COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 34 (2010) 972–1016
Copyright © 2010 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01117.x

The Logical Problem of Language Acquisition: A Probabilistic Perspective

Anne S. Hsu,[a] Nick Chater[b]

[a]*Department of Cognitive, Perceptual, and Brain Sciences, University College London*
[b]*Department of Cognitive, Perceptual, and Brain Sciences and ESRC Centre for Economic Learning and Social Evolution (ELSE), University College London*

Received 12 August 2008; received in revised form 23 December 2009; accepted 30 March 2010

Abstract

Natural language is full of patterns that appear to fit with general linguistic rules but are ungrammatical. There has been much debate over how children acquire these ''linguistic restrictions,'' and whether innate language knowledge is needed. Recently, it has been shown that restrictions in language can be learned asymptotically via probabilistic inference using the minimum description length (MDL) principle. Here, we extend the MDL approach to give a simple and practical methodology for estimating how much linguistic data are required to learn a particular linguistic restriction. Our method provides a new research tool, allowing arguments about natural language learnability to be made explicit and quantified for the first time. We apply this method to a range of classic puzzles in language acquisition. We find some linguistic rules appear easily statistically learnable from language experience only, whereas others appear to require additional learning mechanisms (e.g., additional cues or innate constraints).

Keywords: Child language acquisition; Poverty of the stimulus; No negative evidence; Bayesian probabilistic models; Minimum description length; Simplicity principle; Natural language; Identification in the limit

1. Introduction

A central objective of cognitive science is to understand the mental processes underlying language acquisition. There is significant debate over how children acquire language. The debate began when linguists observed that children do not appear to receive adequate linguistic input to make language acquisition feasible (Baker & McCarthy, 1981; Braine,

Correspondence should be sent to Anne S. Hsu, Department of Cognitive, Perceptual, and Brain Sciences, University College London, London WC1HOAP, United Kingdom. E-mail: ahsu@gatsby.ucl.ac.uk

1971; Chomsky, 1965, 1975). This argument is also known as the poverty of the stimulus (POS).

A significant part of the POS argument focuses on the influential observation that children receive primarily *positive* linguistic data. This means that the child hears only examples of sentences that *are* possible and is never explicitly told which sentences *are not* possible (Bowerman, 1988; Brown & Hanlon, 1970; Marcus, 1993). However, all language speakers, including children, are constantly required to generalize linguistic constructions into new phrases and sentences. This leads to the central question of how the child learns to avoid *overgeneral* linguistic constructions, that is, constructions that are consistent with previous linguistic input but are ungrammatical. We will refer to these rules that prevent linguistic overgeneralization as *linguistic restrictions*. Two examples of this can be seen in the contraction restriction for *want to* and the dative alternation restriction for *donate*.

Example 1: Some linguistic restrictions

(1) a. Which team do you want to beat?
    b. Which team do you wanna beat?
    c. Which team do you want to win?
    d. *Which team do you wanna win?
(2) a. I gave some money to the museum.
    b. I gave the museum some money.
    c. I donated some money to the museum.
    d. *I donated the museum some money.

Sentence (1b) in Example 1 shows a grammatical contraction of *want to* in Sentence (1a). However, the contraction in (1d) of (1c) is not allowed. Thus, there is a restriction on the allowable contractions of *want to*. Similarly, *give* can occur both in the prepositional construction, as shown in (2a), as well as the direct construction, as shown in (2b). However, the similar verb, *donate*, can only appear in the prepositional construction (2c), and the direct construction (2d) is not allowed. We shall call the problem of learning such linguistic restrictions without negative evidence the Problem of No Negative Evidence (PoNNE). The PoNNE can be viewed as a subset of POS because the lack of negative evidence is one way in which the linguistic input can be impoverished.

There has been extensive research on how children learn language in light of the PoNNE and POS (Baker & McCarthy, 1981; Bowerman, 1988; Chater & Vitányi, 2007; Crain & Lillo-Martin, 1999; Elman, 1990; Lightfoot, 1998b; Mac Whinney, 1987; MacDonald, 2002; Perfors, Regier, & Tenenbaum, 2006; Pinker, 1994; Ritter & Kohonen, 1989; Tomasello, 2004; Yang, 2004). On one extreme, traditional nativist linguists have argued that the lack of negative evidence (PoNNE), as well as the more general lack of sufficient linguistic input in general (POS), makes many linguistic constructions impossible to learn without the aid of a large amount of innate language-specific knowledge (Chomsky, 1965; Crain & Lillo-Martin, 1999; Lightfoot, 1998b; Pinker, 1989). They use the POS and PoNNE to argue that the highly specific and seemingly arbitrary nature of many linguistic restrictions implies that our innate ability to learn these restrictions could not rely on

cognition-general mechanisms and instead must come from language-specific knowledge. This presumed innate knowledge of language is typically viewed as a Universal Grammar: linguistic principles that must be shared by all the world's natural languages. From the innateness perspective, researchers have argued that a large variety of specific linguistic constructions are unlearnable without the aid of innate linguistic knowledge.

In response to such innateness claims, many researchers have argued that learning constraints other than innate language-specific ones are sufficient for successful language learning. These can be innate domain-general learning mechanisms or other learning mechanisms that are acquired during development. Here, researchers argue that the language input *is* rich enough to allow for such learning. Researchers have also pointed out that children have access to many additional sources of noninnate linguistic information that could aid language learning, thus alleviating the need for direct negative evidence. For example, distinct word categories have been shown to be inferable from the distribution of words in sentences (Redington, Chater, & Finch, 1998). Other studies have shown that humans are sensitive to statistical patterns in syllables and other phonological cues, which can be used to acquire language (Newport & Aslin, 2004; Spencer et al., 2009). From an emergentist perspective, environmental information can also aid language learning. These include communicational contexts, for example, where a speaker is looking (Tomasello, 2003); prosody, for example, tone of speaker; and gestures, for example, pointing and hand signals (Tomasello, 2003). Computational models, such as connectionist networks, have emphasized that linguistic knowledge can be built from an interplay of different linguistic factors such as vocabulary, morphology, and metaphors (Bates, Marchman, Thal, Fenson, & Dale, 1994; Elman et al., 1996; Goldberg, 1995; Mac Whinney, 1987; Seidenberg, 1997) as well as biological constraints such as memory (MacDonald, 2002).

More recently, a significant line of research has countered innateness claims by using computational models to show that many features of language can be learned based on positive linguistic data and language statistics alone. Here, language statistics means any information that can be obtained from a language corpus (e.g., likely sentence structures, relationships between words). In particular, researchers have shown that language *is* theoretically learnable from language statistics and positive evidence only using the cognition-general principle of simplicity (Chater & Vitányi, 2007). Computational models have simulated positive language learning using a variety of different approaches (Chater, 2004; Chater & Vitányi, 2007; Dowman, 2007; Elman, 1990; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2007; Perfors et al., 2006; Pullum & Scholtz, 2002; Reali & Christiansen, 2005; Regier & Gahl, 2004; Stolcke, 1994; Tomasello, 2003). Connectionist models are one key modeling approach. Here, the patterns of connections in neural networks, often combined with dynamical systems theory, are used to learn behavioral patterns approximating that which humans assume toward syntactic structures (Bates et al., 1994; Christiansen & Chater, 2007; Elman, 1990; Elman et al., 1996; Mac Whinney, 1987; MacDonald, 2002; McClelland & Elman, 1986; Ritter & Kohonen, 1989). A particularly influential network approach uses simple recurrent networks (SRNs) to learn sequential language input (Elman, 1990). SRNs were shown to be capable of learning different categories of words, including transitive versus intransitive verbs. Such models thus are capable of learning restrictions to

overgeneralizations by accurately capturing the real and complex probability distributions present in language (Seidenberg, 1997). Another key modeling approach uses probabilistic models and the cognition-general learning principle of simplicity. Researchers have used these probabilistic models to show that linguistic restrictions can be acquired by directly learning the probability distribution of grammatical sentence structures in the language. These models learn this probability distribution using the cognition-general principle of simplicity (Dowman, 2000, 2007; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Grünwald, 1994; Langley & Stromsten, 2000; Perfors et al., 2006; Reali & Christiansen, 2005; Regier & Gahl, 2004; Stolcke, 1994). Compared with SRNs, probabilistic modeling based on the simplicity principle has the advantage of being a more transparent, tractable, and general methodology. Our current work builds on the probabilistic modeling approach.

