstract rules and representations is false.

The very early acquisition of the determiner system raises important questions: *How* do children learn this particular rule of the English grammar so quickly? Here Zipf's Law poses a significant challenge. The sparsity of language entails that the caretaker's speech can never be fully saturated with linguistic combinations. Recall that the combinatorial diversity in caretaker speech is equally low as in child speech (Valian et al. 2009): that is, only a small fraction of nouns that can combine with both *a* and *the* will do so in the input. So how does the child generalize a property that holds for a small subset of words to all words — as they apparently do in Table 1? To answer this question, we need to confront the problem of learning by generalization: How do children acquire productive rules from lexical examples?

4 Productivity: $\theta_N = N/\ln N$

As Sapir remarked "all grammars leak" (1928, p. 38-39): the balancing act between rules and exceptions is one of the oldest problems in linguistics. While linguists can distinguish rules from exceptions by carrying out grammaticality judgments and fMRI, children's job must be considerably more difficult. Since rules and exceptions are defined in opposition of each other, children seem to face a chicken-and-egg problem, and it is one that needs to be resolved in a few short years, without supervision or feedback, all the while under the sparsity of data befitting Zipf's Law.

It is useful to distinguish two kinds of exceptions to rules. One kind can be called *positive exceptions*. The English past tense system is such an example: children receive overt evidence for the exceptions against the rule, by hearing irregular past tense forms that do not take "-ed". This is a familiar problem. A great many approaches, ranging from generative linguistics (e.g., Aronoff 1976, p. 36) to connectionist modeling (Marchman and Bates 1994) to hybrid models (Pinker 1999), share the same underlying intuition: a rule must "earn" its productivity, in the sense that it must somehow overcome the exceptions. It is frequently observed that a productive rule ought to be the one that covers the most diverse range of items: indeed, "statistical predominance" is traditionally the hallmark for linguistic productivity (e.g., Nida 1949, p. 14). However, (12) provides some illustrative problems from morphology and phonology, which suffice to show that the solution is not so simple:

(12) a. English past tense: A default rule is learned abruptly and results in overregularization, after a protracted stage of rote memorization (Marcus et al. 1992, Yang 2002).

b. English stress: The grammar of English stress (Chomsky and Halle 1968, Hayes 1982, Halle and Vergnaud 1987) is not trochaic with a list of lexical exceptions despite a vast majority of English words bearing stress on the first syllable (Cutler and Carter 1987, Legate and Yang 2013).

c. German noun plurals: A suffix ("-s") can be the productive default despite coverage of fewer nouns than any of its four competitors (Clahsen et al. 1992, Wiese 1996).

d. Russian gaps: That morphological categories need not and sometimes do not have a default, as illustrated by the missing inflections of certain Russian verbs in the 1st person singular non-past (Halle 1973). Such cases are far from rare (Baerman et al. 2010): the absence of the past tense for *undergo* and past participle for *stride* are the more familiar examples from English speakers (Pullum and Wilson 1977, Pinker 1999).

By contrast, *negative exceptions* in language learning seem more paradoxical. These are the cases where children must learn that a rule does *not* apply across to items that it could have. The well-researched English dative constructions illustrate the nature of the problem clearly:

(13) a. John gave the team a prize. John gave a prize to the team.
b. John assigned the students a textbook. John assigned a textbook to the students.
c. *John donated the museum the painting. John donated the painting to the museum.
d. John guaranteed the fans a victory. *John guaranteed a victory to the fans.

The verbs *give* and *promise* can freely alternate between the double object construction and the *to*-dative construction. However, semantically very similar verbs such as *donate* can only appear in the *to*-dative construction, and *guarantee* is exactly the opposite. Because children do not receive negative evidence, how do they learn what not to say in their language? Here a lexically conservative approach cannot work: the productivity of these constructions is evident in child language (Gropen et al. 1989, Conwell and Demuth 2007) and can also be observed when they are extended to novel verbs with appropriate semantic properties: when the verb *text* appeared, its double object form was instantly available as in *I texted them the score*.

Problems such as the acquisition of the dative constructions once dominated the learnability research in the generative tradition (Baker 1979, Berwick 1985, Fodor and Crain 1987, Pinker 1989). In recent years, they have become a major focus of usage-based theories under the tenet of entrenchment (Tomasello 2003, Bybee 2006, Ambridge et al. 2008) or preemption (Stefanowitsch 2008, Boyd and Goldberg 2011): both are a form of indirect negative evidence which takes the absence of evidence as evidence of absence (Pinker 1989). According to a recent formulation (Ibbotson and Tomasello 2016), "if children hear quite often *She donated some books to the library*, then this usage preempts the temptation to say *She donated the library some books*." But use of indirect negative evidence is unproblematic (Pinker 1989). After all, the absence of evidence is *not* evidence of absence; see Yang (2015b, 2017) for an assessment using realistic child-directed language corpus statistics. Most glaringly, such approaches fail to account for one of the most robustly attested errors in children's dative constructions, namely expressions such as "I said her no", which have been studied by some of the most prominent usage-based language researchers (Bowerman 1982, Bowerman and Croft 2008). The communication verb *say*, of course, is always used in the *to*-dative construction, and is among the most frequently used verbs in English — yet this (deeply) entrenched form fails to preempt the double object alternation.

The Tolerance Principle and its corollary the Sufficiency Principle (Yang 2016) provide a unified solution for the problem of rules and exceptions. I will not review the empirical motivation for their development but will simply state:

(14) a. *Tolerance Principle*
Suppose a rule R is applicable to N items in a learner's vocabulary, of which e items do not follow R and are thus exceptions. The necessary and sufficient condition for

the productivity of R is:

$$e \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N}$$

b. *Sufficiency Principle*
Suppose a rule R is applicable to N items in a learner's vocabulary, of which M follow R and no information is available about the remaining $(N - M)$ items. The necessary and sufficient condition for the productivity of R is:

$$(N - M) \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N}$$

The unifying theme of the two principle lies in the quantity of positive evidence — $N - \theta_N$ in both cases — that is necessary to support the productivity of a rule. To understand the intuition behind the rationale, consider an analogous case in a non-linguistic domain. Suppose you have encountered 10 new species off a remote island, of which 8 share a certain property (e.g., phosphorescence). Even though you may not have any information about the other two, or maybe even if the other two are known *not* to have the property, it seems reasonable to form a generalization about the entire class and extend it to the 11th species. By contrast, if the property only holds for 2 of the 10 examples, it seems wise not to rush to any general conclusion. The Tolerance Principle provides a precise weight of evidence, in the form of $\theta_N = N/\ln N$, that warrants productive generalizations.

The mechanism of learning under the Tolerance Principle is schematically illustrated in Figure 1.

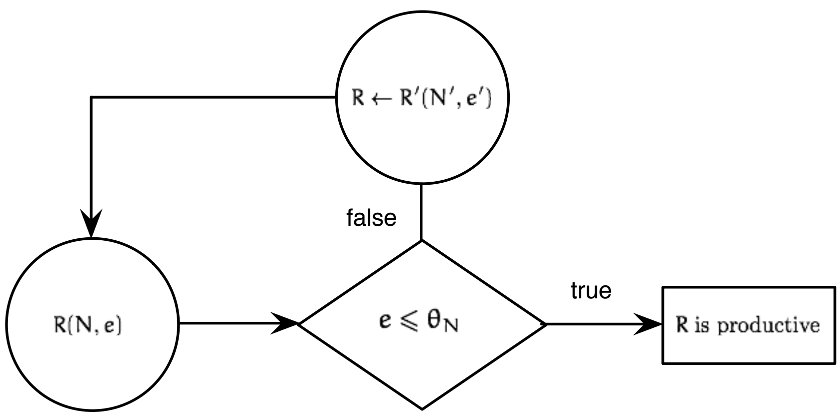

Figure 1: Tolerance Principle guides the search for productive rules in language learning.

