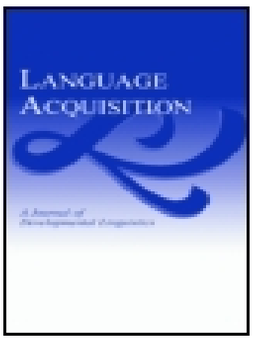

Language Acquisition

ISSN: 1048-9223 (Print) 1532-7817 (Online) Journal homepage: http://www.tandfonline.com/loi/hlac20

Rage against the machine: Evaluation metrics in the 21st century

Charles Yang

To cite this article: Charles Yang (2017): Rage against the machine: Evaluation metrics in the 21st century, Language Acquisition, DOI: 10.1080/10489223.2016.1274318

To link to this article: http://dx.doi.org/10.1080/10489223.2016.1274318

Accepted author version posted online: 17 Mar 2017.
Published online: 17 Mar 2017.

LANGUAGE ACQUISITION
http://dx.doi.org/10.1080/10489223.2016.1274318

Routledge
Taylor & Francis Group

Rage against the machine: Evaluation metrics in the 21st century

Charles Yang

University of Pennsylvania

ABSTRACT

I review the classic literature in generative grammar and Marr's three-level program for cognitive science to defend the Evaluation Metric as a psychological theory of language learning. Focusing on well-established facts of language variation, change, and use, I argue that optimal statistical principles embodied in Bayesian inference models are ill-suited for language acquisition. Specific attention will be given to the Subset Problem: Indirect negative evidence, which can be attractively formulated in the Bayesian framework, is ineffective when the statistical properties of language are examined in detail. As an alternative, I suggest that the Tolerance Principle (Yang 2016) provides a unified solution for the problem of induction and generalization: It bridges the computational and algorithm levels in Marr's formulation, while retaining the commitment to the formal and empirical constraints in child language development.

ARTICLE HISTORY
Received 30 November 2016
Accepted 4 December 2016

1. An I-Language approach

As I understand it, the Evaluation Metric was conceived as a mechanical procedure that chooses the simplest among grammatical analyses and was first applied in Chomsky's master's thesis to provide an account of Modern Hebrew morphophonemics (1951). Similar discussions can be found in *The Logical Structure of Linguistic Theory* (LSLT; Chomsky 1955), but it was in *Aspects* where the Evaluation Metric took a decidedly psychological turn. The *Aspects* model of language acquisition has the following components (Chomsky 1965:§6):

(1) (a) a technique of representing input signals;
(b) a way of representing structural information about these signals;
(c) some initial delimitation of a class of possible hypotheses about language structure;
(d) a method of determining what each such hypothesis implies with respect to each sentence;
(e) a method for selecting one of the (presumably, infinitely many) hypotheses that are allowed by (c) and are compatible with the given primary linguistic data.

The Evaluation Metric, or (1e), becomes part of the language learning algorithm that the child uses to determine the grammar from the primary linguistic data. The study of Evaluation Metrics reached its height in the years that immediately followed: *The Sound Pattern of English* (SPE; Chomsky & Halle 1968), for instance, put forward several detailed technical proposals that were intensively discussed and debated.

It is possible to read too much into a word, but the term Evaluation Metric (or Measure) was rebranded as Evaluation *Procedure* in SPE, perhaps reflecting the change from an information-theoretic conception in LSLT to an explicitly algorithmic commitment from *Aspects* on. But a central feature has

CONTACT Charles Yang charles.yang@ling.upenn.edu University of Pennsylvania, Department of Linguistics & Computer Science, 3401 Walnut Street 315-C, Philadelphia, PA 19104-6305.

remained constant throughout these developments. The Evaluation Metric has always been understood within an empirical framework of linguistic theory, rather than a generic preconception of simplicity or optimality. In LSLT, the selection of a grammar (relative to a corpus) requires the justification of a general theory of linguistic structure: It is not a "mathematical game" detached from the reality of the world's languages (I-3.1), but "one indispensable aspect of the validation of a grammar of a given language is the construction of grammars for other languages" (I-11). In modern terms, an Evaluation Metric is embedded in a particular theory of Universal Grammar (UG). The empirical nature of the Evaluation Metric is strongly emphasized in *Aspects* (§7). Chomsky rejects simplicity as "a general notion somehow understood in advance outside of linguistic theory." The pool of grammars available for evaluation is a psychological object: "(W)e are given, in part, an empirical pairing of certain kinds of primary linguistic data with certain grammars that are in fact constructed by people presented with such data." The requirement of explanatory adequacy places further constraints on the Evaluation Metric: Can it guide the child to the target grammar "within the given constraints of time and access, and within the range of observed uniformity of output?" (54). To use the terminology introduced in Chomsky (1986), an Evaluation Metric is part of one's I-language, an individual's internal system of linguistic knowledge. The empirical, and domain-specific, orientation of the Evaluation Metric in turn has had significant implications for the role of simplicity in the philosophy of science (e.g., Sober 1975).

From the 1970s on, however, the Evaluation Metric all but vanished except in the relatively small field of formal learnability. There are multiple reasons for this, some reasonable and some not, but the net effect is that a mechanistic approach to language, strongly embodied in early generative grammar, received less attention, and the study of language acquisition has suffered as a result. Progress has been limited to what the child knows about her language at which specific stage of development, rather than how such linguistic knowledge is attained. For instance, the language-specific phonetic/phonemic inventory is largely established before the age of 1;0 (Werker & Tees 1984; Kuhl et al. 1992), but there is no successful account of how these linguistic units emerge from the input or how the phonological grammar is selected. Very young children show mastery of language-specific syntactic categories and rules (Brown 1973): Production errors are vanishingly rare (Valian 1986), rule usage is systematically productive (Yang 2013) and can guide the child to deduce the properties of novel words (Shi & Melançon 2010). Yet the distributional learning models for category and rule learning are nowhere near to the requisite level of accuracy. The Principles and Parameter framework (Chomsky 1981), a very successful "delimitation of a class of possible hypotheses about language structure," does take away some of the tasks originally charged to the Evaluation Metric, but we still do not know what kind of learning algorithm can navigate through a realistically complex space of grammars (Yang 2002a; Sakas & Fodor 2012).

The renewed interest in the Evaluation Metric, then, can only be a welcome move to understand how the components of language acquisition in (1) fit together. And we surely are in a better position than the 1960s to investigate these issues. The availability of electronic corpora, including those from child and child-directed language, provides an increasingly accurate measure of the input to as well as the output of language acquisition. Advances in computational learning theory and high-performance hardware have made the investigation of Evaluation Metrics more practical than before. But along with these new developments lurks the danger of taking us further away from understanding child language, unless the empirical nature of the Evaluation Metric remains in sight.

In this article, I would like to reinforce the conception of Evaluation Metrics as a component of the I-language and develop a critique of approaches to language learning that abandon psychological commitments, especially idealized learning models. The argument comes in three parts. Section 2 reviews Marr's influential three-level program for cognitive science, which has often been invoked as the intellectual background for idealized learning models. As we will see, however, the computational level advocated and practiced by Marr, just like the Evaluation Metric in generative grammar, is in fact deeply connected with considerations at the other levels. Section 3 is devoted to the unique difficulties that arise in language learning. Reviewing studies of language variation, change, and use, I suggest that the ideal observer model of statistical inference is illsuited for language acquisition.

Special attention will be given to indirect negative evidence, a powerful learning strategy that has been used to eliminate superset hypotheses in Bayesian inference (e.g., the Size Principle; Tenenbaum & Griffiths 2001; Xu & Tenenbaum 2007). Focusing on a class of English adjectives and its acquisition, I show that the statistical distribution of language blunts the effectiveness of indirect negative evidence regardless of its formulation. Section 4 develops a more positive line of the argument. I review the *Tolerance Principle* (Yang 2002b, 2005, 2016), an Evaluation Metric driven by the principle of computational efficiency and grounded in the empirical study of language structure and use. The Tolerance Principle provides a calculus that allows the learner to determine the correct scope of linguistic generalizations. As a case study, I show that the Tolerance Principle successfully resolves the acquisition of English dative constructions (Baker 1979; Pinker 1989), a classic puzzle in language learnability and development.

2. Poor man's vision and the idealized observer

When developing models of language learning, there is always a delicate balancing act in striving for the most general result while still contributing to the specific studies of language development by children. The maximally general statement "Let there be *N* grammars" is the default starting position for the mathematician: It may lead to formal results applicable across frameworks but can hardly inform the child how to set the head directionality parameter. In my view, a great virtue of the traditional language learnability research (e.g., Wexler & Culicover 1980; Berwick 1985; Dresher & Kaye 1990; Gibson & Wexler 1994; Sakas & Fodor 2001; Yang 2002a) lies in its respect for what is known, or what can be plausibly assumed, about the actual language acquisition process.

The development of an Evaluation Metric, which is now regarded as a component of the language learning model, faces exactly the same challenge. In the extreme case, one may trivialize the delimitation of possible grammars (as in 1) and rely on the Evaluation Metric to carry the full load of language acquisition. One of the earliest examples in this direction is the Bayesian framework of grammatical inference developed by Horning (1969), which has been revived and extended in recent years (e.g., Johnson, Griffiths & Goldwater 2007; Perfors, Tenenbaum & Regier 2011; O'Donnell 2015). In Horning's model, the learner evaluates probabilistic context free grammars (PCFG), a formal language model well known to be inappropriate for human language syntax on formal (Huybregts 1984, Shieber 1985) as well as probabilistic (Manning & Schütze 1999) grounds. No Evaluation Metric can salvage a learning model where the hypotheses are not "in fact constructed by people."

These concerns are not unique to the study of language but are of great importance to anyone interested in understanding how learning actually works. It is worth remembering that the mathematical psychology of learning (e.g., Estes 1950; Bush & Mosteller 1951), which has generated a rich and rigorous body of literature, is equally committed to empirically plausible and testable learning mechanisms even though it generally focuses on much simpler behavioral domains than language. In an article that anticipates many contemporary discussions (e.g., Jones & Love 2011; Bowers & Davis 2012; Marcus & Davis 2013), Pat Suppes (1966) critically assesses the Bayesian approach of learning ("concept formation") and decision making. According to the Bayesian view, the problem of inductive inference is to be replaced with the selection of a probabilistic distribution over "all the possible states of the world and all the possible future histories" (22) which presumably include hypotheses never constructed by people. For Suppes, the problem of concept formation is strikingly similar to the language acquisition model in *Aspects*:

> It is, I would take it, the central problem of a theory of concept formation to provide such a structure and to state the laws by which organisms use the structure to solve the problem confronting them. … On the other hand, it is precisely the imposition of structures that seems to be necessary to bring some order and constraints to discouragingly large number of possible concepts that may be considered in solving even a relatively simple problem. … The core of the problem is that of developing an adequate psychological theory to describe, analyze and predict the structure imposed by organisms on the bewildering complexities of possible alternatives facing them. (43, 47)

In other words, the learner requires "some initial delimitation of a class of possible hypotheses" and "a method of determining what each such hypothesis implies with respect to each sentence": a "UG" for conception formation, in addition to an effective mechanism that chooses the correct hypothesis from the available candidate set. Suppes points out that in behavioral studies, hypotheses are actually *formed* as the subject processes information but not available a priori to be evaluated in Bayesian terms. He also enumerates some major difficulties in relating the Bayesian approach with well-established results in the stimulus-sampling tradition, where again the subject shows adaptive responses that seem best accounted for by online learning as input stimuli are presented.