Previously, probabilistic models have been applied to learning linguistic restrictions in either highly limited (Dowman, 2000, 2007; Elman, 1990; Grünwald, 1994; Langley & Stromsten, 2000; Stolcke, 1994), or artificial (Onnis, Roberts, & Chater, 2002), language datasets. These treatments often involve full models of language learning that are difficult to scale up to the level of natural language for most linguistic constructions (their computational load makes their application to natural language sets intractable). In the context of natural language, there have been a handful of studies that show learnability of specific linguistic cases such as anaphoric one (Foraker et al., 2009), auxiliary fronting (Reali & Christiansen, 2005), and hierarchical phrase structure (Perfors et al., 2006). However, there has been no general account for assessing the learnability of wide-ranging linguistic constructions.

Here, we provide a *general quantitative framework* that can be used to assess the learnability of any given *specific linguistic restriction* in the context of real language, using positive evidence and language statistics alone. Previous learnability analyses could not be applied to natural corpora, making previous natural language arguments prone to error or misinterpretation. We build upon previous probabilistic modeling approaches to develop a method applicable to natural language. Our method provides researchers with a new tool to explicitly explore the learnability in a corpus relative to well-known information-theoretic principles given a grammatical description.

We only aim to quantify learnability and do not aim to build a full model of language learning. This is because it would be intractable to build a full model of language learning that could serve as generally as our framework does in a natural language context. This enables us to provide a general framework for evaluating learnability for a wide range of specific linguistic constructions. When analyzing the learnability of a linguistic construction, there are two main assumptions: (a) The description of the grammatical rule for the construction to be learned (e.g., possibilities are to frame it as a general global rule or specific local rule). (b) The choice of a corpus that approximates the appropriate input. Given the current chosen description of the linguistic construction and an assumed corpus input, our framework provides a method for evaluating whether a construction is present with adequate frequency to make it learnable from language statistics. Our framework is very flexible because it is amenable to variation in these two main assumptions. By making these

assumptions explicit, we can provide a common forum for quantifying and discussing language learnability.

Our framework is analyzed from the perspective of an *ideal* learner, thus establishing an upper bound on learnability. If a linguistic restriction cannot be learned by an ideal learner, there are two possibilities: One, the learner's representation of language is not what we assumed it to be. In this case, a reframing of the learner's language representation could also potentially make a restriction learnable. For example, in the example of *want to* contraction given earlier, the restriction governing contraction can be viewed as either a singular case, that is, contraction is not allowed in this specific local context, or a general linguistic pattern, for example, a global rule such as trace-licensing (Crain, 1991). Such a change in assumed rule description can dramatically shift that regularity's apparent learnability (e.g., there may be many more instances of the more general pattern in the language input). A second possibility is that additional linguistic input is required. Such additional linguistic input could be provided through any of the multiple other sources mentioned above (e.g., more representative corpora, situational contexts, phonological cues, prosody and gestures, innate knowledge). Additionally, a more complex grammar may be preferred in the light of nonlinguistic data, for example, implicit social cues, explicit instruction in school. Our proposed framework is not tied to any particular alternative, but rather to provide a neutral way of quantifying learnability, given whatever assumptions about linguistic representations, prior knowledge, and data available to the learner, that the theorist would like to explore.

We see the primary contribution of this paper as methodological. Given representational assumptions and a chosen corpus of linguistic input, our framework yields estimates of the minimum amount of linguistic data required for learning the grammar rule that prevents overgeneralization of a linguistic restriction.

In addition to our main goal of presenting a methodology, we illustrate our framework by using it to estimate an upper bound on learnability for 19 linguistic constructions, many of which have been commonly cited as being unlearnable (Crain & Lillo-Martin, 1999), and others for which child language data have been previously collected. The quantification of learnability for specific linguistic constructs then provides a predicted order for the acquisition by children, which can be verified in further experiments.

Interestingly, we find that our framework yields very different learnability results for different linguistic constructions. For some linguistic constructions, the PoNNE is not a significant problem: These constructions are readily learnable from a relatively small corpus with minimal prior representational assumptions. For other constructions, the PoNNE appears to require sources of information other than language statistics alone. Where this is true, making a different assumption about the relevant linguistic rule description may aid learnability. Alternatively, additional data (e.g., concerning details of the speech signal, or social or environmental context) might crucially assist learning (indeed, the purpose of our framework is to encourage the comparison of results under different assumptions). These dramatic differences in outcome between constructions are reassuring because they indicate that the qualitative character of the results do not depend on fine, technical details of our assumptions.

Although our current analyses may not be sensitive to the technical details that we choose, they *are* potentially sensitive to our wider theoretical assumptions about how linguistic rules are represented. We chose assumptions that were straightforward and convenient. However, we stress that we do not intend to be strictly tied to these assumptions in our analysis. Instead, our analysis is meant to illustrate our framework, which we hope will be used by others to compare results under different assumptions. The purpose of the framework is to provide a quantitative common ground that can inspire more rigorous and precise proposals from different sides of the language debate.

The structure of the paper is as follows: First, in Section 2, we give an overview of the methods that we use to assess learnability, including the minimum description length (MDL) principle from which we build our framework. Then in Section 3, we describe in greater mathematical detail how MDL can be used to practically assess learnability for an ideal language learner given an assumed grammar rule description and a chosen corpus input. In Section 4, we apply our framework to 19 linguistic constructions and provide learnability results using our assumed grammatical rule descriptions and five English language corpora. Finally, Section 5 highlights the implications of contrasting learnability results and how these results can be used to provide future directions for shedding light on the POS and PoNNE.

2. Assessing PoNNE using the simplicity principle

2.1. Background on simplicity-based models

Our framework is based on a class of probabilistic models that uses the simplicity principle. Simplicity is a cognition-general learning principle that can be used not only for language but also for learning sequences, patterns, and exemplars. The general idea of the simplicity principle is that the learner should choose between models based on the simplicity with which they encode the data. Previous research has shown that models based on the simplicity principle can successfully learn linguistic restrictions using positive evidence alone (Dowman, 2000; Kemp, Perfors, & Tenenbaum, 2007; Langley & Stromsten, 2000; Onnis et al., 2002; Perfors et al., 2006; Stolcke, 1994). Simplicity has also been successfully applied to unsupervised morphological and phonological segmentation of language and speech (Brent, 1999; Goldsmith, 2001; de Marcken, 1996). Simplicity models view language input as streams of data and the grammar as a set of rules that prescribe how the data are encoded. Inherent in these models is the trade-off between simpler versus more complex grammars: Simpler overgeneral grammars are easier to learn. However, because they are less accurate descriptions of actual language statistics, they result in inefficient encoding of language input, that is, the language is represented using longer code lengths. More complex grammars (which enumerate linguistic restrictions) are more difficult to learn, but they better describe the language and result in a more efficient encoding of the language, that is, language can be represented using shorter code lengths. Under simplicity models, language learning can be viewed in analogy to investments in energy-efficient, money-saving

appliances. By investing in a more complicated grammar, which contains a restriction on a construction, the language speaker obtains encoding savings every time the construction occurs. This is analogous to investing in an expensive but efficient appliance that saves money with each use. A linguistic restriction is learned when the relevant linguistic context occurs often enough that the accumulated savings makes the more complicated grammar worthwhile. Because learning is based on how often different linguistic constructions appear, simplicity models are able to learn restrictions based on positive evidence alone (see Fig. 1).