Language learning is a search for productive generalizations that is best characterized as abductive learning (Chomsky 1968, p. 80). Children construct a rule R from the input data guided by linguistic and cognitive constraints and evaluate its productivity according to the associated numerical values (N and e). The rule is deemed productive if the positive evidence is sufficiently high; otherwise learners formulate a revised rule (R') to obtain a new set of values ($N'$and e') and the Tolerance Principle is applied recursively. Thus, the quantitative accumulation of exceptions can lead to the qualitative change in the productivity of rules. If no plausible rule can be found that meets the criterion for productivity, children will lexically memorize each instance that follows R and no productive generalization will be established.

Table 2 provides some sample values of N and the associate threshold values θ_N.

Table 2: The maximum number of exceptions for a productive rule over N items.

N	θ_N	%
10	4	40.0
20	7	35.0
50	13	26.0
100	22	22.0
200	38	19.0
500	80	16.0
1,000	145	14.5
5,000	587	11.7

These thresholds for productivity under the Tolerance Principle are significantly lower than a naïve "majority rule", which has many interesting implications for language acquisition. In particular, the Tolerance Principle asserts that a smaller vocabulary (i.e., smaller values of N) can tolerate a higher percentage of exceptions: all else being equal, productive rules are *easier* to detect for learners who have access to *less* input data. I return to this important theme in the conclusion.

Here I briefly summarize the application of the Tolerance and Sufficiency Principle to two of the most intensively studied problems in language acquisition: the past tense and the dative constructions in English.

In the case of English past tense, a distributional learning of induction must be used to identify the rules of verbal inflection. Many proposals of inductive learning, from many diverse fields (Chomsky 1955, Osherson and Smith 1981, Mitchell 1982, Cohen 1995, Yip and Sussman 1997, e.g.,), are applicable. These models typically operate by forming generalizations over exemplars in some suitable representation. For example, suppose two good baseball hitters can be described with feature bundles [+red cap, +black shirt, +long socks] and [+red cap, +black shirt, +short socks]. The rule "[+red cap, +black shirt] → good hitter" will follow, as the shared features (cap, shirt) are retained and the conflicting feature (sock) is neutralized. This method is thus very capable of identifying the "-ed" suffix as one applicable without all phonological restrictions on the stem: the verbs that take "-ed" are phonologically very diverse, and no restrictions will be identified (Yip and Sussman 1997). Thus, the productivity of "-ed" will be determined by the total number of verbs (N) and the number of irregular exceptions (e) in the learner's vocabulary. The same consideration must also apply to the patterns that govern irregular verbs. For instance, the irregular verbs *bring*, *buy*, *catch*, *fight*, *seek*, *teach*, and *think* all undergo a stem change replacing the rime with [ɔt]. This rule also has no restriction on the stem, because the participating verbs are phonologically very diverse. But it is easy to see that the rule"rime →ɔt" will fare terribly: the seven positive members are easily swamped by hundreds of negative examples that do not change the rime to [ɔt], far exceeding the tolerance threshold. As a result, the rule is not productive and will be lexicalized — to these seven verbs. Other irregular patterns can be analyzed similarly: as shown elsewhere (Yang 2016, chap. 4), all rules except the regular "add *-d*" will be assessed as unproductive, accounting for the near-total absence of over-*ir*regularization errors in child English (Xu and Pinker 1995; see Lignos and Yang 2016 for a cross-linguistic review of similar findings.)

Following the same logic, we can see that the emergence of the "add *-d*" rule will require a long gestation period. Although children can quickly induce its structural description — perhaps using no more than a few dozen verbs (e.g. Yip and Sussman 1997) — irregulars are likely over-represented in children's early vocabulary. For instance, in a corpus of 5 million words of child-directed English drawn from the CHILDES (MacWhinney 2000) database, 76 of the 200 most frequent verbs in past tense are irregular. As θ_{200} is only 37, children with a small vocabulary are unlikely to establish the productivity of "-ed", despite the fact that it may be by far the statistically dominant rule. For very young children, then, verbs marked with "-ed" are in effect on part with irregulars: they are item-based and lexically memorized.

Telltale evidence for productivity comes from the first attested overregularization errors. When longitudinal records are available, the Tolerance Principle can account for the particular juncture at which productivity emerges. As noted earlier, Adam produced his first recorded overregularization error at 2;11 ("What dat feeled like?"). By then, Adam must have acquired a sufficiently large number of regular verbs to overwhelm the irregulars. In Adam's transcripts leading up to 2;11, he used $N = 300$ unique verbs in all, of which $e = 57$ are irregular. This is quite close to the predicted $\theta_{300} = 53$, and the discrepancy may be due to the underrepresentation of his regular verbs, which will be less frequent and thus more likely to be left out of a sample. Thus, Adam acquired a productive "-ed" only after he acquired a "super" majority of regular verbs, consistent with the Tolerance Principle.

The acquisition of the dative constructions is more complex as it involves children overgeneralizing before retreating: young children say "I said her no" but older children and adults do not. But the same procedure of constructing and evaluating generalizations applies here as well. Here I will briefly review the acquisition of the double object construction as the *to*-dative construction can be handled in a similar fashion (Yang 2016, chapter 6). In a five-million-word corpus of child-directed English data, roughly corresponding to one year worth of input, we can find a total of 42 verbs used in the double object construction. Of these, 38 have a very clear semantics of "caused possession", which we assume is identifiable if the learner is equipped with a suitable set of conceptual and semantic primitives (e.g., Pinker 1989, Grimshaw 1990, Jackendoff 1990). The four exceptions, well below the threshold $\theta_{42} = 11$, do not convey caused possession: all are performative verbs (*call*, *consider*, *name*, and *pronounce*, e.g., *I called him a liar*). Thus, the semantic condition necessary for double object construction (Gropen et al. 1989, Pinker 1989, Levin 1993, Goldberg 1995, Pesetsky 1995, Krifka 1999) need not be stated as a UG primitive but can acquired from the language-specific data. The following hypothesis, then, can be formulated under the guidance of the Tolerance Principle:

(15) If a verb appears in the double object construction, then it will have the semantics of caused possession.

At this point, the child may consider the converse of (15), in trying to establish the validity of caused possession as a *sufficient* condition for the double object condition. This amounts to testing if the entire set of caused-possession verbs (N) in the child-directed corpus can be productively used in the double object construction. According to the Sufficiency Principle, this is warranted only if the $M = 38$ items, which are actually attested in the construction, constitute a sufficiently large subset of N.

In the present case, the input corpus contains an additional 11 verbs that belong to the semantic class in (15). But these did not appear in the double object construction:

(16) address, deliver, describe, explain, introduce, return, transport, ship, mention, report, say

This is an interesting list. For some of the items in (16), e.g., *deliver* and *say*, the double object construction is ungrammatical: **John delivered the kids a pizza*, **John said Bill something mean*. But the verb *ship* does allow the double object construction — *John shipped Bill his purchase* — it just was not in the corpus because the caretakers opted for the *to*-dative form instead. Of course, the child does not know *why* the 11 verbs in (16) fail to show. The statistical distribution of language makes it difficult, if not impossible, to distinguish impossible forms from possible but unattested forms, which is the Achilles heel of indirect negative evidence and its entrechment or preemption variant (Yang 2015b, 2017). Nevertheless, $M = 38$ constitutes a sufficiently large subset of $N = 38 + 11 = 49$, and thus the following generalization ensues:

(17) If a verb has the semantics of caused possession, then it can appear in the double object construction.