Suppes's remarks are highly pertinent to the study of language learning, which we develop further in Section 3. The proponents of the Bayesian approach would not deny the delimiting factors on possible concepts or grammars; in fact, they would be happy to include these among the set of prior hypotheses. In practice, however, research energy is typically devoted to calibrating the statistical machinery (e.g., prior probabilities, likelihood functions) that can be fitted across domains, rather than working out or incorporating constraints specific to the topic of study. To take a concrete example, consider a narrow task in the problem of word learning, namely, how to determine the semantic reference of phonological words (that is, /kæt/is that fuzzy thing that meows). There is a class of "global" models that attempts to resolve the problem of referential ambiguity by aggregating situational data from a large number of word occurrences with lexical items (e.g., "cat"). These models, broadly known as cross-situational learning, record the totality of the information presented to the learner to identify the best lexicon from the space of all possible lexicons (Xu & Tenenbaum 2007; Yu & Smith 2007; Frank, Goodman & Tenenbaum 2009, Fazly, Alishahi & Stevenson 2010). Putting aside computational issues for the moment (more in Section 3), these proposals are at odds with a series of experimental results showing that not only do experimental subjects (obviously) fail to evaluate all possible lexicons, they seem to attend to probably no more than two referents for each learning instance (Medina et al. 2011; Trueswell et al. 2013; Köhne, Trueswell & Gleitman 2013; Stevens et al. 2016).

Nevertheless, recent years have seen a surge of interest in idealized models that do not share the traditional concerns in the study of learning. It has become acceptable, or even commendable, to develop idealized learning models (e.g., Chater & Vitányi 2007; Xu & Tenenbaum 2007; Goldwater, Griffiths & Johnson 2009; Perfors, Tenenbaum & Regier 2011; Feldman et al. 2013) that explicitly disavow any claim of psychological mechanisms. Learning is viewed as an idealized observer that chooses the optimal or near-optimal hypothesis among (generally) a vast set of candidates, typically in the framework of Bayesian inference (Tenenbaum et al. 2011).

At this point, the spirit of David Marr (1982) is trotted out. The rhetorical point is that a model needn't be feasible as long as it operates at Marr's computational level, a specification of the problem to be solved (e.g., the input/output conditions of language acquisition), rather than the representational structure, algorithmic process, or neurobiological implementation of how language is actually learned. For instance, a leading group of Bayesian cognitive scientists state that "(T)he probability calculus then describes inferences that can be drawn by combining these beliefs with new evidence, without the need to commit to a process-level explanation of how these inferences are performed (Marr 2010)" (Goodman et al. 2015: 539). Similarly, Tenenbaum et al. (2011) invoke the computational level as a prerequisite to "reverse engineer" human learning and cognitive systems.

For a field such as language acquisition, these recent developments do not seem to represent progress. In the formal study of language acquisition, the absence of a process-level learning model, or unrealistic assumptions about the learner's computational capacity, is a defect not a virtue. Had idealized learning been accepted, the problem of language acquisition would have been solved long ago. For instance, if the learner were to have access to the D-structure of sentences — surely an idealized learner can read off the care taker's intentional states — then we already have the learnability proof of *Aspects*-style transformational grammars (Wexler & Culicover 1980). Similarly, if the requirement of explanatory adequacy is dispensed with, then a learner can embark on a random walk in the land of grammars. Under most modern theories of language (Chomsky 1981; Prince & Smolensky 2004), the finiteness of possible grammars guarantees that they will eventually stumble upon the target grammar eventually, long after everyone on the planet Earth is dead.

Furthermore, I believe that the idealized learning approach as currently proclaimed and practiced is a misreading of Marr's program, in both spirit and letter. In the remainder of this section, I discuss Marr's own work and the research tradition it has engendered, including contemporary assessments offered by practicing (computational) neuroscientists close to Marr's tradition. They do not provide conceptual or methodological underpinning for computational-level models that are detached from the empirical considerations at the other levels.

In a highly influential treatise, David Marr proposes three distinct levels of descriptions in the study of information-processing systems (Table 1).[1]

Having placed the psychophysics and neuroscience of vision at the algorithmic and implementational level, Marr provides a vigorous defense of the computational level, by now familiar to all cognitive scientists. Despite stressing the independence of these three levels, Marr nevertheless maintains how these levels must be integrated to obtain a full understanding of the problem.[2]

> There must exist an additional level [the computational level] of understanding at which the character of the information-processing tasks carried out during perception are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads. That was what was missing — the analysis of the problem as an information-processing task. *Such analysis does not usurp an understanding at the other levels* — of neurons or of computer programs — but it is a necessary complement to them, since without it there can be no real understanding of the function of all those neurons. (Marr 2010: 19; emphasis added)

In Marr's own work, a computational level theory should inform the study of the algorithmic and implementational levels, which in turn provide constraints for the computational level. It is instructive to consider what Marr himself (Marr 2010:111) regards as a computational-level theory, the cooperative algorithm for stereo disparity proposed by Marr & Poggio (1976) and elaborated in subsequent work (Marr, Palm & Poggio 1978; Marr & Poggio 1979).

In vision, objects separated in depth are perceived at slightly different positions by the two eyes. The brain can make use of the angular disparity in the two images to estimate the distance of the object from the viewer. The problem of stereo disparity is to establish the correspondence between the two images by attending to the relative positions of certain locations such as the lines and edges of objects. The problem is one of ambiguity resolution. On a random-dot stereogram of the type studied by Julesz (1971) (and can often be found at airport shops), n dots will be perceived by both the left and the right eyes: The ambiguity lies in which correspond to which out of the logically possible n^2 pairs. Instead of framing the stereopsis problem as the search for the globally optimal solution, Marr & Poggio proposed an iterative algorithm that directly encodes structural constraints that can operationalized in a strictly local fashion. Their theory is schematically outlined as follows, where I have taken the liberty to simplify the notation.

$$C^{t+1} = \sigma\left[\sum_{C\in S} C^t - \sum_{C\in O} C^t\right]$$

Table 1. Description of the three levels (from Marr 2010).

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

[1] I am indebted to Constantine Lignos who, in his dissertation (2013), initiated the discussion of Marr's levels in the setting of language acquisition, with specific reference to the problem of infant word segmentation.

[2] All page numbers in Marr's *Vision* refer to the 2010 MIT Press reprinting of the original 1982 edition.

Here C stands for the state of a cell or processor placed on a two-dimensional grid where the horizontal and vertical lines represent the lines of sights from the left and the right eye. The state of a cell at the time, or C^t, has the value of 1 if two black dots from the two images can match, and 0 otherwise. In the formulation above, C^{t+1} is determined, via a threshold function σ, by the states of cells at time t in its two local neighborhoods on the grid: specifically, the difference between the overall level of excitation and inhibition defined over two local neighborhood regions, S and O, respectively. The definition of S and O directly encodes psychophysical constraints of stereopsis. For instance, the Uniqueness constraint (Rule 2; Marr & Poggio 1976 and Marr 2010:115) holds that a black dot from one image generally cannot match more than one black dot from the other image. This constraint is implemented by asserting that if a particular cell has the state value of 1, then all the cells along the horizontal and vertical lines, i.e., those in the inhibitory neighborhood (O), must have the value of 0. As the algorithm iterates, the state of cells changes, triggering changes in the excitatory and inhibitory neighborhoods all the while. The algorithm converges when the states of the cells are static ($C^{t+1} = C^t$ for all cells).

This is probably more computer vision than warranted in an article on English dative constructions, but I wish to make clear that in Marr's own conception and practice, the computational theory does in fact "commit to a process-level explanation of how these inferences are performed" (*pace* Goodman et al. 2015). The stereo disparity algorithm is not an idealized model but one that builds on psychophysical motivations and comes with attractive computational properties. It converges rapidly and accurately, which was demonstrated by simulation and later formally analyzed (Marr, Palm & Poggio 1978; Grimson 1981). Furthermore, the notion of the cell as a computing device was inspired by the available neurophysiological evidence for disparity detectors in the primary visual cortex. Indeed, further studies in the psychophysics of stereopsis prompted the revision of the Marr & Poggio 1976 model, resulting in the theory of Marr & Poggio (1979), which Marr regarded as superior because it provided a better bridging between the levels (2010:111).

Over the years, however, the interdependence across the three levels seems lost on many of Marr's followers. As the quote from Goodman et al. (2015) shows, modern cognitive scientists appear to hold that, as long as Marr is invoked, a computational theory can operate with impunity, free from the formal and empirical constraints in the domain where the problem is to be solved. This doesn't seem to be the interpretation that Marr had in mind. In a recent tribute to Marr's contribution to computational neuroscience, Willshaw, Dayan & Morris (2015) seek to correct these misunderstandings exemplified in purely bottom-up "hypothesis-free 'omic' neuroscience" such as the Human Brain Project and "implementation-free computational approaches that aim to explore whether the brain functions according to optimality principles."[3]

> A common reading of his later work is that it is appropriate and sufficient to start from the computational level; divorced from implementation considerations, information processing can be readily formulated as optimal inference and control, using ideas from fields such as statistics, operations research, economics and machine learning. However, as recognized throughout the book *Vision*, these accounts are limited. In all but the very simplest circumstances, optimal inferences and control are radically intractable for animal and machine alike, and so are formally limited or even useless. . .. This opens a critical channel of reverse communication between Marr's three levels. (Willshaw, Dayan & Morris 2015:8)

Marr might have foreseen the misunderstandings in the years to come. In any case, he certainly regarded the *actual* human information processing capacity to be the primary criterion for the development of computational theories. In an imaged conversation with a skeptical interlocutor:

> [Question] I cannot really accept that the computational theory is so independent of the other levels. To be precise, I can imagine that two quite distinct theories of a process might be possible. Theory 1 might be vastly superior to theory 2, which may be only a poor man's vision in some way, but it could happen that neural nets have no easy way of implementing theory 1 but can do theory 2 very well. Effort would thus be misplaced in an elaborate development of theory 1.