Previous applications of simplicity-based models to the learning of linguistic restrictions in natural language have been formulated to treat only particular linguistic constructions (Foraker et al., 2009; Perfors et al., 2006). This is because the learning of general linguistic phenomena from a corpus representative of natural language is a too complicated task for current computational simulations. Full learning models are difficult to apply on natural language because the learning algorithm in these models requires a search over all possible grammars, and this search becomes intractable over large datasets. Therefore, we do not aim to build a full model of language learning. However, without the full search, it is still possible to establish the minimum amount of language input required to learn a linguistic

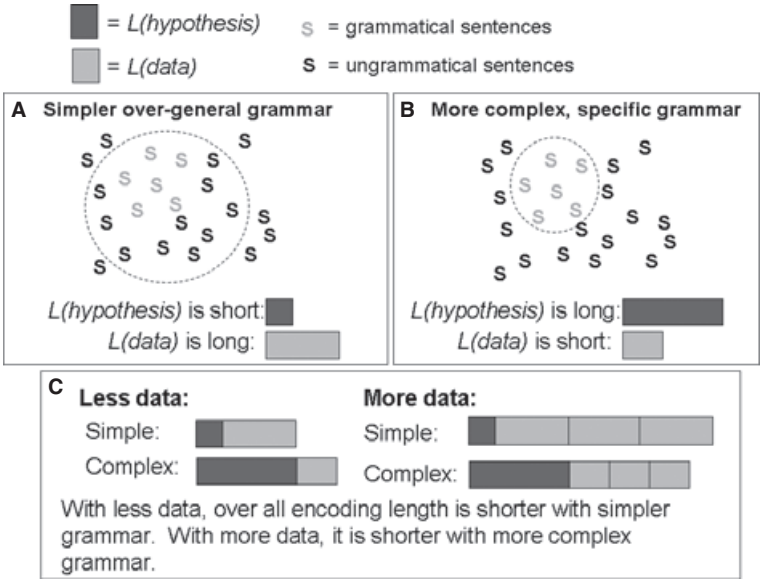

Fig. 1. MDL simple grammar versus efficient language encoding trade-off. (A) A simpler grammar is often overgeneral, that is, allows for ungrammatical sentences as well as grammatical ones. Such an overgeneral grammar is easy to describe (i.e., short grammar encoding length) but results in less efficient (longer) encoding of the language data. (B) A more complex grammar describes the language more accurately, that is, allows only for grammatical sentences and does not allow for ungrammatical sentences. A more complex grammar is more difficult to describe (i.e., longer grammar encoding length) but allows for more efficient (shorter) encoding of language data. (C) Initially, with limited language data, the simpler grammar results in an overall shorter coding length and is preferable under MDL. However, with more language input data, the savings accumulated from having a more efficient encoding of language data makes the more complex grammar worthwhile.

construction and thus quantify its learnability. This is achieved by comparing specific candidate grammars: the original grammar and the more complicated grammar that contains the linguistic restriction that is to be learned. We then can establish the amount of data required for the more complicated grammar to be worth ''investing'' in and determine whether this amount of data is available to the child learner. This method provides an upper bound on learnability because if a specifically chosen construction is not locally learnable based on the available language input, it cannot be learned under the general search of a full learning model.

This approach was used by Foraker et al. (2007) for learning restrictions on the anaphoric one. The authors show that the anaphoric one can be learned by noticing that ''one'' never replaces a noun without its complement (if the complement exists), whereas ''one'' *can* replace a noun without its modifier. Hence ''one'' will never be followed by the complement of the noun it replaces. Here, the two grammars being compared (a grammar that predicts a string of ''one'' + complement and another grammar that does not) were of equal complexity. Thus, here, the ''correct'' grammar is always more efficient for any language set that includes complement phrases with just one example of the anaphoric ''one'' because the incorrect grammar would assign nonzero probability to the nonoccurring string, ''one'' + complement. This example shows that in the special case where candidate grammars have equal complexity, the most accurate grammar is obviously more efficient and hence immediately preferable under simplicity with minimal language exposure. Our framework allows for learnability analysis in the general cases where the candidate grammars are not of equal complexity.

Although we do not presume that our framework models the actual complex process of language learning, we note that it does contain some aspects of the learning problem a child is faced with. Child language learning is likely to be an incremental process where knowledge of rules is built on previous knowledge. This type of incremental learning does not require a comprehensive search over all possible grammars. This greatly reduces the computational load involved in acquiring each new grammatical rule. Thus, our framework may be akin to a learner who uses previous grammatical knowledge to narrow the space of possible grammars to particularly relevant candidates. Our method simulates the step in incremental learning during which specific grammatical rules are (or are not) acquired.

2.2. MDL as a model of language acquisition

In practice, the simplicity principle can be instantiated through the principle of MDL. MDL is a computational tool that can be used to quantify the information available in the input to an idealized statistical learner of language as well as of general cognitive domains (Feldman, 2000). When MDL is applied to language, grammars can be represented as a set of rules, such as that of a probabilistic context-free grammar (PCFG; Grünwald, 1994). An information-theoretic cost can then be assigned to encoding both the rules of the grammar as well as the language under those rules. For our purposes, the language consists of the full language corpus that a speaker has experienced. MDL does not merely select for the simplest grammar, as has been proposed in other theories of language (Chomsky, 1955;

Fodor & Crain, 1987). Instead, MDL selects the grammar that minimizes the *total* encoding length (measured in bits) of both the grammatical description and the encoded language length. MDL is a concrete, practical framework that takes as its input real language data and outputs an optimal choice of grammar. The MDL framework can also be expressed as a corresponding Bayesian model with a particular prior (Chater, 1996; MacKay, 2003; Vitányi & Li, 2000). Here, code length of the model (i.e., grammar) and code length of data under the model (i.e., the encoded language) in MDL correspond to prior probabilities and likelihood terms, respectively, in the Bayesian framework.

2.3. Two-part MDL

As with the previous work of Dowman (2007), the version of MDL that we implement is known as two-part MDL. From here on, for conciseness we will refer to two-part MDL simply as MDL, though in general there are many other formulations of MDL that are not two part. The first part of MDL encodes a probability distribution, and the second part encodes the data, as a sample of that distribution. In the context of language acquisition, the first part of MDL uses probabilistic grammatical rules to define a probability distribution over linguistic constructions, which combine to form sentences. Note that these probabilities are not necessarily the real probabilities of sentences in language, but the probabilities as specified under the current hypothesized grammar (see Section 3.1.1, for equation and mathematical details). The second part of MDL consists of the encoded representation of all the sentences that a child has heard so far (see Section 3.1.2, for equation and mathematical details). According to information theory, the most efficient encoding occurs when each data element is assigned a code of length equal to the smallest integer $\geq -\log_2(p_n)$ bits, where p_n is the probability of the nth element in the data. These probabilities are defined by the grammatical description in the first part of MDL. If the probabilities defined in the grammar are more accurately matched to the actual probabilities in language, the grammar description will be more efficient (see Section 3.1.3, for a concrete linguistic example). In our analogy with energy-saving appliances, the first part of MDL would be evaluation of the cost of the appliance (i.e., cheap for simple grammars, expensive for complicated, more descriptive grammars) and the second part of MDL would be evaluation of the cost of using the appliance. Note that the analogy between MDL and appliances is not absolute because with an appliance, the buyer must project how often it will be used and assess ahead of time whether the more expensive appliance is worth the investment. In contrast, under MDL, the grammatical description is updated to be the most efficient one each time more data inputs are obtained. Savings occur because certain grammatical descriptions result in a more efficient (shorter) encoding of the language data. In general, more complex (i.e., more expensive) grammatical descriptions allow for more efficient encoding of the language data. Because savings accumulate as constructions appear more often, more complex grammars are learned (i.e., become worth investing in) when constructions occur often enough to accumulate a sufficient amount of savings. If there is little language data (i.e., a person has not been exposed to much language), a more efficient encoding of the language does not produce a big increase in savings. Thus, when there is less language data, it is better to make a cheaper investment in

a simpler grammar as there is not as much savings to be made. When there is more language data, investment in a more costly, complicated grammar becomes worthwhile. This characteristic of MDL learning can explain the early overgeneralizations followed by retreat to the correct grammar that has been observed in children's speech (Bowerman, 1988).