This immediately accounts for the overgeneralization errors such as *I said her no.* A search of child speech data in CHILDES also yields errors such as *I delivered you a lot of pizzas* (3;8), a verb not permissible in the construction in the adult language, thereby supporting the productivity of (17). This also accounts for the experimental evidence that children as young as 3;0 have productive usage of the dative constructions upon learning a novel verb with the appropriate semantic properties (e.g., Gropen et al. 1989, Conwell and Demuth 2007).

The retreat from overgeneralization straightforwardly follows the Tolerance/Sufficiency Principle but to do so requires the child to expand their vocabulary. Highly frequent verbs, which populate our child-directed corpus and are among those learned earlier by children, heavily favor the productive use of the rule in (17). I combined several corpora to approximate the vocabulary of ditransitive verbs likely known to most English speakers (Yang 2016, p. 208), and the results are given in Table 3.

Table 3 shows that a child with a very limited vocabulary is bound to conjecture (17) as a productive rule, namely, *all* caused-possession verbs may appear in the double object construction. But as they learn more verbs, the proportion of those used in the construction will continue to drop, eventually below the sufficiency threshold. The child can thus successfully retreat, without the problematic use of indirect negative evidence.

It is interesting to probe the properties of the verbs in Table 3 further, which provides a learning-theoretic account for many regularities in the double object construction and its acquisition (Mazurkewich and White 1984, Gropen et al. 1989, Bley-Vroman and Yoshinaga 1992, Inagaki 1997); see Yang and Montrul (2017) for a review. For example, 50 of the 92 verbs in Table 3 are monosyllabic, of which 42 allow double objects. By comparison, only 10 of the 42 polysyllabic verb can participate in the construction. Thus, when given a novel verb that describes, say, the movement from an object initiated by one individual to another, both children and adults

Table 3: Caused-possession verbs and their availability in the double object construction (adapted from Yang 2016).

top	yes	no	θ_N	productive?
10	9	1	4	Yes
20	17	3	7	Yes
30	26	4	9	Yes
40	30	10	11	Yes
50	34	16	13	No
60	39	21	15	No
70	43	27	16	No
80	46	24	18	No
92	50	42	20	No

are more inclined to accept a short novel verb (e.g., *pell*) in the double object construction than a long one (e.g., *orgulate*). These tendencies are simply the consequences of distributional learning, rather than structural constraints as proposed in some theoretical literature (see Harley and Miyagawa 2016 for a recent review). Similarly, while the entire class of caused-possession verbs cannot categorically participate in the double object construction, productivity can be found in semantic subclasses assuming that such classes can be constructed by language learners (Pinker 1989, Yang 2016): recall that productivity is more likely with smaller values of N.

In sum, the Tolerance Principle appears to embody what some usage-based researchers envision as the key solution to the problem of language learning: "a single mechanism responsible both for generalization, and for restricting these generalizations to items with particular semantic, pragmatic, phonological (and no doubt other) properties (Ambridge and Lieven 2011, p. 267)". Its simplicity yields sharp behavioral predictions, with interesting implications for the apparent differences between the outcome of child and adult language acquisition.

5 Acquisition in Adulthood

The questions addressed by the three equations arise for any problem of language acquisition, by children and adults alike. Here I offer some comments and speculations on how these methods may apply to adult language acquisition.

5.1 Selection and UG Access

Any learning model must consist of precise statements about the hypothesis space that the learner entertains – the initial state – and the mechanisms of learning from data (Chomsky 1965, Yang 2002). A successful model is the combination of the initial stage and the learning mechanisms that explains the specific patterns and changes during language development. These questions are often debated first language acquisition but they arise for second language as well, often in the form of the role concerning UG (e.g., Clahsen and Muysken 1986, Cook and Newson 2014, Epstein et al. 1996, Schwartz and Sprouse 1996, White 2003, Rothman and Slabakova 2017). Adult language learners also have an initial state, which may be partly dependent on their first language, and they also need to integrate data and experience into their linguistic knowledge. I would like to suggest that the variational learning model provides a useful perspective on these theoretical issues.

It is now evident that language, and language learning, obey certain structural constraints: not all formal language-like systems are natural or naturally learnable for by otherwise capable children and adults alike (Crain and Nakayama 1987, Smith and Tsimpli 1995, Newport and Aslin 2004, Tettamanti et al. 2004, Friederici 2017). What remains controversial is the nature of such constraints: what they are and whether they reflect domain specific restrictions on language or follow from more general cognitive principles. If it can be established that the variational model is employed by child and adult language learners alike, essentially by holding the learning mechanism component of the model constant, then the nature of the initial state may be fruitfully investigated.

The variational learning model is very likely implicated in language acquisition throughout the lifespan. In fact, it would be highly surprising if it were *not*, given its ubiquity in probabilistic learning and decision making across domains and species. And there appear to be strong continuities between children and adults in at least certain domains of language. For word learning, the parallels between children and adults are very strong (Markson and Bloom 1997, Bloom 2000), and models of lexical acquisition are frequently tested on adult participants (Yu and Smith 2007, Stevens et al. 2017, e.g.). Furthermore, the model converges to a statistical combination of multiple hypotheses in linguistically heterogeneous environments as in the case of bilingualism and language change (Yang 2000), which further supports its broad applicability. Although there is no denying that adult, non-native grammatical acquisition is different from that of children in path and ultimate attainment, the differences do not have to — and like do not — relate to differences in underlying (cognitive/linguistic) mechanisms availble to each. Let us assume, then, that variational learning is indeed used by language learners at all stages of development and explore how it helps determine the role of UG in adult language acquisition.

Consider, again, the obligatory use of grammatical subjects in languages such as English. As reviewed earlier, L1 acquisition can be modeled as a competition among the options delimited by UG: the telltale evidence is frequent occurrence of null subjects in adjunct Wh questions (3) and near absence in argument Wh questions (4), both characteristic of a topic-drop grammar. How would an adult learner acquire a second language which uses the grammatical subject in a different way from their their first language? Here it is important to briefly review the distributional evidence that can disambiguate the three broad classes of grammars from the perspective of the language learner.

As reviewed earlier, obligatory-subject languages such as English can only be uniquely identified by non-referential expletive subjects, whose rarity in the child-directed input, and conversational speech quite generally, accounts for the extended subject drop stage during language acquisition. Consider now the properties of the topic-drop and pro-drop grammars, which I will illustrate with examples from Chinese and Italian respectively.

In languages such as Chinese, the omission of the subject is due to discourse conditions under which the subject is the topic that is saliently retrievable (Huang 1984). An associated property is the possibility of object drop, when the object happens to the topic (*e* below):

(18) Top_e [Zhangsan kanjian-le *e*].
Top_e [Zhangsan saw-ASP *e*]
'Zhangsan saw *e*'

This option is unavailable to pro-drop languages such as Italian, which licenses subject omission on the basis of subject-verb agreement morphology. Thus, null objects serve as unambiguous evidence for the topic-drop grammar.