[3]Tommy Poggio, Marr's closest collaborator, similarly cautions against the detachments of the three levels and strongly emphasizes the need for reintegration (Marr 2010: 365).

> [Answer] Yes, this could certainly happen, and I think it already has in the case of deriving shape from shading. I would not be at all surprised if it was unreasonably difficult to solve Horn's integral equations for shape shading with neural networks, yet the equations can be solved on a computer for simple cases. Human ability to infer shape from shading is very limited, and it may be based on simplistic assumptions that are often violated — a sort of theory 2 of the kind you mentioned. Nevertheless, I doubt that the effort put into a deep study like Horn's was misplaced, even in the circumstances. Although it will not yield direct information about human shape-from-shading strategies, it probably provides indispensable background information for discovering the particular poor man's version that we ourselves use. (Marr 2010:339)

We are in the business to understand the poor man's vision, not that of an idealized observer. Likewise, we are interested in how children learn human languages, not how an all-powerful computing machine may stumble on the meaning of "cat."

Before we return to the problem of language learning, I should state up front that the Bayesian inference model has made important contributions to the study of learning and decision making and will continue to do so. Yet as we will see, the empirical studies of language turn up some unique and serious challenges. My central claim is that optimal statistical principles, as embodied in the current Bayesian inference models, are problematic computational level theories for language acquisition, even if their detachment from the other levels were granted.

3. Bayesian optimality and linguistic reality

It is easy to see why the Bayesian inference model offers a natural implementation for the Evaluation Metric. A set of candidate grammars is provided, and the Bayesian framework offers a principled method for selecting the winner. Most Bayesian inference models in language and cognitive model adopt the maximum a posteriori estimator and have the following components:

(2) (a) a set of hypotheses to be selected H;
(b) the *prior* probabilities of the hypotheses $P(H)$;
(c) a likelihood function that calculates the probability of the observed learning data D under the hypotheses H, or $P(D|H)$, to obtain
(d) the *posterior* probabilities of the hypotheses $P(H|D) = P(H)P(D|H)/P(D)$ (Bayes's rule), which are normalized by $P(D)$ to select the optimal hypothesis $\hat{H} = \underset{H}{\text{argmax}}\ P(H)P(D|H)$.

Consider a nonlinguistic example. Suppose that there are two coins, one fair and one biased with heads on both sides. We randomly select one and would like to know whether it is fair or biased on the basis of its behavior. Here the prior probabilities of $P(H = \text{fair})$ and $P(H = \text{biased})$ are both $1/2$. The data, D, consists of 100 trials of coin toss where we obseved only heads. The likelihoods are: $P(D|H = \text{fair}) = 1/2^{100}$ and $P(D|H = \text{biased}) = 1$. Multiplying with the prior probabilities, it is clear than $H = \text{biased}$ is overwhelmingly favored over $H = \text{fair}$.

So far so reasonable, and Bayesian inference has found successful applications in many problems of science and engineering. Indeed, for language researchers, including those in the generative tradition, this appears to be one of the most attractive features of the Bayesian framework. Surely we can put all of our favorite UG assumptions — principles, parameters, constraints, etc. — into the hypothesis set. Furthermore, varying their prior probabilities allows for a principled way to state markedness, defaults, learning biases, and even typological tendencies. In short, put whatever you want in the priors, and trust the good Reverend Bayes to pick the winner.

Unfortunately, in the domain of language, and quite unlike the coin toss example, we have little idea about the priors and likelihoods about the hypotheses that correspond to actual linguistic data. A simple twist on the coin toss problem illustrates the need to independently justify all the components of the Bayesian program (2). Instead of picking one of the two coins randomly, consider the case where someone hands us one of them, and the same sequence of 100 heads ensues. If you still think that the coin is almost surely biased, then you are not a very good Bayesian — or more interestingly, Bayesian inference is not a

very good model of you, a psychological object. In fact, we cannot be confident at all about any conclusion. The reason is that we do not know anything about the prior probabilities. Suppose that the coin-giver *always* hands over the fair coin: The posterior probability of $P(H = \text{fair}\,|D = 100\ \text{heads})$ is of course 1!

In practice, then, we see many Bayesian models of language where prior probabilities and likelihood functions are chosen arbitrarily or on the basis of unverified intuition. For instance, grammars with fewer symbols or lexicons with shorter words are favored with higher prior probabilities (Goldwater, Griffiths & Johnson 2009; Perfors, Tenenbaum & Regier 2011; Perfors, Tenenbaum & Wonnacott 2010), even though there is no evidence that word length, which varies across languages due to their phonotactic properties, affects the vocabulary development (e.g., Tardif, Shatz & Naigles 1997), or that languages with more rules (e.g., presumably those with freer word order) are harder to learn than languages with fewer rules (Brown 1973, Slobin 1997). More worryingly, model parameters are often manually tuned to produce the desired outcome, to maximize performance scores (e.g., Frank, Goodman & Tenenbaum 2009; Dillon, Dunbar & Idsardi 2013, Feldman et al. 2013) or to fit experimental findings (e.g., Culbertson, Culoerfsan, Smolensky & Legendre 2012, Myslín & Levy 2016) without further validation of testing on unseen data. The choice of likelihood function is also often adhoc because there has been little empirical study of the actual likelihood functions used in language (if they are used in language). What the Bayesian model needs is the probability of sentences such as "Thank you" and "He drinks coffee" under a specific grammar, but these questions have never been seriously investigated. (After all, the key task for the child, and the acquisition researcher, is to formulate a grammar that produces "I am hungry," not how often "I am hungry" is produced.) Of course, there have been many formulations of specific probabilistic language *models* (e.g., N-grams, Markov chains, PCFGs, etc.) under which the probabilities of strings can be calculated, usually under various independence assumptions. The question, then, concerns the empirical legitimacy of these models, which are almost always multiplicative in nature: Shorter sentences are favored exponentially over longer ones, and language is in effect a finite system, which is by no means a universally accepted assumption (Chomsky 1957).

In what follows, I will largely sidestep these details of Bayesian implementation. Rather, I will focus on some empirical aspects of language that pose more fundamental challenges for the principles of Bayesian approach regardless of its specific formulation.

3.1. *Optimal inference and language variation*

The optimal principles of Bayesian inference seem too rigid to account for the fact of language variation, change, and use.

Consider an idealized setup that is strongly analogous to the previous coin toss example. Suppose the learner is to choose between two opposite values for a syntactic parameter such as verb raising. Here the value is [+] for French (H_+) and [–] for English (H_-), and the two languages can be distinguished by the position of finite main verbs relative to negation and certain adverbs: "Jean voit$_t$ souvent/pas t Marie" is possible for French but not English. Given a sample of French data that presumably contain such disambiguating sentences, $P(D|H_-)$ will be lowered: $P(D|H_+)$ will be favored by the child learner, and the parameter value for French can be correctly set. In this idealized setting of language learning, the Bayesian model is imminently applicable.[4] But significant difficulties arise when a fuller range of linguistic facts is taken into account.

First, it is now firmly established that linguistic environments, and the individual's linguistic knowledge, are both inherently variable (Weinreich, Labov & Herzog 1968). There are numerous instances of stable linguistic variation, where the speakers in a speech community show remarkably uniform levels of variable rule or grammar use (Labov 1972, 2007). In the much-studied phenomenon known as "t/d-deletion" (Labov 1989), for instance, English speakers probabilistically omit the word-final alveolar stops in a structurally

[4]Though how well it works must be compared against alternative formulations, some of which are domain specific (e.g., triggering; Gibson & Wexler 1994), while others are domain general (e.g., reinforcement learning such as the Bush & Mosteller (1951) model of linear reward penalty scheme used in Yang (2002a) for parameter setting).

well-defined set of words ("just" becomes /jus/, "bagged" becomes /bæg/), and their phonological, morphological, and grammatical environments have strong influence on the rate of deletion. Young children by the age of 3;00 can successfully acquire both the structural and probabilistic constraints on these linguistic variables (Roberts & Labov 1995; Smith, Durham & Fortune 2009). Similar results have been found for the acquisition of morphology, syntax, and semantics under inconsistent or unstable input conditions (Miller & Schmitt 2012; Yang, Ellman & Legate 2015, Han, Musolino & lidz 2016).

The situation of language change is similar. When historical linguists use terms such as "erosion" or "optionality," they are describing the gradual competition of alternative linguistic forms over time. One of the most detailed studies concerns the rise of the periphrastic *do* in the history of English (Kroch 1989), where the older grammar of moving the main verb to Tense (similar to Modern French noted earlier) was gradually replaced by the new form that places *do* in the Tense position. Over the span of several centuries, sentences such as "Quene Ester *looked* never with swich an eye," where the finite verb ("looked") preceded the adverb ("never"), slowly disappeared. Detailed analysis of historical data shows clearly that an individual speaker/writer had simultaneous command of multiple grammatical options at the level of phonology, morphology, and syntax (Taylor 1994; Pintzuk 1999; Sankoff & Blondeau 2007). For instance, in Santorini (1992)'s study of 16th-century Yiddish texts, an author demonstrated simultaneous use of an older INFL-final and an newer INFL-medial grammar — highlighted by the boldfaced finite verbs in (3) — on consecutive pages of the manuscript.

It is difficult to reconcile the Bayesian inference model with the fact of linguistic variation. In fact, it is difficult to see how language could ever change under Bayesian inference. When a new, and ultimately

(3) (a) vas er zeyn tag fun zeynm r[ebe] gilernt **hat**
what he his day from his rabbi learned **has**
'what he learned from his rabbi in his day'
(b) d[a]z der mensh **git** erst oyf in di hikh
that the human **goes** first up in the height
'the people first grow in height'

winning, linguistic variant first emerged, it necessarily constitutes a low probability event in the environment and is therefore subject to immediate elimination under the Bayesian optimality assumption, resulting in no change at all.[5] In the few studies of language change in the Bayesian framework, these issues are overlooked. For instance, in the iterated learning model (e.g., Kirby, Dowman & Griffiths 2007), the child only learns from a single individual in the previous/parent generation. As Niyogi & Berwick (2009)'s formal analysis shows, the iterated learning model is never able to converge on a shared language for the whole community; the uniformity of language variation and change that can be observed across a speech community, however, is striking (Labov 1972; Kroch 1995). Furthermore, the Bayesian character of iterating learning implies that the child will only acquire one grammar out of those in a linguistically heterogeneous environment, which is contrary to the reality of language variation and change just reviewed. At the minimum, then, a Bayesian learner must be able to converge on a distribution of mutually incompatible hypotheses in the environment: The optimal statistical principle of selecting the maximum a posteriori hypothesis need to be relaxed or abandoned.