2.4. New contributions: MDL evaluation in natural language

A previous work (Dowman, 2007) has applied a full-learning MDL model to small artificial corpora. Full-learning models of MDL involve a search over all grammars that can possibly describe the input corpus. This makes full-learning models of MDL unfeasible for large natural-language corpora. Our current work presents a new extension of previous methods that allows MDL to be applied to natural language learning. The following new developments of our proposed method enable MDL evaluation of natural language. First, we show that specification of the entire grammar is not necessary. Instead, learnability can be estimated using just the relevant portions of the two grammars. This is not an obvious result because traditional MDL evaluation requires knowledge of the speaker's entire grammar. Here, we show how MDL differences between grammars can be evaluated without knowledge of the full grammar, by assuming that old and new grammars only differ in specific local features that are critical to the construction being learned (derivation detailed in Section 3.2.3). Second, though we do not need to specify the full grammar, we still need to approximate some general form of a speaker's grammar in order to enumerate the parts of grammar relevant for calculating grammar differences. Here, we provide simple general frameworks for explicitly representing a speaker's grammar. This formulation is flexible enough to represent the learning of a variety of linguistic rules. Although this formulation is far from the only possible one, it is a general starting point, which can be adapted as needed in future work. Third, we present a method for estimating language-encoding length differences between new versus old grammars, given a chosen language corpus.

3. Methods

In practical contexts, MDL application requires choosing a representational format for the grammar. The particular representation used will affect the grammar length as well as the data description length. As mentioned in Section 1, this representation-dependent feature of practical MDL is useful because it provides a way to compare the effectiveness of different representations of grammar.[1] A large number of researchers have taken the perspective that language learning involves acquiring a symbolic grammar (Chomsky, 1975; Crain & Lillo-Martin, 1999; Fodor, Bever, & Garrett, 1974), although a range of other perspectives on the acquisition problem have also been proposed (Elman et al., 1996; Goldberg, 2003; Lakoff, 1987; Langacker, 1991; Tomasello, 2003). Here, we will take the symbolic grammar representation perspective and illustrate our framework using a PCFG. This framework is able to capture a wide variety of the linguistic patterns found in language. Although we use PCFGs to illustrate our method, we stress that our framework can be implemented in

any grammar formalism, which is a particular strength of our approach: If different formalisms yield significantly different results, this would suggest that the nature of a child's language representation is important for theories of acquisition.

In this section, we will first illustrate how MDL is used to evaluate language encoding, along with an example application of MDL to a limited language set. We will then describe our new method, which uses the MDL framework to assess learnability in natural language contexts.

3.1. Basic evaluation of MDL

3.1.1. The MDL code, Part 1: Encoding length of grammatical description

The first part of MDL consists of evaluating the encoding length of the grammatical description. As mentioned above, we encode our grammar using a PCFG representation. Some examples of PCFG rules are S->NP VP #, VP-> (tense) V#, and so on. Here, we use an end symbol, #, to indicate the end of a rule. Because our rules allow for a variable number of left-hand-side symbols, an end symbol is necessary in our code to establish where one rule ends and another begins. Alternatively, if all rules had the same number of left-hand-side symbols, the end symbol would not be necessary. In order to encode these rules, we must have an encoded representation for each the grammar symbols used, that is, S, NP, VP, (tense), V, #. An obvious possible code for these symbols would be the basic ASCII character set encoding that uses 8 bits per symbol (allowing for 2^8 = 256 different symbols). However, this is not an optimal code for our grammar symbols because we may not need exactly 256 different symbols, and the symbols will be used with different frequencies in the grammatical description. To encode the grammar most efficiently, information theory again applies: The most efficient encoding occurs when each symbol in the grammar definition is assigned a code of length of about $-\log_2(p_n)$ bits, where p_n is the probability in the grammatical description of the nth symbol. Thus, for an optimally encoded grammar, we will need to tally up the frequencies of all grammar symbols used in our grammatical description. Each grammar symbol is then encoded with a binary string of length approximately equal to $-\log_2(f_s/F_{\text{total}})$, where f_s is equal to the frequency of occurrence of symbol s in the grammatical description and $F_{\text{total}} = \sum_s f_s$ is the total frequency of all symbols that occur in the grammatical description.

Finally, our grammatical description needs to include a list of the usage probabilities for each symbol in the grammar and each PCFG rule in the encoded language. The probability of each grammar symbol is calculated directly from the grammar description. Symbol probabilities are required to map each symbol to a code element, that is, its efficient binary string representation described above. The probability of each PCFG rule is calculated by estimating how often each rule will be used to encode the language, that is, how often different linguistic constructions and sentences appear. These probabilities will be used in the second part of MDL to construct the code for representing specific linguistic constructions and rules. The more accurately these probabilities reflect those in real language, the more efficiently the language will be encoded. However, it is not practical to encode all these probabilities to infinite values. Instead, following Dowman (2007), we assume that all

occurrence probabilities will be encoded to fixed accuracy, for example, two decimal places. Probabilities with accuracies of two decimal places can be encoded as integers 0 through 99. Thus, according to standard coding theory (Grünwald, 1994), all probability values will require a constant encoding length of approximately $-\log_2(1/100) = 6.6$ bits. In summary, our grammatical description will include the following: the list of PCFG rules, the probabilities of all grammar symbols used to enumerate these rules, and the probabilities with which each rule occurs in language. See Fig. 2 for an example of a grammatical description based on PCFG rules.

A Sample general-syntax rules

code	Rule	prob
1	S -> NP VP #	1
1	NP ->[Det] N' #	1
1	N '-> AP N' #	0.33
2	N' -> N' PP #	0.33
3	N' -> N #	0.33
1	VP -> t V' #	1
1	V' ->V' NP #	0.33
2	V'->V' PP #	0.33
3	V' -> V #	0.33
1	PP' -> P' #	1
1	P' -> P' NP#	0.33
2	P' -> P' VP #	0.33
3	P' -> P #	0.33

…

Encoding of general-syntax grammar:
S, NP, VP, #, 1, 2, NP, (det), N', #, 1…

NP=Noun Phrase, N=Noun,
D=determiner, VP=Verb Phrase,
V=Verb, t=Verb tense
PP=Preposition Phrase,
P=Preposition, AP=Adjective Phrase

*General-use probabilities are not reflective of real language statistics and for illustration purposes only

B Sample specific-situation rules

[situation definition give]
[direct-dative] V'->V NP NP #
[prepositional-dative] V'->V NP PP #
[dative-alternation give verb] give#
[end]

[situation] [dative-alternation give verb]
1 [direct-dative] 0.75
2 [prepositional-dative] 0.25
[end]

[situation definition want to]
[want to] want to #
[end]

[situation] [want to]
1 [contract] 0.25
2 [no-contract] 0.75
[end]

Sample encoded sequence of specific-situation grammar rules:

[situation definition donate/give], [direct-dative], V, V', NP, NP, #, [prepositional-dative], V,V', NP, PP,# …

C Sample vocabulary rules

Code	Rule	prob
1	Det ->[empty] #	0.1
2	Det ->the #	0.1
3	Det ->a #	0.1
…		
1	N -> I #	0.05
2	N -> you #	0.05
3	N -> home #	0.05
4	N -> money #	0.05
5	N -> [empty] #	0.05
…		
1	V -> go #	0.01
2	V -> want #	0.01
3	V -> give #	0.01
…		
1	t -> present #	0.2
2	t -> past #	0.2
…		
1	P -> to #	0.1
2	P -> from #	0.1