What constitutes unambiguous evidence for the pro-drop option? There is a rich literature on this topic (Jaeggli and Safir 1989, Biberauer et al. 2010) but it has been difficult to uncover precisely what makes it available. While "rich" agreement morphology does appear a necessary condition without which the identity of the omitted pronoun cannot be recovered, it is by no means a sufficient one: languages such as Icelandic have richly agreement morphology but do not allow pro-drop and indeed make use of expletive subjects. From a language acquisition perspective, however, we only need to identify the distributional evidence that sets Italian apart from its competitors (i.e., Chinese and English). Under the variational learning scheme, every instance of omitted subject for the Italian learner will obviously penalize the English grammar which requires the grammatical subject. But the Chinese option is more difficult to rule out. Here the argument vs. adjunct asymmetry in topic-drop languages, which was discussed extensively in Section 2, becomes relevant. Because pro-drop is licensed by verbal agreement, its availability should not be sensitive to the discourse configuration of topics at all. In other words, Italian should allow null subjects regardless the fronted Wh-word is an argument (19a and 19b) or an adjunct (19c), in contrast to the Brazilian Portuguese examples reported in (7) where the null subject is licensed by the topic-drop option.

(19) a. Chi_2 e_1 ha baciato t_2?
Who_2 has_{3SGM} kissed t_2?
'Who has he kissed?'

b. Chi_3 e_1 credi che e_2 ami t_3?
Who_3 e_1 $think_{2SG}$ that e_2 loves-3SGF t_3?
'Who do you think she loves?'

c. $Dove_2$ hai e_1 visto Maria t_2?
$Where_2$ $have_{2SG}$ e_1 seen Maria t_2?
'Where have you seen Maria?'

Thus, the critical evidence that uniquely identifies the three types of grammars is summarized in (20), along with their frequencies in child-directed input:

(20) a. Chinese: Null objects (11.6%; Wang et al. 1992)
b. Italian: Null subjects in object wh-questions (10%; Yang 2002)
c. English: Non-referential expletive subjects (1.2%; Yang 2002)

The amount of unambiguous evidence for the Chinese and Italian type grammars is quite high. It is in fact, higher than the unambiguous evidence – about 7% (Yang 2002) – for the correct placement of the finite verb, which children acquire very early and are essentially error-free (Pierce 1992). Thus, we predict very early acquisition of topic drop and pro drop. Indeed, studies of the acquisition of Italian (Valian 1991), Catalan/Spanish (Grinstead 2000), Chinese (Wang et al. 1992), Korean (Kim 2000), etc. have consistently found that children reach adult-level use of subjects around age 2, considerably earlier than the consistent use of grammatical subjects by English-learning children which typically takes place by age 3 or later.

Note that these findings about child language only follow if the initial stage consists of hypotheses, and their associated properties, laid out in (20). Because virtually all grammatical subjects are also thematic, the protracted null subject stage in child English must be linked to the low frequency of expletive subjects, which are the only type of input that distinguishes the competing options in UG. If the initial state consists of surface rules such as $S \rightarrow$ NP VP or their probabilistic instantiations, children would be predicted to quickly converge to the target grammar. Furthermore, the fact that not all languages are learned at the same time runs counter of the performance-based accounts according to which the child's grammar is identical to the adult's. If English-learning children's grammar were adult-like and the null subjects are due to performance constraints that prevent the production of fully formed sentences (Pinker 1984, Bloom 1990, Valian 1990), such constraints must apply equally to children learning pro-drop and topic-drop languages: surely children's maturational schedule is not language dependent. But that would result in *more* omitted subjects than what young children acquiring these languages actually produce — at adult level.

With this background on first language acquisition, we can turn to the question of adult second language acquisition. In fact, the published literature already points to strong parallels between children and adults: the acquisition of pro drop and topic drop appears "easier" than the acquisition of the obligatory subject use. For instance, Phinney (1987)'s classic study finds that while advanced L1 Spanish learners of English do not consistently use expletive subjects, even beginning L1 English learners of Spanish show excellent command of pro drop. Later studies confirmed the excellent command of the pro-drop option by L2 learners (e.g., Pérez-Leroux and Glass 1999). The topic-drop option is likewise easily acquired. For example, Kanno (1997) finds that L1 English learners of Japanese have close-to-native command of null subjects across a number of syntactic and discourse contexts. By contrast, even near-native L2 learners of English fail to consistently use the expletive subject (Judy 2011), the true hallmark of the obligatory subject grammar. The variational learning model provides a straightforward account for these cross-linguistic findings. As shown in (20), the advantage of the topic/pro-drop grammars over the obligatory subject grammar is afforded by the more abundant disambiguating evidence in the input language, because variational learning is gradual, probabilistic, and quantity sensitive.

The variational interpretation of the second language acquisition of the subject can help make an even stronger claim about the role of UG. If adult language acquisition mirrors child language acquisition, then one must conclude that the parametric options of UG are available to children and adults alike, which amounts to a very strong form of UG access hypothesis most similar to the position of Epstein et al. (1996). To establish this claim requires the same kind of evidence from child language acquisition under the variational model: we need to find the footprints of non-target yet UG-consistent grammatical hypotheses. We thus turn again to the argument vs. adjunct asymmetry: null subjects are possible in adjunct Wh-questions but not possible in argument Wh-questions.

Such an asymmetry, if found, may not be very surprising for a learner whose L1 is a topic-drop grammar, but would provide strong evidence for the full accessibility of UG for learners whose L1 is a pro-drop language, which licenses null subjects by agreement and does not show the distributional restrictions. I am not aware of any study of adult acquisition of the English grammatical subject that specifically targets these distributional properties. And there is preciously little data on second language acquisition in the public domain. In what follows, I can only offer a preliminary analysis based on a corpus provided by Klein and Perdue (Perdue 1993, available at talkbank.org).

There are four adult English learners whose first language is Italian, which provide the suitable opportunity to examine the distribution of null subjects in their Wh-questions. I extracted all of their Wh-fronted questions. There are 35 (object) argument questions, only one missing the subject ("what doing in here"). There are 72 adjunct questions, with 16 missing the subject; some examples are given below:

(21) where is?
why shouldn't?
how much cost?
why no have appuntamento (appointment) in the evening?
when come in the school.

The sample size is small but there is a statistically significant difference ($p = 0.01$) between the null subject rates for argument and adjunct Wh-questions. The predictions here are straightforward and can be easily verified in future research with larger datasets. The findings are suggestive: adult learners have access to a UG option – topic drop á la Chinese – that is neither in their first (Italian) or second (English) language.

The continuity between child and adult language, if true, points to potential intervention strategies. Although second language learners of English are frequently reminded of the fact that English requires the subject, only expletive subjects truly serve the purpose of driving the learner toward the target form, much like the acquisition of English by young children. Because expletive subjects are relatively infrequent in language use, amplifying the amount of such input may result in accelerated acquisition of the subject, similar to the improvement in L2 speech perception and production under targeted input (e.g., Bradlow et al. 1999). In addition, different intervention strategies may be designed to probe the nature of adult learners' initial state. For example, if adult learners more rapidly acquire the obligatory use of English subjects on the basis of expletive subjects rather than merely lexical or pronominal subjects as in the rule $S \rightarrow$ NP VP, it would provide further evidence for the continuity between child and adult language acquisition and amplify the role of UG. The variational learning model, due to its generality across individuals and domains, provides a level playing field for different theoretical approaches to language acquisition.