In my earlier work (Yang 2000, 2002a), I have suggested that the heterogeneity and dynamics of language variation can be straightforwardly accounted for by reinforcement learning models from mathematical psychology. [6] For instance, a rodent can gradually converge on the probabilities of reward/penalty available at the two branches of a T-maze. The behavior of probability matching is suboptimal: The rodent is better off heading toward the branch with the higher probability of receiving a reward, rather than

[5]Or immediately favored, for whatever reason, eliminating the old form instantly, which is also inconsistent with the record of variation in language change.

[6]That is not to say that the learner will always recapitulate the statistical distribution of linguistics variants in the environment. In fact, the distribution may be gradually altered by the relative "fitness" of the competing variants, resulting in language change; see Yang (2000) for details.

spreading its odds over multiple options. As it stands, Bayesian inference models, which favor optimal actions, must be supplemented with ancillary assumptions to reproduce probability matching behavior (Eberhardt & Danks 2011; Bowers & Davis 2012).[7] In some cases, strikingly non-Bayesian heuristic strategies are adopted to approximate Bayesian optimization. For instance, the "win-stay, lose-shift" strategy (Steyvers, Lee & Wagenmakers 2009; Bonawitz et al. 2014) is straight from reinforcement learning (Robbins 1952; Sutton & Barto 1998): It seems more informative to call a spade a spade.

Finally, language throws up even more puzzling facts under optimal statistical principles. Linguists have long noticed that many linguistic expressions are simply ineffable. In a classic paper, Halle (1973) draws attention to morphological "gaps," the absence of inflected words for no apparent reason. For instance, there are about 70 verbs in Russian that lack an acceptable first person singular non-past form:

(4) *lažu 'I climb'
*pobežu (or *pobeždu) 'I conquer'
*deržu 'I talk rudely'
*muču 'I stir up'
*erunžu 'I behave foolishly'

Such defective gaps are quite widespread across languages; see Baerman, Corbett & Brown 2010 for a recent survey. Even in English, not known for its morphological complexity, the past participle of *stride* seems absent, as neither *strode* nor *stridden*, never mind *strided*, is generally unacceptable to native speakers (Pinker 1999). In most cases of gaps, the affected items are relatively few in number and concentrated in various corners of the morphological system. It is thus interesting to find a spectacular failure of language in the singular genitive system of Polish masculine nouns, which affects hundreds of items. There are two suffixes for the singular genitive (- *a* and - *u*), but neither passes the standard tests for default as in the English *-ed* past tense (Dąbrowska 2001). Of the two suffixes, one is necessarily more frequent, covering a statistical majority of words: In any case, the Bayesian optimality framework would favor one of the two options. Yet the Polish speakers we have surveyed are at a loss when asked to supply an ending for the following masculine nouns:

(5) drut 'wire'
rower 'bike'
balon 'balloon'
karabin 'rifle'
autobus 'bus'
lotos 'lotus flower'
Sometimes even the best is not good enough.

3.2 *Optimal inference and computational complexity*

From a computational perspective, the Bayesian detachment from psychological mechanisms is understandable. The problem of Bayesian inference is well known to be intractable (Dagum & Luby 1993; Chickering, Heckerman & Meck 2004), and even approximation algorithms are prohibitively expensive (e.g., Kwisthout, Wareham & van Rooij 2011). To see the scale of the problem, consider the Bayesian model of word learning that requires the child to select the optimal lexicon out of the space of all conceivable lexicons (Xu & Tenenbaum 2007; Frank Goodman & Tenebaum 2009). If there are n phonological words and m referents, the learner needs to consider all the possible mappings between all possible subsets of n and all possible subsets of m. A few minutes of child-directed speech in a crowded living room may result in more lexicons than the number of particles in the observable universe.

[7] Or, the measure of optimality, which is generally construed as expected payoff/penalty in the behavioral studies, would have to be reconceptualized.

As noted earlier, the Bayesian inference approach to language learning is not new, and its computational intractability was recognized by Horning (1969) himself in his pioneering work. Recent years have seen the development of efficient approximation methods (e.g., Geman & Geman 1984; Gilks, Richardson & spiegelhalter 1996; Blei, Ng & Jordan 2003) that enables practical optimization, although how well a Bayesian model works must be assessed on a case-by-case basis. The advantage of Bayesian models is not obvious.[8] Consider several comparisons in the domain of language learning. In syntactic acquisition, a hierarchical Bayesian learning model for English dative constructions (Perfors, Tenenbaum & Wonnacott 2010) does not perform better than a much simpler reinforcement learning model (Villavicencio et al. 2013). In the domain of morphology, Sirts & Goldwater (2013) developed a semisupervised Bayesian model that uses a corpus of morphologically annotated words. The training time is a week on 50 thousand words, but it provides very modest performance gain over a completely unsupervised and psycholinguistically inspired online learning model (Lignos 2010), which churns through almost 900 thousand words in half an hour. The Bayesian model of word learning proposed by Frank, Goodman & Tenenbaum (2009) carries out approximation inference over the set of all possible lexicons: 500 child-directed utterances took 500 hours of computing time, yet it is outperformed by a resource-constrained online model ("Pursuit") that considers no more than two meaning candidates at any instance of learning (Stevens et al. 2016), which processed the same 500 utterances in under one second. Interestingly, these alternatives to the Bayesian approach are all instances of reinforcement learning. It is conceivable that these psychological learning mechanisms, which are evolutionarily ancient and widely available through the animal kingdom, may have shaped the way language is structured and used. As a result, they are more adept at detecting linguistic signals in the stochastic environment of language use than idealized learning models conceived in an ecological vacuum.

I now turn to a particular application of Bayesian inference that has proved very attractive for a long-standing puzzle in language acquisition: indirect negative evidence as a strategy for the Subset Problem.

Indirect negative evidence (Chomsky 1981) is a powerful tool in language learning. As soon as Gold presented his overwhelmingly negative learnability results, he offered three strategies that may lead to positive learnability (Gold 1967:454):

(6) (a) restrictions on the space of hypotheses;
(b) negative evidence;
(c) restrictions on the distribution of learning data.

(6a) and (6b) are familiar: The former has been rigorously pursued in both generative grammar (Chomsky 1965, 1981; Prince and Smolensky 2004) and the identification of learnable language classes (Angluin 1982; Kanazawa 1998; Clark and Eyraud 2007), and the latter has been long recognized as infeasible, ineffective, and certainly unnecessary for the successful acquisition of language (Brown & Hanlon 1970; Heath 1983). (6c) is indirect negative evidence. In Gold's own words:

There is an a priori restriction on the class of texts which can occur, such as a restriction on the order of text presentation. The child may learn that a certain string is not acceptable by the fact that it never occurs in a certain context. This would constitute a negative instance. (Gold 1967: p454)

That is, absence of evidence can be used as evidence of absence.

Most language acquisition researchers recognize the appeal as well as pitfalls of indirect negative evidence. The Subset Problem, which has always received considerable attention in the formal studies of learning, clearly illustrates why indirect negative evidence would be desirable if justified. Consider the schematic illustration in Figure 1:

[8]Due to the stochastic nature of approximation methods, there are some practical difficulties in assessing the performance of Bayesian models: How long should one allow the search to run? How close does the best solution (when the search terminates) approach the true global optimum? The complexity of Bayesian inference can easily overwhelm most research groups' computing resources.

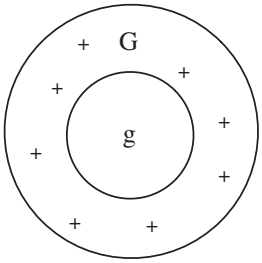

Figure 1. The target, and smaller, hypothesis *g* is a proper subset of the larger hypothesis *G*.

Suppose the target hypothesis is *g* but the child has instead conjectured a superset hypothesis *G*. They will never be contradicted by the learning data because every instance of *g* is also compatible with the more general *G*. Indirect negative evidence to rescue: If the learner fails to observe + forms that are expected under *G* but not *g*, then it can retreat back to *g* in response. However, the effectiveness of indirect negative evidence has always been questioned. For instance, Pinker's wide-ranging study notes that to use indirect negative evidence, one needs to specify "under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical" (1989:40): Every proposal under review is shown to be problematic, leading him to reject indirect negative evidence as "virtually a statement of the original learning problem."

Berwick's Subset Principle (1985) approaches the Subset Problem by requiring the learner to attend to the smaller/subset hypothesis first. But the computational complexity of using the Subset Principle is too high to be practical. In order to determine the subset-superset relations, the learner may need to compare the extensions of these hypotheses (i.e., the potentially infinite sets of strings they generate), which may not even be computable (Osherson, Stob & Weinstein 1986). In a study of a linguistically realistic domain of syntactic parameters, Fodor & Sakas (2005) find that the computational cost of detecting subset relations among a finite set of grammars is prohibitively high. Without a feasible means of computing the expectations of hypotheses — that the + expressions in the nonoverlapping region of $(G - g)$ — indirect negative evidence is unusable.

Probabilistic learning models such as Bayesian inference provides a natural formulation of indirect negative evidence. Failing to observe a sentence in a corpus of 1,000 words is one thing — one may not have got around to use it — but its absence in a corpus of 50 million words is more conspicuous and a potentially stronger clue for its ungrammaticality. But a probabilistic formulation of indirect negative evidence does not fundamentally change the nature of the computational problem (Niyogi 2006). In practice, the likelihood function needs to be specially formulated to produce the desired outcome: Again, many have followed Horning's formulation (1969) in treating longer sentences as in effect ungrammatical, and thus language is finite. Another widely used strategy in recent Bayesian learning models, including those applied to language acquisition, is the *Size Principle* (Tenenbaum & Griffiths 2001). Take Xu & Tenenbaum's word learning model (2007):

> Consider a hypothesis about the word's extension that picks out a finite set of K objects. The likelihood of picking any one object at random from this set of size K would be $1/K$ and for n objects (sampled with replacement), $1/K^n$. This reasoning leads to the following likelihood function:
>
> $$p(X|h) = \left[\frac{1}{size(h)}\right]^n$$
>
> if $x_i \in h$ for all i, and 0 otherwise. We refer to [the] equation as the *size principle* for scoring hypotheses: Hypotheses with smaller extensions assign greater probability than do large hypotheses to the same data, and they assign exponentially greater probability as the number of consistent examples increases. (Xu & Tenenbaum 2007:252)

As Xu & Tenenbaum note (252), this formulation is exactly the statistical formulation of the Subset Principle (Berwick 1985; Wexler & Culicover 1980) so we are back to the intractable problem of comparing the extensions of hypotheses. Note that in the context of word learning, the Size Principle here is also a statistical version of the mutual exclusivity constraint in word learning (e.g., Markman & Wachtel 1988), also known as the Uniqueness Principle (Wexler & Culicover 1980) or Principle of Contrast (Clark 1987), all of which fall under the rubric of indirect negative evidence (Pinker 1989:15). The Bayesian Size Principle provides a different formulation of indirect negative evidence, but it does not in general make it more computationally feasible.