…

Sample encoded sequence of vocabulary grammar rules:

Det, the, #, Det, a, #, Det, an, #, N, I, #, N, you, #…

Fig. 2. Examples of grammatical description. Here is a sample of the type of grammatical description that we used in our analysis. The grammatical description is described in more detail in Appendix S1. For illustration purposes, we associate each specific-usage rule with a unique index number in place of the unique code element, which would be used in the actual grammar. This index number is only unique relative to the specified situation (i.e., the specific symbol on the left-hand side). (A) Basic syntax rules include basic phrase structure grammatical rules and their usage probabilities. (B) Specific-situation rules will be used to specify concepts related to specific linguistic situations and their usage probabilities under specific situations, such as when contraction may occur. Each specific usage rule comes with a specific situation definition that describes the relevant linguistic situation. (C) Vocabulary rules represent the different words in a speaker's vocabulary. Although we use this hypothetical setup of grammatical knowledge, we wish to stress that our results do not heavily depend on the exact setup chosen. The setup of these grammars is discussed further in the Appendix S1.

The encoding length of a grammar (i.e., the investment cost) will be given by L(grammar). Here, we use a formula similar to that in Dowman (2007)[2] :

$$L(\text{grammar}) = -\sum_s f_s \log_2 \frac{f_s}{F_{\text{total}}} + \left(N_{\text{symbols}} + N_{\text{rules}}\right) C_{\text{prob}} \text{ bits} \quad (1)$$

Here, f_s is the occurrence frequency of symbol s. $F_{\text{total}} = \sum_s f_s$, the total occurrence frequencies of all symbol types in the grammatical description. N_{rules} is the number of PCFG rules in the grammar. N_{symbols} is the number of different symbol types in the grammar. C_{prob} is the constant value of encoding probabilities to fixed decimal accuracy.

Intuitively, this equation sums the encoding cost contributions from (a) all the symbols used in the grammatical description; (b) the probabilities of each grammar rule's usage in the language data; and (c) the probabilities of each symbol in the grammatical description. The first component in Eq. 1, $-\sum_s f_s \log_2(f_s/F_{\text{total}})$, is the length of encoding associated with listing all the PCFG rules. Note that f_s/F_{total} is the probability of occurrence for symbol s in the grammatical rules. The second component, $N_{\text{rules}}C_{\text{prob}}$, is the length of encoding the probabilities with which each rule occurs in language (where C_{prob} is the code length for a single rule). The third component, $N_{\text{symbols}}C_{\text{prob}}$, is the length of encoding the probabilities with which the grammar symbols occur in the grammatical description. The same constant, C_{prob}, is used for both symbols and rules probabilities because we assume both are encoded to the same decimal accuracy.

3.1.2. The MDL code, Part 2: Encoding length of the language data

The second part of the MDL code consists of evaluating the length of encoding the language data. The language is encoded relative to the grammar defined in the first part of MDL. Here, we assume that this length will include the encoding of all sentences experienced, including repeats.[3] For example, if the sentence *How are you?* required 12 bits to encode, and is experienced 400 times by a language learner, the total code length required to encode these occurrences would be 400 × 12 = 4,800 bits. Sentences are encoded as derivations from a grammar using a sequence of PCFG rules. In PCFG, each rule is expanded in sequence until a sentence results. An example is provided in Section 3.1.3. Again, for optimal encoding, each PCFG rule is represented with a code length that is about $-\log_2(p_n)$ bits, where p_n is the probability of the nth rule being used in language. A notable point is that not all constructions are allowed in all linguistic contexts. Thus, the probability of a linguistic construction is calculated relative to all other possible linguistic constructions that can occur in the given context. In terms of PCFG rules, this means that not all expansions are allowed for a given left-hand-side grammar symbol. Thus, the probability of any PCFG rule is calculated only relative to all other PCFG rules that have the same left-hand-side symbol. The length of the encoded data, as prescribed by the grammar, L(data) will be estimated using (Dowman, 2007):

$$L(\text{data}) = -\sum_r f_r \log_2 \frac{f_r}{t_r} \text{ bits} \quad (2)$$

Here f_r is the frequency (in the language data) of rule r, and t_r is the total frequency (in the language data) of all rules with the same symbol on their left-hand side as rule r.

Thus, f_r/t_r is the probability of rule r being used relative to all other possible rules that could be used at that point of the phrase expansion. Thus, the same binary code element will refer to different PCFG rules depending on the current grammar symbol being expanded.

3.1.3. MDL evaluation: A simple concrete example

Let us consider some concrete examples of language encoding under MDL. A key feature of practical MDL is that only the length of the representation is important. Thus, only the representation lengths need to be determined in order to select the best grammar. For illustration purposes, we will show encoding length evaluation of a single artificial grammar. In practice, MDL involves evaluating the description length of multiple (in our implementation two) grammars and then choosing the one with the shortest description length. Following Dowman (2007), we use PCFGs with binary branching or nonbranching rules as shown in Example 2. These grammars contain both terminal symbols (words) and nonterminal symbols (any other nodes in a syntactic tree, such as lexical or phrasal categories). Within this formalism, a valid sentence is any string of terminal symbols that can be derived by starting with the symbol *S*, and repeatedly expanding symbols from left to right using any of the rules in the grammar.

Example 2: Encoding language using PCFGs

Sample phrase structure grammar:

(1) S->NP VP # 1.0
(2) NP->N # 1.0
(3) VP-> V PP # 1.0
(4) PP-> P N # 1.0
(5) N-> John # 0.25
(6) N->Mary # 0.75
(7) P-> at # 1.0
(8) V-> smiled # 1.0

Sample language data:

Mary smiled at John.
Mary smiled at Mary.

Here, we use the standard symbols *S* (Sentence), *VP* (Verb phrase), *NP* (Noun phrase), *PP* (Prepositional phrase), and we use # to indicate an end symbol. The numbers following the # symbol are the relative probabilities for which these rules occur in our example artificial language. Remember, these probabilities are defined relative to all other possible expansions of a rule, that is, all other rules with the same left-hand-side symbol. Thus, because Rules 1–4, 7, and 8 are the only possible expansions given the symbols on their left-hand side (*S*, *NP*, *VP*, *PP*, *P*, *V*, respectively), their relative probability of occurrence is 1. Rule 5 is applied only one of four times and thus has the probability .25. Rule 6 is applied three of four times and thus has probability .75. That is, to derive the first sentence in the language data in Example 2: *S* is expanded to *NP VP*, which is expanded to *N V PP*, which is expanded to, *N V P N*, which can be then expanded to *Mary smiled at John.* This sentence would be represented by the following ordered list of the rules: 1, 2, 3, 4, 6, 8, 7, 5. That is:

S->NP VP #, NP->N#, VP->V PP#, PP->P N#, and so on. In this highly artificial example, there are only a few possible sentences in the ''language'' prescribed by this set of grammatical rules. In fact, not all possible allowed sentences are present in this ''language'' of two sentences, that is, *John smiled at Mary* and *John smiled at John* are not in the language input and thus this example is actually a case of an overgeneral grammar.

Table 1 breaks down the calculation of encoding length for the grammar in Example 2. The symbol occurrence frequencies in the second column reflect the fact that there is one token of symbol type S (occurring in Rule 1), two tokens of symbol type NP (occurring in Rule 2), and so on. The total of the first column shows that there are 12 symbol types, $N_{\text{symbols}} = 12$. The total of the second column shows that there are 27 symbol tokens, $F_{\text{total}} = \sum_s f_s = 27$. The total of the third column is the cost of enumerating grammar rules, $-\sum_s f_s \log_2(f_s/F_{\text{total}}) = 86.38$ bits. The number of rules can be read directly from the grammar in Example 2, $N_{\text{rules}} = 8$. Finally, we need to estimate the free parameter, C_{prob}, which is the number of bits used to encode the probabilities of each symbol type and each rule. Here, we assume probabilities will be encoded to an accuracy of two decimal places. Thus, we set $C_{\text{prob}} = -\log_2(1/100)$ bits = 6.6 (as explained in Section 3.1.1). Substituting these values into Eq. 1, we get a total grammatical description length of 218 bits. That is the first part of the MDL evaluation.