5.2 Rules and Cognitive Capacity

The statistical test for assessing grammatical productivity can be ported straightforwardly to the study of second language: Do L2 speakers go through a stage where the grammar is lexically specific and lacks abstract generalization? Again, these questions are empirical ones and are not inherent or unique features of any specific theoretical approach. However, claims of usage/item-based learning from first language study (e.g., Tomasello 2000, 2003) appear to have been imported wholesale into second language research (Ellis and Larsen-Freeman 2009): low frequency and diversity of syntactic combinations are taken as evidence for lexically based prototypes and exemplars and the absence of a fully abstract system. As discussed in Section 3, these claims cannot be taken at face unless they are evaluated against rigorously formulated statistical hypotheses, including the null hypothesis that language – of children and adults – is in fact fully productive.

In this section, I apply the methods developed by assessing child language to adult language. The inferential problem is the same: given a linguistic corpus, what is the nature of the underlying mechanism that generates the production? Is it a fully productive system, or is it lexically specific? The case study focuses on seven adult learners of English studied by Klein and Perdue (Perdue 1993, available at talkbank.org). The methods are very simple and can be automated with simple natural language processing tools (see Yang 2013a and Silvey and Christodoulopoulos 2016for details).

Table 4: Empirical and expected combinatorial diversity in L2 English (data from talkbank)

Subject	L1	Sample size (S)	Types (N)	Empirical	Expected
Andrea	Italian	355	171	4.1%	9.8%
Lavinia	Italian	822	295	19.3%	22.6%
Santo	Italian	398	193	3.1%	9.6%
Vito	Italian	314	139	6.5%	11.0%
Ravinder	Punjabi	121	75	5.3%	8.6%
Jainail	Punjabi	283	148	8.1%	9.3%
Madan	Punjabi	237	102	3.9%	11.7%

The L2 learners' use of determiner-noun combinations is significantly below the diversity level expected under a fully productive grammatical rule ($p < 0.001$). Note however the samples in Table 4 are considerably smaller than the datasets from first language acquisition and the results must be taken with a grain of salt.[9] But Table 4 does appear to reveal a usage-based stage of L2 acquisition in which learners have not mastered a simple grammatical rule, which L1 learners command with ease at a very early age (Table 1).

To understand the discrepancies between L1 and L2 acquisition, we must first address how young children acquire the determiner-noun rule in the first place. Because the rule is language specific, it must be learned distributionally from the input data. An examination of the child-directed input data turns up an interesting puzzle — one which may shed light on why adult language learners struggle with grammatical rules.

Consider Adam, for the last time. He produced 3,729 determiner-noun combinations in his speech with 780 distinct nouns. Of these, only 32.2% appeared with both determiners, which is similar to the expected value of 33.7%; see Table 1. Adam's mother, whose speech was also transcribed in the same corpus, produced a diversity measure of 30.3% out of 914 nouns. Even among the 469 nouns used at least twice, which provided opportunities to be used with both determiners, only over half (260) did so. To appreciate the logic of learning, consider a baseball

[9] This is also a plea for wider dissemination of L2 data for public use.

analogy: the interchangeablity of *a* and *the* for a noun can be viewed a batter's ability to switch hit (i.e., batting both left- and right-handed). Suppose a scout has been sent to evaluate a team of players, about a third of whom switch-hitted in the batting practice. It would seem crazy to conclude *all* players in the squad are switch-hitters, but that is apparently what Adam did: he generalized the interchangeability of *a* and *the* from a third of nouns to all nouns. Such inductive leap seems absurd. It is certainly not sanctioned by the Tolerance/Sufficiency Principle, which asserts that a rule — the interchangeability of *a* and *the* — can be extended to a class of words only if the rule holds for an overwhelming majority of the words.

A promising, and perhaps the only, way out of this dilemma is to make use of a key property of the Tolerance Principle: rule learning is easier, and more tolerant of exceptions, when the learner has a smaller set of items in their vocabulary (Table 2). The developmental literature offers the idea of "less is more" (Newport 1990, Elman 1993, Kareev 1995, Cochran et al. 1999): the maturational constraints place a limit on the processing capacity of young children, which may turn out to be beneficial for language acquisition. If children's vocabulary is smaller, the odds of acquiring productive rules improve considerably.

Consider again the determiner-noun combinations produced by Adam's mother. Only 277 out of the 914 nouns are used with both *a* and *the*, which is nowhere near the requisite threshold for generalization ($\theta_{914} = 134$). But if Adam were only to learn from the 50 most frequent nouns, he would notice that almost all of them — 43 to be precise — are paired with both determiners. On this much smaller subset of data where $N = 50$, there is sufficient evidence for generalization: the 7 nouns that appear exclusively with only one determiner are below the tolerance threshold $\theta_{50} = 12$. For the top $N = 100$ nouns, 83 are paired with both determiners: the 17 loners are again below the tolerance threshold $\theta_{100} = 23$. At the time when children acquire the productive rule for determiners, their vocabulary size does not exceed a few hundred (Fenson et al. 1994, Hart and Risley 1995). It is thus highly like that they have acquired the rule on a very small set of high frequency nouns, almost all of which will show interchangeability with both determiners; the rest is just noise.

This line of thinking naturally leads us to speculate why adults tend to be worse at language learning than children, when they are better at pretty much everything else. There are of course many differences between children and adults but I would put forward a simple but bold possibility. By the virtue of having greater cognitive capacities, adult language learners may have developed a lexicon too large for their own good.

Table 5 gives the token/type ratio for the words used by the L1 and L2 learners, whose productivity measures have been given in Table 1 and 4. The ratio provides a rough measure of the language user's vocabulary. A ratio of X means that, on average, the speaker produces a new word type every X words; thus, a small token/type ratio is an indication of a larger vocabulary, a long-standing practice in language research (Miller 1981, Huttenlocher et al. 1991). It is evident that the adults have considerably larger vocabularies than the children.[10]

I suggest that the L2 learners's apparent inability to use a simple rule fully productively is because they know *too many* words. Young children have no choice but to learn from a small set of high frequency words for which the evidence for productive rules is sufficiently strong. Again, toddlers' vocabulary size has been estimated to be no more than just over a thousand

[10] The data is scanty but even at the early testing sessions shortly after their arrival (Perdue 1993), the adult learners' token/type ratios are still considerably lower than young children's.

Table 5: Vocabulary size estimate and productivity across lifespan.

Subject	Token	Type	Token/Type
Adam	140,793	3,811	36.94
Eve	27,147	1,582	17.16
Naomi	35,459	2,274	15.59
Nina	87,933	2,553	34.44
Peter	62,006	1,936	32.03
Sarah	73,951	3,658	20.22
Andrea	8,097	1,013	7.99
Lavinia	16,941	1,673	10.13
Santo	10,705	1,284	8.34
Vito	9,344	1,296	7.21
Ravinder	9,980	954	10.46
Jainail	10,017	913	10.97
Madan	8,672	909	9.54

(Huttenlocher et al. 1991, Hart and Risley 1995), yet children have perfect command of a productive system including the determiner rule reviewed earlier and many aspects of morphology and syntax (see Guasti 2004 for a review). Thus, a small vocabulary must be sufficient and, if I am correct, necessary, for the acquisition of the essential components of language. Adults, by contrast, are in fact handicapped by their considerably larger vocabulary size due to their mature cognitive capacities. A larger value of N has the inadvertent consequence of raising the threshold for productivity, thereby making rule learning much more difficult. Perhaps the smart move for adult learners is to hold them back, rather than intensive vocabulary drills or the memorization of conjugation tables.