In sum, the Bayesian inference model involves highly complex computational operations. Simulation results on several problems in language learning show that it does not provide superior performance than much simpler traditional models. However, the true test of learning models lies not in conceptual arguments but how it fares under realistic conditions of language acquisition that involve specific linguistic details. Although the general problem of using indirect negative evidence (e.g., via the Size Principle) is still intractable, there may be special cases where the extensions of the superset and subset grammars can be directly compared. In the remainder of this section, I use a concrete example to show that indirect negative evidence, regardless of how it is formulated, is ineffective when situated in a realistic setting of language acquisition.

3.3. *Optimal inference and the statistics of language*

The empirical case study concerns the properties of a certain class of English adjectives in (7):

(7) (a) The cat is asleep. ??The asleep cat.
(b) The boss is away. ??The away boss.
(c) The dog is awake. ??The awake dog.
(d) The child is alone. ??The alone child.
(e) The troops are around. ??The around troops.

The adjective here can be used predicatively like typical adjectives, but their attributive usage in a prenominal position inside an NP is ill-formed. These items all start with a schwa *a-*, which will be referred to as a-adjectives.

The a-adjectives offer a perfect application for indirect negative evidence. The learner is to acquire a subset hypothesis (g) that only allows predicative usage and to reject a superset hypothesis (G) that admits both predicative and attributive usage, which will allocate some probability mass for patterns such as "the asleep cat." For a corpus of English data (D) that contains a-adjectives, the expected attributive usage under G fails to show, thereby gradually lowering the likelihood of $P(D|G)$. Given enough time, the posteri probability $P(G|D)$ will lose out to $P(g|D)$, allowing the learner to adopt the correct, subset, hypothesis g. A search in the roughly 2 million words of child English in the public domain (MacWhinney 2000) yields about 2,300 tokens of a-adjectives: Not a single instance of attributive usage is found. Although the child data are pooled from a large number of subjects, the average age of the learner is just over 2;10. English-learning children evidently acquire the syntactic properties of a-adjectives very early, and the distributional evidence must be robustly available.[9]

Previous accounts of a-adjective acquisition follow exactly this argument even though they are not always explicitly formulated in a Bayesian framework. For frequency-based accounts such as Stefanowitsch (2008), if a sufficiently frequent adjective fails to appear in an attributive position, it would constitute as evidence for its ungrammaticality. Similarly, Boyd & Goldberg (2011) invoke the strategy of *statistical preemption*: Instead of using nonoccurrence as cues for ungrammaticality, the

[9]It is logically possible, though I believe unlikely, that the superset hypothesis of attributive usage is already ruled before the child has uttered a word. If so, then the onus is on the advocate of indirect negative evidence to demonstrate the reality of this very brief stage.

child assumes that the syntactic forms that realize the same meaning are mutually exclusive. As Pinker (1989) notes, this is also a form of indirect negative evidence as suggested by Wexler & Culicover (1980) and Clark (1987). More concretely, if the semantic expression "the asleep cat" is called for, and children consistently observe attestations in the relative clause (e.g., "the cat that is asleep"), then they will over time conclude that the attributive form is impossible. That is, the relative clause form preempts the attributive form.

It is easy to cast these proposals in a Bayesian formulation. In fact, recent studies of Bayesian learning have made use of *overhypothesis* (Kemp, Perfors & Tenenbaum 2007), adopted from Goodman's well-known discussion (1955), a form of knowledge abstracted over a class of individual items. Hierarchical Bayesian models can make inference over multiple levels of abstraction, from individual words and sentences to lexical casses and rules to universal constraints on language; see Perfors, Tenenbaum &Wonnacott 2010 for an application in syntactic learning. In the present study, one may wish to consider the a-adjectives as a natural class, and there are in fact linguistic and developmental evidence that children must do so; see Yang 2015 for extensive discussion. If so, the indirect negative evidence would be stronger and presumably more effective: The Size Principle will disfavor the superset hypothesis more significantly because "the number of consistent examples" (Xu & Tenenbaum 2007) is larger when evaluated on the entire class than on any specific lexical item.

But these proposals do not seem to work, whether or not the a-adjectives are evaluated as individual items or collectively as a class under the overhypothesis formulation. The fundamental problem can be started simply: The superset hypothesis cannot be effectively ruled out due to the statistical properties of child-directed English. As such, indirect negative evidence leads to very poor learning results such that the a-adjectives cannot be distinguished from typical adjectives.

Our empirical study draws from two sets of data from the public domain. The first part is a parsed corpus of approximately 180,000 child-directed sentences about 440,000 words in all (Pearl & Sprouse 2013). The parsed corpus facilitates the search for specific syntactic structures that will be important for evaluating adjectives. The second part is a 4.5 million- word corpus of child-directed English in the CHILDES database (MacWhinney 2000), which corresponds to about a year's data that many English-learning children receive (Hart & Risley 1995). Both sets of data contain exactly 12 a-adjectives:

(8) across, afraid, ahead, alike, alone, apart, around, ashamed, asleep, awake, aware, away

Consider the use of indirect negative evidence that exploits frequency and absence. All the 12 a-adjectives are relatively common words so as to appear in a modest 440,000 word corpus. But none of them is sufficiently frequent such that their failure to appear attributively would be remarkable. In the parsed corpus, there are 517 predicatively used adjectives, including the 12 a-adjectives, with an average frequency of 13.75. There are also 575 attributively used adjectives with a noun phrase, with an average frequency of 14.73: The a-adjectives, as expected, never appear there. The intersection of the two sets produces 198 adjectives that are used both predicatively and attributively, with an average frequency of 57.7. This is also expected, because higher frequency adjectives have more opportunities to be used in both constructions. But only one of the 12 a-adjectives (*afraid*, with a frequency of 73 out of 440,000) falls into this higher frequency range; many of the other 11 a-adjectives appear only once or twice, and their absence of attributive use is not at all conspicuous. At the same time, even a cursory search reveals that the corpus contains many typical adjectives (e.g., *careful, sorry, ready*) that are much more frequent than *afraid* but appear exclusively predicatively: Unlike the a-adjectives, these adjectives *can* appear attributively.

Evaluating the a-adjectives as a class under the overhypothesis approach (Kemp, Perfors & Tenenbaum 2007, Perfors, Tenenbaum &Wonnacott 2010) does not help. In the parsed corpus, there are collectively 143 a-adjectives. But even as a class, the a-adjectives are still less frequent than typical adjectives such as *sorry* and *careful*, which do not appear attributively: Thus the Size Principle

cannot accurately identify the a-adjectives as a class either. In fact, even in the 4.5 million-text corpus, a part-of-speech tagger shows that the a-adjectives are, collectively, still less frequent than *sorry* and *careful*, which still do not appear attributively.[10]

The strategy of statistical preemption by the relative clause paraphrase (Boyd & Goldberg 2011) fares far worse. I direct the reader to Yang 2015 for details. The problem is that adjectives are very rarely used in relative clauses to modify noun phrases (on average, only once almost every 3,000 utterances). Only 3 out of the 12 a-adjectives are used in a relative clause at all in 4.5 million words of child-directed English, and there are also many typical adjectives that, when modifying noun phrases, are exclusively used in the relative clause form. The rates of false positives and false negatives under the paraphrase preemption are extremely high.

Taken together, it is very unlikely for indirect negative evidence to reveal the syntactic properties of the a-adjectives. Once the superset hypothesis (attributive plus predicative usage) is introduced, there is no sufficient statistical evidence to rule it out. The failure of indirect negative evidence can be attributed to the inherent statistical distribution of language. Under Zipf's law, which applies to linguistic units (e.g., words) as well as their combinations (e.g., N-grams, phrases, rules; see Yang 2013), it is very difficult to distinguish low probability events and impossible events. In the present case, the statistical distribution of language cannot separate the a-adjectives that resist attributive usage by design from the typical adjectives that fail to show attributive usage by chance.

The correct identification of a-adjectives, which I develop in a recent article (Yang 2015), is to turn indirect negative evidence on its head. Note that the superset hypothesis is defined in terms of the distributional *differences* between a-adjectives and typical adjectives, but these differences are undetectable in realistic linguistic input as we have just seen. The alternative strategy is a positive one, as it exploits the distributional *similarities* between a-adjectives and other linguistic units, specifically locative participles such as *on, off, here, there*, etc., and prepositional phrases, both of which resist attributive usage ("the light is on/*the on light," "the car is on the road/*the on the road car"). Under this approach, the superset hypothesis is never available to the learner, and there is no need to rule it out.[11]

To summarize the discussion thus far: Neither generative grammar nor the Marrian levels, two highly successful paradigms in cognitive science, provides philosophical and methodological support for idealized models. None of these specific criticism in the domain of language — variation, change, gaps, a-adjectives — is necessarily decisive: The Bayesian model can be amended with other assumptions, and even formal complexity arguments can be circumvented by appealing to the current ignorance of how the brain works and what the child is ultimately capable of computing. But I hope to have made clear that the empirical aspects of language and language acquisition do pose significant challenges: At the minimum, it is worth pursuing alternative solutions.

4. A Minimalist evaluation metric

Since the beginning of formal learning research (Gold 1967), the Subset Problem has always been formidable. Naturally, one way to limit the damage is to ensure the Subset Problem never, or rarely, arises in the acquisition of language. Constraining the space of possible grammar has proved successful from both typological (Baker 2001) and developmental perspective (Crain & Thornton 2000; Yang 2002b). In the case of syntactic parameters, evidence for one value does seem to constitute evidence against the opposite value. Another strategy is to provide empirical evidence that the subset grammar is in fact a default option from which the learner starts, and the larger grammar is only accessed upon the presentation of data that contradict the small grammar. A good

[10]Presumably, there are more than two such typical adjectives, but here a cursory evaluation is sufficient.