The second part is to quantify the encoding length of the language data. The first sentence, *Mary smiled at John*, is encoded with the following list of the rules: 1, 2, 3, 4, 6, 8, 7, 5 and the second sentence, *Mary smiled at Mary*, is encoded with the following rules: 1, 2, 3, 4, 6, 8, 7, 6. Here, in our simple grammar, rules 1, 2, 3, 4, 7 and 8 occur with probability 1 and therefore require 0 bits to encode. The only contribution to encoding length comes from Rules 5 and 6. Rule 5 requires $-\log_2(0.25) = 2$ bits and Rule 6 requires $-\log_2(0.75) = 0.4$ bits. Thus, 2.4 bits are required for encoding *Mary smiled at John* (Rules 5 and 6 occurring once each) and 0.8 bits are required for encoding *Mary smiled at Mary*

Table 1
Encoding costs for grammar in Example 2

Symbol Type	Symbol Occurrence Frequency f_s	Symbol Encoding Cost (bits) $f_s \log_2(f_s/F_{\text{total}})$
S	1	$-1 \times \log_2(1/27) = 4.75$
NP	2	$-2 \times \log_2(2/27) = 7.5$
VP	2	$-2 \times \log_2(2/27) = 7.5$
PP	2	$-2 \times \log_2(2/27) = 7.5$
V	2	$-2 \times \log_2(2/27) = 7.5$
N	4	$-4 \times \log_2(4/27) = 11.02$
P	2	$-2 \times \log_2(2/27) = 7.5$
Mary	1	$-1 \times \log_2(1/27) = 4.75$
smiled	1	$-1 \times \log_2(1/27) = 4.75$
John	1	$-1 \times \log_2(1/27) = 4.75$
at	1	$-1 \times \log_2(1/27) = 4.75$
#	8	$-8 \times \log_2(8/27) = 14.04$
Column totals		
$N_{\text{symbols}} = 12$	$F_{\text{total}} = 27$	$-\sum_s f_s \log_2(f_s/F_{\text{total}}) = 86.38$

(Rule 5 occurring twice). Thus, the total code length for the grammar and the language data in Example 2 is 218 bits + 2.4 + 0.8 bits = 221.2 bits. Note that by encoding the more commonly used Rule 6 with fewer (0.4) bits, and the more rarely used Rule 5 with more (2) bits, we have a more efficient code than if both rules were encoded using 1 bit each. Fig. 2 shows further examples of PCFG grammars. Figs. 3 and 4 show sentences encoded using the grammar specified in Fig. 2.

For illustration purposes, Example 2 showed the encoding length evaluation for a single grammar only. However, MDL involves description length evaluation of multiple grammars (in our case two grammars), and choosing the grammar that allows the shortest overall encoding length of both grammar and language data, that is, we want to find, $\min_{\text{grammar}}[L(\text{grammar}) + L(\text{data}|\text{grammar})]$, where $L(\text{grammar})$ is the length of the grammatical description and $L(\text{data}|\text{grammar})$ is the length of the encoded data under that grammar. MDL serves as a principled method of choosing between shorter grammars that result in longer encoded data lengths versus longer grammars that result in shorter encoded data lengths: That is, with little language data, $L(\text{grammar})$ will have a larger contribution to the total length. However, as the amount of data increases (i.e., a speaker has heard a lot of language), the contribution of $L(\text{grammar})$ will become small relative to $L(\text{data}|\text{grammar})$.

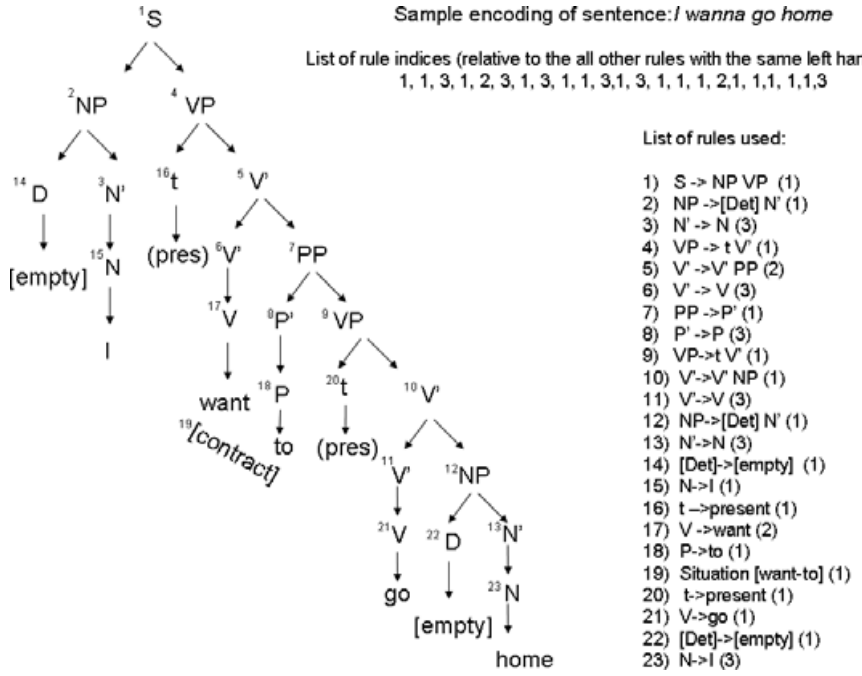

Fig. 3. Sample encoding of sentence: *I wanna go home.* Sentence is encoded using the code numbers that represent each of the grammatical rules shown in Fig. 2. In practice, the code number would be replaced with its equivalent binary number with string length approximately equal to its log probability. Probabilities are calculated relative to all other rules with the same left-hand side. Thus, the same code number refers to different rules depending on the grammar symbol being expanded. The full list of rules used is also shown.

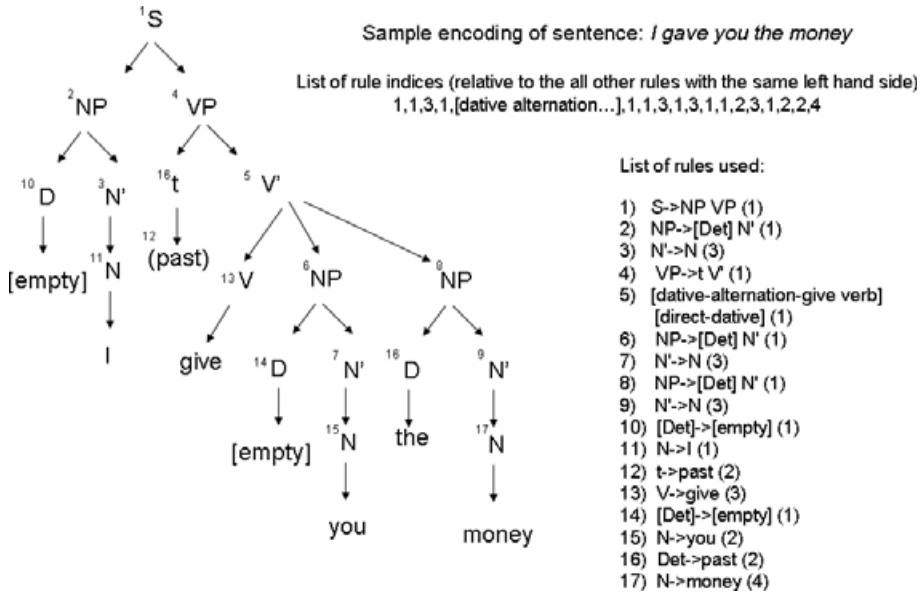

Fig. 4. Sample encoding of sentence: *I gave you the money.* Sentence is encoded using the index numbers that represent each of the grammatical rules shown in Fig. 2. The full list of encoded rules is also shown.