Finally, a word about the application of the Tolerance Principle in adult language acquisition. The Principle is a method by which the learner evaluates potentially productive hypotheses about language. That these hypotheses are abstractions from input data, and to be evaluated against (further) input data entails that a quantitative measure of the input data is absolutely crucial. At this point, it is worth emphasizing that the values of N and e pertain to the vocabulary composition of specific learners, which necessarily vary on an individual basis. Thus, some learners may discover productive rules before others, and there is also the possibility that the terminal state of individuals' grammars varies with respect to the productivity of certain rules; see Yang (2016, Ch. 4) for case studies. This poses some practical challenges for the language acquisition researcher, as getting accurate measures of an individual learner's vocabulary is quite difficult: the analysis of how "Adam" developed his past tense rule in Section 4 was enabled by the unusually large sample of his speech and can only be regarded as an approximation. And it is virtually impossible to obtain a complete record of any individual learners' input data. Fortunately for first language acquisition, the problem is partly ameliorated by the striking uniformity in the grammar attained by individuals in a same speech community, down to the quantitative details of language variation and change (Labov 1972, 2007). Thus, one can speak of a "typical" child, whose vocabulary can be approximated by pooling a large amount of child-directed input data and using word frequency thresholds (Nagy and Anderson 1984) to gauge the learner's vocabulary size at various stages of language development – as illustrated in the study of the dative constructions (Section 4).

Obtaining precise quantitative measures of the input data is obviously much hard for adult language learners. But it may be easier to obtain more accurate offline vocabulary measurements for adults, who are likely to be more cooperative than toddlers. Ultimately it is the individual's *internalized* vocabulary that determiners the productivity of rules. Furthermore, it is possible to devise experiments where one can have precise control over the individual's vocabulary with the use of artificial language.

For example, in Schuler et al. (2016)'s study, young children learn nine novel nouns. In one condition, five nouns share a plural suffix ("regulars"), and the other four are idiosyncratic ("irregulars"). In the other condition, the mixture is three regulars and six irregulars. The choices of 5/4 and 3/6 are by design: the Tolerance Principle predicts the productive extension of the regular suffix in the 5/4 condition because four exceptions are below the threshold ($\theta_9 = 4.2$) but there is no generalization in the 3/6 conditions. In the latter case, despite the statistical dominance of the regular suffix, the six exceptions exceed the threshold. When presented on additional novel items in a Wug-like test, almost all children in the 5/4 condition generalized in a process akin to the productive use of English -ed, and none in the 3/6 condition did, much like speakers trapped in morphological gaps (12); see Halle (1973) and Baerman et al. (2010).

A follow-up study (Schuler 2017) was able to test the Tolerance Principle at the individual level, which is perhaps more relevant to the study of adult language acquisition where the individual variation would be considerably greater than child language acquisition. Children also received nine nouns in the input but due to manipulations of exposure frequency, not all nine nouns were learned by all children. For example, there are children who learned eight nouns (thus $N = 8$). But some of these children happened to learn five regulars and three irregulars, whereas others learned four regulars and four irregulars. Strikingly, the Tolerance Principle makes accurate predications about the individual's grammar. Because $\theta_8 = 3.8$, the first group of children were predicted to have a productive rule and the second group of children were predicted not to. Despite the identical vocabulary size, the regular-irregular composition of the lexicons is different across learners, and the predictions are again nearly categorically confirmed. Taken together, these corpus and experimental studies suggest that the Tolerance Principle may be directly testable on adult language learners, yielding precise and individual-level predictions; see Yang and Montrul (2017) for specific suggestions concerning the dative constructions.

5.3 Final Remarks

In my view, precise quantitative models are the inevitable next step in the study of language acquisition as the understanding of linguistic structures and learning mechanisms reaches an appropriate level of depth, especially in the light of evolution (Chomsky 2001, 2005, Berwick and Chomsky 2016, Yang et al. 2017). But such models must be simple, even at the risk of being too simple, to be useful. Recall the rise of modern linguistics and cognitive science was enabled by the formal methods introduced by generative grammar: abstraction and idealization over complex phenomena, followed by deductive analysis of nontrivial depth. The worst one could wish for are "interactionist" and "emergentist" proposals where all conceivable factors are thrown into a stew just so no correlation could ever be missed. These approaches do not enlighten but only obfuscate; see Yang (2015a) for a critical assessment of a recent proposal centering around the role of frequency in language acquisition.

Needless to say, more research, and especially more data, will be needed to further test the three equations discussed here and to verify their applicability to adult language acquisition. All the same, I hope to have conveyed the importance of formal methods to the study of language acquisition. The equations may turn out to be wrong. But one of the most appealing aspects about language is that it is tractable, and even mechanical. We can make progress even by making mistakes so long as the mistakes are precisely formulated.

References

Ambridge, B. and Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches.* Cambridge University Press, Cambridge.

Ambridge, B., Pine, J. M., Rowland, C. F., and Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135.

Aronoff, M. (1976). *Word formation in generative grammar.* MIT Press, Cambridge, MA.

Baerman, M., Corbett, G. G., and Brown, D., editors (2010). *Defective paradigms: Missing forms and what they tell us.* Oxford University Press, Oxford.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.

Bates, E. and Elman, J. (1996). Learning rediscovered. *Science*, 274(5294):1849.

Berwick, R. (1985). *The acquisition of syntactic knowledge.* MIT Press, Cambridge, MA.

Berwick, R. C. and Chomsky, N. (2016). *Why only us: Language and evolution.* MIT Press, Cambridge, MA.

Biberauer, T., Holmberg, A., Roberts, I., and Sheehan, M., editors (2010). *Parametric variation: Null subjects in minimalist theory.* Cambridge University Press, Cambridge.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning. *Linguistic perspectives on second language acquisition*, 4:1–68.

Bley-Vroman, R. and Yoshinaga, N. (1992). Broad and narrow constraints on the english dative alternation: Some fundamental differences between native speakers and foreign language learners. *University of Hawai'i Working Papers in ESL*, 11:157–199.

Bloch, B. (1947). English verb inflection. *Language*, 23(4):399–418.

Bloom, L. (1970). *Language development: Form and function in emerging grammar.* MIT Press, Cambridge, MA.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, pages 491–504.

Bloom, P. (2000). *How children learn the meanings of words.* MIT Press, Cambridge, MA.

Blum, L. and Blum, M. (1975). Toward a mathematical theory of inductive inference. *Information and control*, 28(2):125–155.

Borer, H. and Wexler, K. (1987). The maturation of syntax. In Roeper, T. and Williams, E., editors, *Parameter setting*, pages 123–172. Reidel, Berlin.

Bowerman, M. (1982). Reorganizational process in lexical and syntactic development. In Wanner, E. and Gleitman, L. R., editors, *Language acquisition: The state of the art*, pages 319–346. Cambridge University Press, New York.

Bowerman, M. and Croft, W. (2008). The acquisition of the English causative alternation. In Bowerman, M. and Brown, P., editors, *Crosslinguistic perspectives on argument structure: Implications for learnability*, pages 279–307. Erlbaum.

Boyd, J. K. and Goldberg, A. E. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, 87(1):55–83.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (1999). Training japanese listeners to identify english /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, 61(5):977–985.

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

Bush, R. R. and Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 68(3):313–323.

Bybee, J. L. (2006). *Frequency of Use and the Organization of Language.* Oxford University Press, Oxford.

Chomsky, N. (1955). The logical structure of linguistic theory. Ms., Harvard University and MIT. Revised version published by Plenum, New York, 1975.