[11]Can the "a-adjective as PP" hypothesis developed in Yang (2015) be cast in terms of Bayesian inference? Absolutely: The Bayesian framework is extremely flexible. But doing so entails the acknowledgment that the Bayesian formulation of indirect negative evidence, which is its central appeal in the study of linguistic generalization, is no longer at play, and the Bayesian framework is superfluous.

deal of formal and empirical research in language development is devoted to working out such nested hierarchies in the grammatical options (Chomsky 1981; Prince & Smolensky 2004) available to the learner (Berwick 1985; Roeper & Williams 1987; Dresher & Kaye 1990), including the acquisition of semantics and pragmatics (Crain 2012). The child would not overgeneralize, and the Subset Problem does not arise.

But innate parameters and defaults cannot solve all of the learning problems. There are clearly rules and generalizations that are language or even lexically specific — such as the a-adjectives and the dative constructions — and must be acquired on an inductive basis. Nevertheless, the reservation with indirect negative evidence has prompted researchers to reformulate the hypotheses available to the child such that the superset hypothesis is not postulated and the Subset Problem is a nonissue; see, for instance, the treatment of a-adjectives in Yang 2015 briefly reviewed previously. This impetus seems lost with the rise of probabilistic learning models. The apparent ease with which indirect negative evidence can be incorporated into the Bayesian inference framework (e.g., the Size Principle) gives the false impression that the Subset Problem has been resolved. But as we have seen in the case of a-adjectives, the problem remains intractable, and inherent, due to the statistical distribution of language.

In some cases of language learning, however, it is clear that the superset hypothesis *is* entertained, and children do backtrack from it. The most clear, and best studied, case is the acquisition of the dative constructions in English first discussed in a classic paper by C. L. Baker (Baker1979).

(9) (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* (9a) and *promise* (9b) 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. How to draw just the right level of dative generalizations has been known as Baker's Paradox: It has been a major problem in theoretical syntax as well as language acquisition, with a substantial cross-linguistic literature devoted to its resolution.

It has been long noted that children produce overgeneralization errors during the course of dative acquisition. According to Gropen et al. (1989)'s quantitative analysis of English children's production data, about 5% of all dative constructions are overregularizations such as those in (10) (see also Bowerman 1988):

(10) I said her no.
I whisper you something.
Mattie demonstrated me that yesterday.

These errors presumably will be eliminated as children grow older. So the superset hypothesis is cognitively available, developmentally accessed, and from which the child must gradually retreat. The Subset Problem needs to be solved — *without* indirect negative evidence.

In what follows, I will briefly review the *Tolerance Principle*, a learning theoretic device that guides the child to discover productive processes in language. Evaluating child-directed English data, I show that the Tolerance Principle provides a simple and effective solution to Baker's Paradox.

4.1. *The Tolerance principle*

The Tolerance Principle was initiated to address an apparently simple problem: How does an English-learning child recognize that the *-ed* rule is the productive rule applicable to an unlimited number of verbs? We follow a strong intuition that is shared by almost all researchers of linguistic productivity (Aronoff 1976): Namely, a rule is productive if it applies to most items to which it's applicable. But if a rule has too many exceptions, then the learner will decide against its productivity. This naturally invites an Evaluation Metric based approach: The learner will value highly a rule that has fewer exceptions than one that has more exceptions, and there exists a "tipping point", so to speak, such that the exceptions to a rule overwhelm its productivity.

The critical question, then, is how many exceptions are enough. In computational terms, there are two, and only two, measures that figure into complexity considerations: space and time. Both metrics are in principle valid for the study of language, and the choice must be based on their empirical merits. Recently, there has been much interest in the distributional learning of language, often couched in the *Minimum Description Length* (MDL) framework (Rissanen 1978) or other formally equivalent or similar approaches (Goldsmith 2001; Tenenbaum & Griffiths 2001; Chater & Vitányi 2007), following developments in natural language processing (e.g., de Marcken 1996). The MDL framework views the grammar as a data compression device, and is indeed an Evaluation Metric used in the earliest work in generative grammar (Chomsky 1951, 1955). It strives to minimize the structural description of data, to eliminate redundancies, and to obtain the simplest and most elegant statement of the grammar. For instance, the regular *-ed* rule obviously contributes to the economy of storage such that thousands of regular verbs needn't be individually listed for past tense. One can conceivably devise a scheme such that the postulation of a rule is justified when it achieves more spacesaving than lexical listing.

At the present time, however, an MDL approach seems a recipe for ad hocism. Because we understand precious little about the constraints on linguistic memory and computation, we have no principled basis to evaluate the cost of storing a word, or the cost of storing a rule, or the cost of storing the mapping between words and the rules that apply to them. Without having an independent measure of these quantities, we do not have a well-motivated currency to quantify the minimization of storage under different organizations of language. To make the matter worse, we have few concrete clues on the child learner's computational power, without which we cannot be certain how much compression can be squeezed out of the data. If pursued to the limit, the "M" in MDL may yield highly abstract descriptions of language such that the storage for words is minimized but the derivational complexity of the inflected forms is maximized. As it stands, the MDL approach does not move us closer to an empirical understanding of language acquisition.

Instead of space, consider time. The Tolerance Principle provides an evaluation metric that quantifies real-time language processing. In particular, we suggest that the learner always chooses the more efficient, i.e., faster, organization of word formation. A productive rule is postulated if it speeds up language processing; otherwise the learner resorts to lexical listing. In a sense, the Tolerance Principle can be viewed as a "third factor" component in the Minimalist approach to language:

> (a) principles of data analysis that might be used in language acquisition and other domains; (b) principles of structural architecture and developmental constraints that enter into canalization, organic form, and action over a wide range, including principles of efficient computation, which would be expected to be of particular significance for computational systems such as language. It is the second of these subcategories that should be of particular significance in determining the nature of attainable languages. (Chomsky 2005:6)

Specifically, we conjecture that the Elsewhere Condition (Anderson 1969; Kiparsky 1973), a general principle for the organization of linguistic information, simultaneously acts as an algorithmic process of linguistic processing. According to the Elsewhere Condition, also known as the Blocking Principle in psycholinguistics (Pinker 1999), exceptions to a rule are evaluated prior to the application of the rule. For instance, when English speakers are to inflect the past tense of a verb, they first

examine if the verb is one of the irregulars: If so, an irregular past tense will be generated; otherwise the regular *-ed* rule applies. Thus, as the number of exceptions to a rule increases, rule-following items will have to "wait" before they are inflected. By contrast, the cost of lexical listing without a productive rule is simply the time complexity of processing each item weighted by its frequency. The reader is referred to Chapter 3 of Yang 2016 for review of the psycholinguistic literature that establishes the serial nature of the Elsewhere Condition in online language processing. By comparing the cost of full listing and the cost of a rule plus listing exceptions, we can derive the threshold at which it is cheaper, i.e., faster, to maintain productivity. Under general assumptions about word frequencies, it is possible to prove the following:

(11) **Tolerance Principle**:

If R is a productive rule applicable to N candidates in the learning sample, then the following relation holds between N and e, the number of exceptions that could but do not follow R:

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

Note that the Tolerance Principle is a "parameter-free" learning model. Neither the researcher nor the child needs to calibrate the model before applying the learning data, and it always makes a specific prediction for any two values of N and e: The rule is either productive or not, which strongly corresponds to the cross-linguistic acquisition of productivity. In this sense, the Tolerance Principle is quite unlike current Bayesian models of learning as well as generalization models from psychology (e.g., Anderson 1991; Nosofsky, Palmeri & McKinley 1994), all of which require the fitting of parameter values against data or experimental results.

In Yang (2016), I present dozens of empirical case studies to show that the Tolerance Principle makes accurate predictions about where productivity arises in language and where it collapses. For productivity to emerge, the slow growth of the $N/\ln N$ function suggests that there must be an overwhelming number of rule-following items in order to overcome exceptions. Recent experiments using the artificial language paradigm with young children have produced near-categorical support for the numerical predictions of the Tolerance Principle (Schuler, Yang & Newport 2016). In one condition, young children learn nine novel nouns: Five share a plural suffix, and the other four are idiosyncratic. In another condition, the mixture is three with the same suffix and six idiosyncratic ones. The choices of 5/4 and 3/6 are deliberate: The Tolerance Principle predicts the productive extension of the shared 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 shared 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. Here I give only two simple applications to illustrate the mechanics of the Tolerance Principle in action.

First, consider the acquisition of English past tense. Suppose an English learner knows $e = 120$ irregular verbs; the productivity of the *-ed* rule is guaranteed only if there are many more regular verbs. Specifically, there must be N verbs, including both regulars and irregulars, such that $\theta_N = N/\ln N \geq 120$. The minimum value of N is 800. In other words, if there are at least 680 regular verbs in English, the *-ed* rule can tolerate 120 irregular verbs. Since there are clearly more than 680 regular verbs in English, the learner will be justified to conclude that the *-ed* rule is productive and can be extended to novel items (Berko 1958).

In fact, the prediction can be made more precisely, potentially at the individual level. After all, productivity is determined by the numerical values of N and e, which are determined by a child learner's vocabulary — and individuals' vocabulary size and composition necessarily differ to some

extent. Consider "Adam," the poster child for English past tense acquisition (Pinker 1999). Adam produced the first instance of overregularization error — "What dat feeled like?" — at the age of 2;11. In the transcript of almost a year prior to that point, not a single irregular verb past tense was used incorrectly. If we take overregularization as the onset of the productivity of *-ed*, then it must be the case that at this point, Adam has acquired a sufficiently large number of regulars to overwhelm the irregulars. To test this prediction, I extracted every verb stem in Adam's speech until 2;11. There are $N = 300$ verbs in all, out of which $e = 57$ are irregulars, which is very close to the predicted $\theta_{300} = 53$, and the small discrepancy may be due to the undersampling of the regular verbs that tend to be lower in frequency and are more likely to be missed in a modest sample. The critical point to note here is that Adam apparently needed a filibuster-proof majority of regular verbs to acquire the *-ed* rule: This is strongly consistent with the predictions of the Tolerance Principle.