3.2. MDL applied to quantifying PoNNE and POS in natural language acquisition

In order to make MDL evaluation tractable in the context of natural language, we replace a general search over all grammars with a local comparison between specific candidate grammars. This allows us to assess the acquisition of any specific linguistic constructions in the context of real language. Under MDL, the chosen grammar is the one that optimizes the trade-off between grammar cost and encoding savings. This natural trade-off inherent to MDL allows us to make comparisons between candidate grammars without considering the whole of language. Thus, we do not have to define generative models of grammar nor conduct searches over the whole of language, as would be required in a full model of learning. Instead, our framework assumes that the new and old grammars differ only in local features that are critical to the construction being learned. Thus, all other parts of the two grammars are assumed to remain the same. Because we are only comparing the difference between local features of candidate grammars, we only need explicit descriptions of the portions that differ between these grammars. These will be the portions that describe the language constructions being examined. The candidate grammars will be chosen to correspond to possible grammatical hypotheses children could have during learning. Although we use two candidate grammars in our current implementation, several candidate grammars may also be used.

The candidate grammars being compared consist of a simple more general grammar and another more complex grammar. The more complex grammar will include the general rule from the simple grammar as well as an additional restriction to the general rule. The simpler

grammar would represent the overgeneralizations in a child's grammar, and the complex grammar would represent the correct grammar that we all achieve by adulthood (Bowerman, 1988). The more complex grammar has a longer grammar description length, but it results in a shorter encoding length of the language data. We can now evaluate MDL by assessing whether the complex grammar is worth investing in. This is carried out by evaluating the trade-off between the encoding length costs resulting from a more complex grammar description versus the encoding length savings resulting from using the more complex grammar to encode the language data. A linguistic restriction becomes learnable when the higher cost grammar with the restriction becomes worth investing in. This means there is enough language data such that the encoding length savings offered by the more complex grammar exceeds its additional cost relative to the simpler grammar.

3.2.1. Approximation of speaker's grammar

Under our framework, we assume that the learner has already acquired significant previous grammatical knowledge. This is consistent with an incremental approach to language learning. The assumption of a certain amount of prior linguistic knowledge is reasonable because language acquisition is an incremental process where new knowledge is continuously built on previous knowledge. For example, we assume that the language speaker has a certain amount of semantic and syntactic knowledge such as knowledge of the distinction between verbs and nouns. We also note that we assume the child has some representation of sentence boundaries and can distinguish noncontracted from contracted phonological forms (*want to* vs. *wanna*). Our analysis does not require that children have formal grammatical understanding of these concepts. Instead, they only need to be able to represent these concepts with the relevant patterns of phonological strings. For example, a child only need be aware that the word *want to* can be replaced with *wanna*, and the child does not need to have a formal understanding of contraction. Thus, the symbol representing contraction, defined as [contract] in our grammar, simply represents the replacement of two words by one other. Similarly, a child could recognize the presence of a sentence boundary, based on an intonations contour and/or a following pause. Any additional specific conceptual knowledge will be discussed individually under the analysis for each construction. For our purposes, the bulk of the child's previously acquired grammar does not need to be defined in explicit detail. However, a general idea of its form must be assumed so that the parts relevant to the construction being learned can be written down explicitly. Next, we describe some rough assumptions that we make of a child's grammar and describe how to specify the portion of two candidate grammars that are relevant to the acquisition of a particular construction. We emphasize that we are not tied to this particular grammatical description. It is only one example of many possible grammatical descriptions that can be used within our framework.

In our current analysis, we will choose a grammar that consists of three types of rules, *general-syntax rules*, *vocabulary rules*, and *specific-situation rules.* The specific form of these rules is chosen to make the analysis as convenient and tractable as possible rather than to reflect any of the wide range of current syntactic theories. General-syntax rules depict basic sentence syntax structures that apply generally to English language. Vocabulary rules represent the words in a speaker's vocabulary. Specific-situation rules describe concepts and

knowledge needed for identifying specific linguistic situations along with the usage probabilities for these specific situations. Most of the grammatical descriptions portions that are relevant to our analysis will be contained in the specific-situation rules. Sample specific-situation rules are shown in Fig. 2b. Next, we explain this representation of grammatical rules further detail.

3.2.1.1. General-syntax rules: General-syntax rules include basic phrase structure grammatical rules. As in Example 2, a probability is specified for each general-syntax rule, which will determine its coding length when it is used to encode language data in general situations. The end symbol, #, demarcates the end of a rule and the first symbol following # will be the general probability of that rule. These general probabilities may be overridden by the probabilities specified for specific linguistic situations by specific-situation rules.

3.2.1.2. Vocabulary rules: Vocabulary rules represent the different words in a speaker's vocabulary. Vocabulary rules can only be introduced with special almost-terminal symbols such as N,V, and so on. Vocabulary rule grammars will be encoded in the same way as general-syntax rule grammars and are shown in Fig. 2c.

3.2.1.3. Specific-situation rules: Specific-situation rules will be used to specify specific linguistic situations that require particular syntax rules or other rules such as contraction. These specific situations will each have associated usage probabilities. The specific linguistic situation and their associated concepts will be defined using a *specific situation definitions* structure. New probabilities will be specified for each specific situation in a *situation* structure. Remember, for the purposes of our MDL analysis, the precise symbols used (e.g., numbers or letters) are not important. All that needs to be known is the total number of symbol tokens and their frequencies of appearance in the grammar.

Example 3: Specific situation definitions

[situation definition verb1/verb2]
    [direct-dative] VP->V NP NP #
    [prepositional-dative] VP->V NP PP #
    [dative-alternation verb1/verb2] verb1 verb2 #
[end]
[situation] [dative-alternation verb1/verb2]
    [direct-dative] *probability*
    [prepositional-dative] *probability*
[end]

Example 3 shows such structures for two similar verbs, verb1 and verb 2, that are both assumed to undergo the dative alternation. A specific situation definition structure would begin with a symbol indicating the beginning of a set of definitions (e.g., [situation definition verb1/verb2]). The end of the set of definitions would be signified by an [end] symbol. The specific situation definitions would contain descriptions of specific linguistic situations and map them to associated symbols, here depicted within brackets. These definitions

include words or syntactic structures used to recognize the specific situation (e.g., [dative verb1/verb2]) as well as the general-syntax rules relevant to the situation (e.g., [direct-dative], [prepositional-dative]). An *end* symbol, #, appears at the end of each concept, signifying the beginning of a new concept. In specific situation definitions, *end* symbols are not followed by probabilities as these will be defined within the situation structure for each specific situation (see Example 3). The usage probabilities for each specific linguistic situation will be encoded beginning with the symbol, [situation], followed by the concept-symbol corresponding to the situation (e.g., [dative verb1/verb2]). This will then be followed by a list of symbols corresponding to the relevant linguistic situation (e.g., [direct-dative], [prepositional-dative]) followed by their relative probabilities under that specific situation.

Note that the dative alternation in Example 3 could just as easily have been represented using basic PCFGs. This is true for verb alternations. However, many linguistic restrictions require enumeration of more complex linguistic situations, such as that restricting the contraction of *want to* shown in Sentence 1d in Example 1. Specific-situation rules allow for easy and efficient representation of more complicated linguistic concepts that are not so readily described using PCFGs. For our purposes, the acquisition of a new linguistic restriction can be formulated as the acquisition of additional specific situation definitions along with the new probabilities associated with these new definitions. This will be explained in detail in the following section. Also see Appendix S1 for further details.