Chomsky, N. (1958). [Review of Belevitch 1956]. *Language*, 34(1):99–105.

Chomsky, N. (1965). *Aspects of the theory of syntax.* MIT Press, Cambridge, MA.

Chomsky, N. (1968). *Language and mind.* Harcourt, Brace and World.

Chomsky, N. (2001). *Beyond explanatory adequacy.* MIT Working Papers in Linguistics, Cambridge, MA.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, 36(1):1–22.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English.* MIT Press, Cambridge, MA.

Clahsen, H. and Muysken, P. (1986). The availability of universal grammar to adult and child learners-a study of the acquisition of german word order. *Interlanguage studies bulletin (Utrecht)*, 2(2):93–119.

Clahsen, H., Rothweiler, M., Woest, A., and Marcus, G. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45:225–255.

Cochran, B. P., McDonald, J. L., and Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, 41(1):30–58.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning, Lake Tahoe, California.*

Conwell, E. and Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, 103(2):163–179.

Cook, V. and Newson, M. (2014). *Chomsky's universal grammar.* Blackwell, Oxford.

Crain, S., Koring, L., and Thornton, R. (2016). Language acquisition from a biolinguistic perspective. *Neuroscience & Biobehavioral Reviews.*

Crain, S. and Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, pages 522–543.

Culicover, P. W. (1999). *Syntactic nuts: Hard cases, syntactic theory, and language acquisition.* Oxford University Press, Oxford.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2(3–4):133–142.

Demuth, K. (1989). Maturation and the acquisition of the sesotho passive. *Language*, pages 56–80.

Demuth, K. (1996). The prosodic structure of early words. In Morgan, J. L. and Demuth, K., editors, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pages 171–186. Psychology Press.

Diessel, H. (2013). Construction grammar and first language acquisition. In Hoffmann, T. and Trousdale, G., editors, *The Oxford handbook of construction grammar*, pages 347–364. Oxford University Press, Oxford.

Ellis, N. C. and Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, 59(s1):90–125.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Epstein, S. D., Flynn, S., and Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Behavioral and Brain Sciences*, 19(4):677–714.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Fodor, J. D. (2001). Parameters and the periphery: Reflections on syntactic nuts. *Journal of Linguistics*, 37(2):367–392.

Fodor, J. D. and Crain, S. (1987). Simplicity and generality of rules in language acquisition. In MacWhinney, B., editor, *Mechanisms of language acquisition*, pages 35–63. Lawrence Erlbaum Associates, Hillsdale, NJ.

Friederici, A. D. (2017). *Language in Our Brain: The Origins of a Uniquely Human Capacity*. MIT Press, Cambridge, MA.

Gerken, L. (1994). A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of child language*, 21(03):565–584.

Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25(3):407–454.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

Goldberg, A. E. (1995). *Constructions*. University of Chicago Press, Chicago.

Goldin-Meadow, S. and Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, 391(6664):279–281.

Goldin-Meadow, S. and Yang, C. (2017). Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience and Biobehavioral Reviews*, 81(Part B):150 – 157.

Grimshaw, J. (1990). *Argument structure*. MIT Press, Cambridge, MA.

Grinstead, J. (2000). Case, inflection and subject licensing in child catalan and spanish. *Journal of child language*, 27(1):119–155.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., and Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, 65(2):203–257.

Guasti, M. T. (2004). *Language acquisition: The growth of grammar*. The MIT Press, Cambridge, MA.

Hadley, P. A., Rispoli, M., and Holt, J. K. (2017). Input subject diversity accelerates the growth of tense and agreement: Indirect benefits from a parent-implemented intervention. *Journal of Speech, Language, and Hearing Research*, 60(9):2619–2635.

Hadley, P. A. and Walsh, K. M. (2014). Toy talk: Simple strategies to create richer grammatical input. *Language, Speech, and Hearing Services in Schools*, 45(3):159–172.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.

Halle, M. and Vergnaud, J.-R. (1987). *An essay on stress.* MIT Press, Cambridge, MA.

Harley, H. and Miyagawa, S. (2016). Ditransitives.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, Baltimore, MD.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Hayes, B. (1982). Extrametricality and English stress. *Linguistic Inquiry*, 13(2):227–276.

Hoff, E. (2014). *Language development.* Wadsworth Cengage Learning, Belmont, CA.

Hornstein, N. and Lightfoot, D. (1981). Introduction. In Hornstein, N. and Lightfoot, D., editors, *Explanation in linguistics: The logical problem of language acquisition*, pages 9–31. Longman.

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic inquiry*, 15(4):531–574.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., and Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27(2):236.

Hyams, N. (1986). *Language acquisition and the theory of parameters.* Reidel, Dordrecht.

Hyams, N. (1991). A reanalysis of null subjects in child language. In Weissenborn, J., Goodluck, H., and Roeper, T., editors, *Theoretical issues in language acquisition: Continuity and change in development*, pages 249–268. Psychology Press.

Hyams, N. and Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry*, pages 421–459.

Ibbotson, P. and Tomasello, M. (2016). Evidence rebuts chomsky's theory of language learning. *Scientific American*, 315(5).

Inagaki, S. (1997). Japanese and Chinese learners' acquisition of the narrow-range rules for the dative alternation in English. *Language Learning*, 47(4):637–669.

Jackendoff, R. S. (1990). *Semantic structures.* MIT Press, Cambridge, MA.

Jaeggli, O. and Safir, K. J. (1989). The null subject parameter and parametric theory. In *The null subject parameter*, pages 1–44. Springer, Berlin.

Jelinek, F. (1998). *Statistical methods for speech recognition.* MIT Press, Cambridge, MA.

Judy, T. (2011). L1/l2 parametric directionality matters: More on the null subject parameter in l2 acquisition. *EUROSLA Yearbook*, 11(1):165–190.

Kanno, K. (1997). The acquisition of null and overt pronominals in japanese by english speakers. *Second Language Research*, 13(3):265–287.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, 56(3):263–269.

Kim, Y.-J. (2000). Subject/object drop in the acquisition of korean: A cross-linguistic comparison. *Journal of East Asian Linguistics*, 9(4):325–351.

Knuth, D. E. (2002). *Selected papers on computer languages.* CSLI Publications, Stanford, CA.

Kowalski, A. and Yang, C. (2012). Verb islands in child and adult language. In Biller, A. K., Chung, E. Y., and Kimball, A. E., editors, *BUCLD 36: Proceedings of the 36th annual Boston University Conference on Language Development*, volume 1, pages 281–289.

Krifka, M. (1999). Manner in dative alternation. In *West Coast Conference on Formal Linguistics*, volume 18, pages 260–271.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English.* Brown University Press, Providence.

Labov, W. (1972). *Sociolinguistic patterns.* University of Pennsylvania Press, Philadelphia.

Labov, W. (2007). Transmission and diffusion. *Language*, 83(2):344–387.

Lardiere, D. (2009). Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research*, 25(2):173–227.

Legate, J. A. and Yang, C. (2007). Morphosyntactic learning and the development of tense. *Language Acquisition*, 14(3):315–344.

Legate, J. A. and Yang, C. (2013). Assessing child and adult grammar. In Berwick, R. and Piattelli-Palmarini, M., editors, *Rich languages from poor inputs: In honor of Carol Chomsky*, pages 168–182. Oxford University Press, Oxford.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago.

Lieven, E. V., Pine, J. M., and Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of child language*, 19(02):287–310.