Second, consider the absence of a productive suffix in the Polish masculine singular genitive system reviewed earlier (3.1). Recall that there are two suffixes, *-a* and *-u*, but neither is productive. According to the Tolerance Principle, it must be the case that neither *-a* nor *-u* applies to a sufficiently large number of nouns so that one becomes the productive rule, and the other applies only to a list of exceptions. To test this prediction, Margaret Borowycz and I examined the distribution of Polish nouns from the child-directed Polish as made available in the CHILDES corpus (MacWhinney 2000). There are 837 masculine nouns that take *-a* in the singular genitive and 516 that take *-u*. For a total of $N = 837 + 516 = 1,353$ nouns, a productive suffix can only emerge if there are fewer than $\theta_{1353} = 187$ exceptions, which is clearly not met here. Thus, we correctly predict that the Polish learner can do no more than learning the suffix for every single noun: The absence of a productive suffix means that for nouns whose inflected forms are not already in language use, speakers will be unsure about which suffix is applicable as shown in (3.1) .

4.2 *From necessity to sufficiency to retreat*

Before considering the application of the Tolerance Principle to the acquisition of dative constructions, let us consider a hypothetical example, the kind that is likely to arise in any instance of inductive learning problems.

Imagine having shipwrecked on a desert island. If you come across 10 exotic species, 7 of which are tame and friendly, you'll probably assume the next encounter is harmless: Seven of 10 seems pretty good odds. But 1 of 10? 2 out of 10? You'd be well-advised to proceed with caution.

Generalizations evidently require weight of evidence: Seven exemplars appear sufficient, but one or two probably won't do. This is not to say that the sufficiency of evidence is fail-proof: The other three species, with which we have had no direct experience, may well turn out to be dangerous. Nor does sufficiency guarantee permanence: If an additional 20 species have come our way, suddenly 7 out of 30 no longer inspire confidence.

In my view, while the Subset Problem in Baker's Paradox is a genuine challenge, its severity has been exaggerated by linguists and psycholinguists. Indeed, if the child were to receive only the four verbs in (4; *give, assign, donate*, and *promise*) and still must generalize correctly, the Subset Program would be intractable. But as the desert island example shows, it would be absurd for the child to generalize the properties attested in two examples (*give* and *assign*) to an entire class. In the spirit of the Tolerance Principle, we propose the following corollary:

(12) **The Principle of Sufficiency**:

Let R be a generalization over N items in the learning sample, of which M items are attested to follow R. R can be extended to all N items if and only iff:

$$N - M < \theta_N \quad \text{where} \quad \theta_N := \frac{N}{\ln N}$$

Before the positive evidence is sufficient, that is, M sits below the sufficiency threshold, the learner lexicalizes all M items and does not generalize beyond them. Without any kind of generalization, the problem of overgeneralization does not even arise. That is, Baker's Paradox involving the verb *donate* as in (4) is a nonissue unless the child has observed a significant majority of similar verbs attested in the double object and *to*-dative construction — which, as we will see, probably won't happen during the first few years of language acquisition. Only when M crosses the sufficiency threshold does R become a truly productive rule.

The Principle of Sufficiency has a built-in mechanism for retreating from overgeneralization. Suppose $N = 50$ and the learner has accumulated $M = 40$ instances to warrant a generalization ($50 - 40 = 10 < \theta_{50} = 12$). But further down the road, N starts to increase again. Suppose that N has grown to 70 with 20 more items but M has been standing still at 40: Now the rule will cease to be productive ($70 - 40 = 30 > \theta_{70} = 16$), and the learner will lexicalize all 40 items and the once productive generalization will be abandoned. The dynamics of learning under the Principle of Sufficiency is exactly the same as under the Tolerance Principle, which we have seen extensively studied in the preceding pages.

Although the formalisms for the Principle of Sufficiency and Tolerance Principle are similar, an important logical and empirical difference remains. The Tolerance Principle keeps track of exceptions to a rule R, that is, attested items that explicitly defy R. The Principle of Sufficiency, by contrast, asserts that unless the sufficiency threshold has been crossed, the learner is in a state of ambivalence regarding the $(N - M)$ items that S/he knows nothing about: "I'm not sure" is an OK answer. That is, in contrast to the use of indirect negative evidence, the Principle of Sufficiency does not conclude that unattested forms are ungrammatical — or grammatical, for that matter. Absence of evidence is *not* evidence of absence.

The Principle of Sufficiency provides a straightforward solution for the acquisition of datives; see Yang (2016) for details. For the ease of presentation, we will only consider the acquisition of the double object construction: children's initial overgeneralization ("she said me no") and then subsequent retreat from it. Following the methodological practice of situating learning models in actual language acquisition data, we examine the distribution of verbs that participate in the double object construction ("verb NP NP"). In a 5 million-word corpus of child-directed English, I identified a total of 42 verbs used in the double object construction. Of these, 38 have a very clearly identifiable semantics of "caused possession," which we assume is identifiable if the learner is equipped with a suitable set of semantic primitives (e.g., Grimshaw 1990; Jackendoff 1990). The four exceptions, well below the threshold $\theta_{42} = 11$, do not convey caused possession but are all performative verbs (*call, consider, name*, and *pronounce*). Thus, it seems that the semantic condition necessary for double object construction, now widely recognized by many researchers (e.g., Gropen1989; Pinker1989; Levin1993; Pesetsky1995; Krifka1999), needn't be stated as a UG primitive but can be acquired from the language-specific data under a principled learning model.

(13) In English, 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 (13), in trying to establish the validity of caused possession as a *sufficient* condition for the double object condition. Again pouring over the child-directed English corpus, we establish the set of caused possession verbs (N) to see if the subset of $M = 38$, which are actually used in the double object construction, constitutes sufficient evidence for generalization to the entire set. In the present case, there are in fact an additional 11 verbs that belong to the semantic class but fail to appear in a double object construction:

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

This is an interesting list. For some of the items in (4.2), e.g., *introduce* and *say*, the double object construction is ungrammatical:

(15) *John introduced the kids a new dish.
*John said Bill something mean.

Whereas others do allow the double object construction but did not have the opportunity to do so:

(16) John shipped Bill his purchase.

But of course the child does not know *why* the verbs in (14) fail to appear in the double object construction (ungrammaticality or lack of opportunity). Nevertheless, a sufficiently large number of verbs, namely $M = 38$, are able to trigger the following generalization on the basis of the child-directed English corpus: $11 < \theta_{49} = 12$ satisfies the Principle of Sufficiency.

(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 in (10) such as "She said me no," as well as experimental evidence that children as young as 3;00 have productive usage of the dative constructions upon learning a novel verb with the appropriate semantic properties (e.g., Conwell & Demuth 2007).

Now the critical part of the Subset Problem: How does the child retreat from the (over) generalization in (17)? After exhausting the child-directed English data in the public domain, I do not believe the problem is likely to arise for a young child. As noted by Gropen et al. 1989, verbs such as *donate* are far too rare to be learned early. Indeed, latinate vocabulary, which includes *donate* and other verbs at the heart of the Subset Problem, is generally acquired in a school setting (Tyler & Nagy 1989; Jarmulowicz 2002).

According to Levin's encyclopedic survey (1993:45–48), there are in fact *more* caused possession verbs (138) that resist the double object construction that those that allow it (115): 115 out of 253 certainly does warrant any productive generalization. However, many of these verbs are rare and probably will not enter into the calibration of productivity for most English learners. By making suitable frequency estimates (see Yang 2016 Section 6.3 for details), one can "trim" the verbs to a relatively common set of 92. Still, only 52 can be expected to appear in the double object construction.[12] But this falls short of the threshold imposed by the Principle of Sufficiency: A productive generalization for 92 items requires $92 - \theta_{92} = 72$ positive instances. The generalization in (17) is no longer productive, and the learner will now need to lexicalize every verb that appears in the double object construction: Errors such as those in (10) will be eliminated as the learner, after acquiring a sufficiently large vocabulary, successfully retreats from a superset hypothesis.

There are additional moves available to the learner. For instance, the learner may subdivide the 92 caused possession verbs into finer classes such as "ballistic motion," "manner of speaking," etc. (Gropen et al. 1989; Pinker 1989; Levin 1993; Pesetsky 1995), assuming that the semantic attributes associated with such categorization, which may include both linguistic and nonlinguistic factors, are

[12]Unsurprisingly, the verbs that allow the construction are considerably more frequent than those that do not. This accounts for the justified productivity of (17) when evaluated on a corpus of child-directed speech that contains the relatively high-frequency words.

accessible to the learner. The Principle of Sufficiency can be applied recursively, detecting productive generalizations in some of these subclasses. One such class may be verbs of "telecommunication": *fax, phone, telegram*, etc., leading to the immediately availability of the construction when verbs such as *email* and *text* entered into the English lexicon. Certain phonological and morphological constraints of verbs (Green 1974; Oehrle 1976) can be successfully identified.

5. Conclusion

Although much of this article has been a critique of the Bayesian program, my main purpose has been to articulate a clear, and I think promising, direction for language acquisition. In the age of big data and big machines, it is still important to regard learning as a psychological theory within the means of plausible cognitive processes and realistic input conditions. One should of course always strive for the most general solution, but not at the expense of the specific issues at the heart of the empirical domain. As I have tried to make clear, the distributional properties of language — inherent variation, morphological gaps, Zipf's Law, etc. — poses many interesting puzzles that do not easily yield to generic formulations of learning such as optimal statistical inference. Further, the ecological condition of language acquisition (e.g., no negative evidence) restricts the computational resources that would be available in other domains. Especially acute is the Subset Problem. The Bayesian approach gives the appearance of a principled solution, but the additional machineries are of no benefit when evaluated in a realistic setting of language acquisition.

At the same time, I submit that the Tolerance Principle, and its Sufficiency corollary, provide a unified treatment for productivity and generalization in language. All things being equal — even though they rarely are — simpler learning models that bridge the Marrian levels should be preferred over those that only operate in an idealized setting. By maintaining the constraint of simplicity and building on the commitment to empirical issues in language learning and use, I hope to have provided a 21st-century upgrade for the Evaluation Metric, a traditional concept from the birth of modern linguistics. May it live long and prosper.

Acknowledgment

I would like to thank Noam Chomsky, Alex Clark, Stephen Crain, Randy Gallistel, Steve Isard, Mark Johnson, Norbert Hornstein, and Lisa Pearl for helpful discussions of the materials presented here.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409.
Anderson, S. R. (1969). *West Scandinavian vowel systems and the ordering of phonological rules*. PhD thesis, MIT.
Angluin, D. (1982). Inference of reversible languages. *Journal of the ACM*, 29(3):741–765.
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.
Baker, M. (2001). *The atoms of language: The mind's hidden rules of grammar*. Basic Books, New York.
Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2–3):150–177.
Berwick, R. (1985). *The acquisition of syntactic knowledge*. MIT Press, Cambridge, MA.
Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.
Bonawitz, E., Denison, S., Gopnik, A., and Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, 74:35–65.
Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In Hawkins, J. A., editor, *Explaining language universals*, pages 73–101. Basil Blackwell, Oxford.
Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389.

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.
Brown, R. (1973). *A first language: The early stages*. Harvard University Press, Cambridge, MA.
Brown, R. and Hanlon, C. (1970). Derivational complexity and the order of acquisition in child speech. In Hayes, J. R., editor, *Cognition and the development of language*, pages 11–53. Wiley, New York.
Bush, R. R. and Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 68(3):313–323.
Chater, N. and Vitányi, P. (2007). Ideal learning of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163.
Chickering, M., Heckerman, D., and Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330.
Chomsky, N. (1951). Morphophonemics of Modern Hebrew. Master's thesis, University of Pennsylvania. Published by Garland, New York, 1979.
Chomsky, N. (1955). The logical structure of linguistic theory. Ms., Harvard University and MIT. Revised version published by Plenum, New York, 1975.
Chomsky, N. (1957). *Syntactic structures*. Mouton, The Hague.
Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.
Chomsky, N. (1981). *Lectures in government and binding*. Foris, Dordrecht.
Chomsky, N. (1986). *Knowledge of language: Its nature, origins, and use*. Praeger, New York.
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.
Clark, A. and Eyraud, R. (2007). Polynomial identification in the limit of context-free substitutable languages. *Journal of Machine Learning Research*, 8:1725–1745.
Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B., editor, *Mechanisms of language acquisition*, pages 1–33. Erlbaum, Hillsdale, NJ.
Conwell, E. and Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, 103 (2):163–179.
Crain, S. (2012). *The emergence of meaning*, volume 135. Cambridge University Press.
Crain, S. and Thornton, R. (2000). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. MIT Press, Cambridge, MA.
Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122 (3):306–329.
D abrowska, E. (2001). Learning a morphological system without a default: The Polish genitive. *Journal of Child Language*, 28(3):545–574.
Dagum, P. and Luby, M. (1993). Approximating probabilistic inference in bayesian belief networks is np-hard. *Artificial intelligence*, 60(1):141–153.
de Marcken, C. (1996). *Unsupervised language acquisition*. PhD thesis, MIT.
Dillon, B., Dunbar, E., and Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, 37(2):344–377.
Dresher, B. E. and Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition*, 34:137–195.
Eberhardt, F. and Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, 21(3):389–410.
Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological review*, 57(2):94.
Fazly, A., Alishahi, A., and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.
Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778.
Fodor, J. D. and Sakas, W. G. (2005). The subset principle in syntax: Costs of compliance. *Journal of Linguistics*, 41 (3):513–569.
Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.
Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6(6):721–741.
Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25(3):407–454.
Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.
Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447–474.
Goldsmith, J. (2001). Unsupervised learning of morphology of a natural language. *Computational Linguistics*, 27 (2):153–198.
Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.
Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press, Cambridge, MA.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., and Hamrick, J. B. (2015). Relevant and robust a response to marcus and davis (2013). *Psychological science*, 26(4):539–541.
Green, G. M. (1974). *Semantics and syntactic regularity*. Indiana University Press.
Grimshaw, J. (1990). *Argument structure*. MIT Press, Cambridge, MA.
Grimson, W. E. L. (1981). *From images to surfaces: A computational study of the human early visual system*. MIT press.
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.
Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.
Han, C.-h., Musolino, J., and Lidz, J. (2016). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences*, 113(4):942–947.
Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing, Baltimore, MD.
Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge University Press, Cambridge.
Horning, J. J. (1969). A study of grammatical inference. Technical report, Stanford University, Stanford, CA.
Huybregts, R. (1984). The weak inadequacy of context-free phrase structure grammars. *Van periferie naar kern*, pages 81–99.
Jackendoff, R. S. (1990). *Semantic structures*. MIT Press, Cambridge, MA.
Jarmulowicz, L. (2002). English derivational suffix frequency and children's stress judgements. *Brain and Language*, 81 (1–3):192–204.
Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Bayesian inference for pcfgs via markov chain monte carlo. In *HLT-NAACL*, pages 139–146.
Jones, M. and Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? on the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(?):169–231.
Julesz, B. (1971). *Foundations of cyclopean perception*. University of Chicago Press, Chicago.
Kanazawa, M. (1998). *Learnable Classes of Categorial Grammars*. Center for the Study of Language and Information, Stanford, CA.
Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 10(3):307–321.
Kiparsky, P. (1973). Elsewhere in phonology. In Anderson, S. R. and Kiparsky, P., editors, *A festschrift for Morris Halle*, pages 93–106. Holt, Rinehart and Winston, New York.
Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.
Köhne, J., Trueswell, J. C., and Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In *Proceedings of the 35th Annual meeting of the* Cognitive Science Society. Austin, TX: *Cognitive Science Society*.
Krifka, M. (1999). Manner in dative alternation. In *West Coast Conference on Formal Linguistics*, volume 18, pages 260–271.
Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1(3):199–244.
Kroch, A. (1995). Dialect and style in upper class Philadelphia. In Guy, G., Feagin, C., Schiffrin, D., and Baugh, J., editors, *Towards a social science of language: Papers in honor of William Labov*, volume 1, pages 23–45. John Benjamins, Philadelphia.
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.
Kwisthout, J., Wareham, T., and van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, 35(5):779–784.
Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.
Labov, W. (1989). The child as linguistic historian. *Language Variation and Change*, 1(1):85–97.
Labov, W. (2007). Transmission and diffusion. *Language*, 83(2):344–387.
Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.
Lignos, C. (2010). Learning from unseen data. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 35–38.
Lignos, C. (2013). *Modeling words in the mind*. PhD thesis, University of Pennsylvania.
MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.
Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge.
Marcus, G. F. and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, 24(12):2351–2360.
Markman, E. M. and Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2):121–157.
Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, MA. Originally published in 1982 by Freeman, San Francisco, CA.

Marr, D., Palm, G., and Poggio, T. (1978). Analysis of a cooperative stereo algorithm. *Biological Cybernetics*, 28 (4):223–239.
Marr, D. and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194(4262):283–287.
Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328.
Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019.
Miller, K. L. and Schmitt, C. (2012). Variable input and the acquisition of plural morphology. *Language Acquisition*, 19 (3):223–261.
Mysln, M. and Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition*, 147:29–56.
Niyogi, P. (2006). *The computational nature of language learning and evolution*. MIT Press, Cambridge, MA.
Niyogi, P. and Berwick, R. C. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106(25):10124–10129.
Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101(1):53.
O'Donnell, T. (2015). *Productivity and reuse in language*. MIT Press, Cambridge, MA.
Oehrle, R. T. (1976). *The grammatical status of the English dative alternation*. PhD thesis, Massachusetts Institute of Technology.
Osherson, D. N., Stob, M., and Weinstein, S. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists*. MIT Press, Cambridge, MA.
Pearl, L. and Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.
Perfors, A., Tenenbaum, J. B., and Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118 (3):306–338.
Perfors, A., Tenenbaum, J. B., and Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(3):607–642.
Pesetsky, D. (1995). *Zero syntax: Experiencer and Cascade*. MIT Press, Cambridge, MA.
Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press, Cambridge, MA.
Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books, New York.
Pintzuk, S. (1999). *Phrase structures in competition: Variation and change in Old English word order*. Routledge.
Prince, A. and Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. MIT Press, Cambridge, MA.
Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.
Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.
Roberts, J. and Labov, W. (1995). Learning to talk Philadelphian: acquisition of short *a* by preschool children. *Language Variation and Change*, 7:101–112.
Roeper, T. and Williams, E. (1987). *Parameter setting*. Springer, Berlin.
Sakas, W. G. and Fodor, J. D. (2001). The structural triggers learner. In Bertolo, S., editor, *Language acquistion and language learnability*, pages 172–233. Cambridge University Press.
Sakas, W. G. and Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, 19(2):83–143.
Sankoff, G. and Blondeau, H. (2007). Language change across the lifespan: /r/in Montreal French. *Language*, 83 (3):560–588.
Santorini, B. (1992). Variation and change in yiddish subordinate clause word order. *Natural Language & Linguistic Theory*, 10(4):595–640.
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.
Shi, R. and Melançon, A. (2010). Syntactic categorization in French-learning infants. *Infancy*, 15(5):517–533.
Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8(3):333–343.
Sirts, K. and Goldwater, S. (2013). Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.
Slobin, D. I. (1997). *The crosslinguistic study of language acquisition*, volume 4. Psychology Press.
Smith, J., Durham, M., and Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change*, 21(1):69–95.
Sober, E. (1975). *Simplicity*. Oxford University Press, New York.
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. (2016). The pursuit of word meanings. In *Cognitive Science*. doi: 10.1111/cogs.12416.

Steyvers, M., Lee, M. D., and Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179.
Suppes, P. (1966). Concept formation and bayesian decisions. In Hintkikka, J. and Suppes, P., editors, *Aspects of inductive logic*, pages 21–48. North-Holland.
Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge University Press.
Tardif, T., Shatz, M., and Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language*, 24(3):535–565.
Taylor, A. (1994). Variation in past tense formation in the history of English. In Izvorski, R., Meyerhoff, M., Reynolds, B., and Tredinnick, V., editors, *Penn Working Papers in Linguistics 1*, pages 143–158. Penn Linguistics Club, Philadelphia.
Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity and bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.
Trueswell, J. C., Medina, T. N., Hafri, A., and Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.
Tyler, A. and Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, 28(6):649–667.
Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, 22(4):562.
Villavicencio, A., Idiart, M., Berwick, R. C., and Malioutov, I. (2013). Language acquisiiton and probabilistic models: Keeping it simple. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1321–1330.
Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W., editor, *Directions for historical linguistics: A symposium*, pages 95–195. University of Texas Press, Austin.
Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.
Wexler, K. and Culicover, P. (1980). *Formal principles of language acquisition*. MIT Press, Cambridge, MA.
Willshaw, D., Dayan, P., and Morris, R. (2015). Memory, modelling and marr: a commentary on marr (1971) 'simple memory: a theory of archicortex'. *Phil. Trans. R. Soc. B*, 370(1666):20140383.
Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245.
Yang, C. (2000). Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250.
Yang, C. (2002a). *Knowledge and learning in natural language*. Oxford University Press, Oxford.
Yang, C. (2002b). A principle of word storage. Manuscript: Yale University.
Yang, C. (2005). On productivity. *Linguistic Variation Yearbook*, 5(1):333–370.
Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327
Yang, C. (2015). 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., Ellman, A., and Legate, J. A. (2015). Input and its structural description. In Ott, D. and Gallego, A., editors, *50th anniversary of Noam Chomsky's Aspects of the Theory of Syntax*. MITWPL.
Yu, C. and Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420.