Lignos, C. and Yang, C. (2016). Morphology and language acquisition. In Hippisley, Andrew R. abd Stump, G., editor, *The Cambridge handbook of Morphology*, chapter 28, page Forthcoming. Cambridge University Press, Cambridge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In Jackson, W., editor, *Communication theory*, volume 84, pages 486–502. Betterworth.

Marchman, V. A. and Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21(2):339–366.

Marcus, G., Pinker, S., Ullman, M. T., Hollander, M., Rosen, J., and Xu, F. (1992). *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

Markson, L. and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385(6619):813–815.

Mazurkewich, I. and White, L. (1984). The acquisition of the dative alternation: Unlearning overgeneralizations. *Cognition*, 16(3):261–283.

McClelland, J. L. and Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11):465–472.

Miller, G. A. (1957). Some effects of intermittent silence. *American Journal of Psychology*, 70(2):311–314.

Miller, J. F. (1981). *Assessing language production in children: Experimental procedures*. Edward Arnold, London.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, 18(2):203–226.

Nagy, W. E. and Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3):304–330.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Newport, E. L. and Aslin, R. N. (2004). Learning at a distance: I. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2):127–162.

Nida, E. A. (1949). *Morphology: the descriptive analysis of words*. University of Michigan Press, Ann Arbor, 2nd edition.

Osherson, D. N. and Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58.

Perdue, C., editor (1993). *Adult language acquisition: Field methods*, volume 1. Cambridge University Press, Cambridge.

Pérez-Leroux, A. T. and Glass, W. R. (1999). Null anaphora in Spanish second language acquisition: probabilistic versus generative approaches. *Second Language Research*, 15(2):220–249.

Pesetsky, D. (1995). *Zero syntax: Experiencer and Cascade.* MIT Press, Cambridge, MA.

Phinney, M. (1987). The pro-drop parameter in second language acquisition. In Roeper, T. and Williams, E., editors, *Parameter setting*, pages 221–238. Kluwer, Dordrecht.

Pierce, A. (1992). *Language acquisition and syntactic theory: A comparative analysis of French and English.* Kluwer, Dordrecht.

Pine, J. M., Freudenthal, D., Krajewski, G., and Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Pine, J. M. and Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(2):123–138.

Pinker, S. (1979). Formal models of language learning. *Cognition*, 7(3):217–283.

Pinker, S. (1984). *Language learnability and language development.* Harvard University Press, Cambridge, MA.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* MIT Press, Cambridge, MA.

Pinker, S. (1995). Why the child holded the baby rabbit: A case study in language acquisition. In Gleitman, L. R. and Liberman, M., editors, *An invitation to cognitive science, Vol. 1: Language*, pages 107–133. MIT Press, Cambridge, MA.

Pinker, S. (1999). *Words and rules: The ingredients of language.* Basic Books, New York.

Pinker, S. and Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Science*, 6(11):456–463.

Pizzuto, E. and Caselli, M. C. (1994). The acquisition of Italian verb morphology in a cross-linguistic perspective. In Levy, Y., editor, *Other children, other languages*, pages 137–187. Lawrence Erlbaum, Hillsdale, NJ.

Pullum, G. K. and Wilson, D. (1977). Autonomous syntax and the analysis of auxiliaries. *Language*, 53(4):741–788.

Rizzi, L. (1990). *Relativized minimality.* The MIT Press, Cambridge, MA.

Rothman, J. and Slabakova, R. (2017). The generative approach to sla and its place in modern second language studies. *Studies in Second Language Acquisition*, pages 1–26.

Saffran, J. R., Aslin, R. N., and Newport, E. (1996). Statistical learning by by 8-month-old infants. *Science*, 274:1926–1928.

Sakas, W. G. and Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, 19(2):83–143.

Sakas, W. G., Yang, C., and Berwick, R. (2017). Parameter setting is feasible. *Linguistic Analysis*, 41(1):(In press).

Sapir, E. (1928). *Language: An introduction to the study of speech.* Harcourt Brace, New York.

Schuler, K. (2017). *The acquisition of productive rules in child and adult language learners.* PhD thesis, Georgetown University, Washington, D.C.

Schuler, K., Yang, C., and Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *The 38th Cognitive Society Annual Meeting*, Philadelphia, PA.

Schwartz, B. D. and Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second language research*, 12(1):40–72.

Silvey, C. and Christodoulopoulos, C. (2016). Children's production of determiners as a test case for innate syntactic categories. In Roberts, S., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Fehér, O., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11).*

Smith, N. V. and Tsimpli, I.-M. (1995). *The mind of a savant: Language learning and modularity.* Blackwell Publishing, Oxford.

Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22.

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19(3):513–531.

Stevens, J., Trueswell, J., Yang, C., and Gleitman, L. (2017). The pursuit of word meanings. *Cognitive Science*, 41:638–676.

Straus, K. J. (2008). *Validation of a probabilistic model of language acquisition in children.* PhD thesis, Northeastern University.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, 29(1):103–114.

Terrace, H. S. (1987). *Nim: A chimpanzee who learned sign language.* Columbia University Press.

Terrace, H. S., Petitto, L.-A., Sanders, R. J., and Bever, T. G. (1979). Can an ape create a sentence? *Science*, 206(4421):891–902.

Tettamanti, M., Alkadhi, H., Moro, A., Perani, D., Kollias, S., and Weniger, D. (2004). Neural correlates for the acquisition of natural language syntax. *NeuroImage*, 17:700–709.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Harvard University Press, Cambridge.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Tomasello, M. (2003). *Constructing a language.* Harvard University Press, Cambridge, MA.

Trueswell, J. C., Sekerina, I., Hill, N. M., and Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73(2):89–134.

Valian, V. (1990). Null subjects: A problem for parameter-setting models of language acquisition. *Cognition*, 35(2):105–122.

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

Valian, V., Solt, S., and Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, 36(4):743–778.

Wang, Q., Lillo-Martin, D., Best, C. T., and Levitt, A. (1992). Null subject versus null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition*, 2(3):221–254.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua*, 106(1):23–79.

Wexler, K. and Culicover, P. (1980). *Formal principles of language acquisition.* MIT Press, Cambridge, MA.

White, L. (1985). Is there a" logical problem" of second language acquisition? *TESL Canada Journal*, 2(2):29–42.

White, L. (1990). The verb-movement parameter in second language acquisition. *Language Acquisition*, 1(4):337–360.

White, L. (2003). *Second language acquisition and universal grammar.* Cambridge University Press.

Wiese, R. (1996). *The phonology of German.* Clarendon, Oxford.

Xu, F. and Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, 22(3):531–556.

Yang, C. (2000). Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250.

Yang, C. (2002). *Knowledge and learning in natural language.* Oxford University Press, Oxford.

Yang, C. (2012). Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(2):205–213.

Yang, C. (2013a). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327

Yang, C. (2013b). Who's afraid of George Kingsley Zipf? Or: Do children and chimps have language? *Significance*, 10(6):29–34.

Yang, C. (2015a). For and against frequencies. *Journal of Child Language*, 42(2):287–293.

Yang, C. (2015b). Negative knowledge from positive evidence. *Language*, 91(4):938–953.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break rules of language*. MIT Press, Cambridge, MA.

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24(2):100–125.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., and Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, 81(Part B):103 – 119.

Yang, C. and Montrul, S. (2017). Learning datives: The tolerance principle in monolingual and bilingual acquisition. *Second Language Research*, 33(1):119–144.

Yip, K. and Sussman, G. J. (1997). Sparse representations for fast, one-shot learning. In *Proceedings of the National Conference on Artificial Intelligence*, pages 521–527.

Yu, C. and Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.