3.2.2. Specifying new versus original grammars

Now that we have an idea of the grammar representation, we will show how to write down explicitly the relevant portions of two candidate grammars. These will be from the original grammar, where the linguistic restriction for the considered construction has not been acquired, and the new grammar, where the restriction rule has been acquired. Let us consider a specific instantiation of the dative alternation of *donate* versus *give* as shown in the second group of sentences in Example 1. Here, the original grammar would be the equivalent of replacing verb1 and verb2 in Example 3 with *give* and *donate*. This overgeneral original grammar would assume that both verbs could alternate (as defined under the concept ([dative-alternation give/donate]). On the contrary, the new grammar would only allow *give* to alternate ([dative-alternation give]) while restricting *donate* to the prepositional construction ([prepositional-only donate]). The explicitly written-out original versus new grammar for the dative alternation restriction on *donate* is shown in Example 4. Here, the occurrence probabilities of the direct-dative and prepositional-dative constructions within original and new grammars are estimated from the spoken portion of the British National Corpus (BNC).

Example 4: Original versus new grammar for restriction on dative alternation of donate

Original grammar:

```
[situation definition donate/give]
    [direct-dative] VP->V NP NP #
    [prepositional-dative] VP->V NP PP #
    [dative-alternation give/donate] donate give
[end]
```

[situation] [dative-alternation give/donate]
[direct-dative] 0.8
[prepositional-dative] 0.2
[end]

New grammar:
[situation definition donate/give]
[direct-dative] VP->V NP NP #
[prepositional-dative] VP->V NP PP #
[dative-alternation give] give #
[prepositional-only donate] donate #
[end]
[situation] [dative-alternation give]
[direct-dative] # 0.87
[prepositional-dative] # 0.13
[end]
[situation] [prepositional-dative-only donate]
[prepositional-dative] # 1.0
[end]

3.2.3. Calculating cost differences for new versus original grammars

Notice that the new grammar is more complex: It has more symbols and is longer and hence will require a longer encoding length to define. Given explicit descriptions of new versus original grammars, we now need to use Eq. 1 to evaluate encoding length differences between new versus original grammatical descriptions. The evaluation of Eq. 1 on the entirety of a child's grammar would be very difficult because it requires knowing the symbol occurrence frequency, f_s, for all symbols used in the grammatical description. This would require enumerating the child's full grammar, which would be an overwhelmingly unwieldy task. It would also require knowing the values of N_{rules} and N_{symbols}, the number of rules and the number of symbols, respectively, in the entire grammar. This would again be very hard to estimate without enumerating the child's full grammar. However, the calculation of grammar encoding length *differences* is much easier than the calculation of absolute grammar encoding length. For example, as shown in Eq. 3, the calculation of the encoding length differences between two grammars only requires knowledge of *symbol occurrence frequencies that differ* between the two grammars. This means we only have to estimate occurrence frequencies for a small subset of grammar symbols in the grammatical description (the ones that differ between the grammars being compared). Similarly, we also only need the differences in N_{rules} and N_{symbols} between the two grammars, which can be obtained directly from the relevant portions of the two grammatical descriptions. Finally, we will also need to approximate F_{total} for the two grammars.

Now we will describe how to make approximations of the variables needed to calculate grammatical description length differences. First, we need to approximate the

variables $F_{\text{total,orig}}$ and $F_{\text{total,new}}$, which are the frequency total of all symbols used in the whole of the speaker's original and new grammars. This actually means only approximating $F_{\text{total,orig}}$ because we can calculate $F_{\text{total,new}}$ from $F_{\text{total,orig}}$ by adding the total number of additional symbols present in the new versus original grammatical description. The approximation of $F_{\text{total,orig}}$ heavily depends on whether vocabulary words and syntax rules are encoded separately or together. If these are encoded together, the number of symbols will be heavily dominated by the number of vocabulary words. Estimates of child–adult vocabulary size range from a thousand to tens of thousands (Beck & McKeown, 1991; Nation & Waring, 1997). Alternatively, if we were to encode syntax rules and vocabulary words separately, we would estimate $F_{\text{total,orig}}$ to be around a few hundred symbols. Fortunately, due to the behavior of logarithmic growth, whether $F_{\text{total,orig}}$ is estimated to be large or very small does not significantly affect the results. Thus, we will report results for a lower and upper estimate of $F_{\text{total,orig}}$ from 200 to 100,000. Note from Eq. 1 that the larger $F_{\text{total,orig}}$ is, the more costly the grammar. Given an assumption for $F_{\text{total,orig}}$, we can then calculate $F_{\text{total,new}}$ based on differences in the total number of symbols between the new versus original grammar. The next variable that needs approximating is the frequency of symbols that change between the original versus new grammar. As mentioned above, the specific individual values of f_s for symbols that do not differ between the two grammars, $f_{\text{s,irrelevant}}$, do not affect grammar length differences. Only their total value, $\sum_{s,\text{irrelevant}} f_{s,\text{orig}}$, is important. This total value can be calculated from the difference between the above approximated $F_{\text{total,orig}}$ and the known sum $\sum_{s,\text{relevant}} f_{s,\text{orig}}$, which can be obtained from the explicit descriptions of relevant grammar portions.

Equation 3 shows that the contributions to grammar length differences from the frequencies of symbols that do not change depends only on their total summed value.

$$\begin{aligned} &- \sum_{s,\text{irrelevant}} f_{s,\text{orig}} \left(\log_2 \frac{f_{s,\text{orig}}}{F_{\text{total,new}}} - \log_2 \frac{f_{s,\text{orig}}}{F_{\text{total,orig}}} \right) \\ &= - \left(\log_2 \frac{F_{\text{total,orig}}}{F_{\text{total,new}}} \right) \sum_{s,\text{irrelevant}} f_{s,\text{orig}} \\ &= - \left(\log_2 \frac{F_{\textit{total},\text{orig}}}{F_{\text{total,new}}} \right) \left(F_{\text{total,orig}} - \sum_{s,\text{relevant}} f_{s,\text{orig}} \right) \end{aligned} \tag{3}$$

The left-hand side of Eq. 3 is the difference in contribution to grammar length from symbols that do not change between new and old grammars, that is, difference in the first component of Eq. 1, summed over symbols that do not change, *s, irrelevant* evaluated for new and old grammars:

$$\sum_{s,\text{irrelevant}} f_{s,\text{new}} \log_2 \frac{f_{s,\text{new}}}{F_{\text{total,new}}} \quad \text{and} \quad \sum_{s,\text{irrelevant}} f_{s,\text{old}} \log_2 \frac{f_{s,\text{old}}}{F_{\text{total,old}}}.$$

Using the fact that $f_{\text{new}} = f_{\text{orig}}$ (so these terms drop out of the equation) and by taking constants independent of s out of the summation, we arrive at the middle of Eq. 3. Finally, we use the fact that

$$\sum_{s,\text{irrelevant}} f_{s,\text{orig}} = \left(F_{\text{total,orig}} - \sum_{s,\text{relevant}} f_{s,\text{orig}} \right)$$

to arrive at the right-hand side of Eq. 3. Now, Eq. 1 can be used directly to write out the difference between new and old grammars. We then use Eq. 3 to replace the sum over symbols that do not change between grammars, $\sum_{s,\text{relevant}} f_{s,\text{orig}}$, with a sum over symbols that do change, $\sum_{s,relevant} f_{s,\text{orig}}$ so that the resulting equation only depends on symbols that differ between grammars. Thus, using Eqs. 1 and 3, we are able to use the following equation to evaluate coding length differences between new and original grammars, Δ_{grammar}:

$$\begin{aligned}\Delta_{\text{grammar}} = &- \sum_{s,\text{relevant}} \left(f_{s,\text{new}} \log_2 \frac{f_{s,\text{new}}}{F_{\text{total,new}}} - f_{s,\text{orig}} \log_2 \frac{f_{s,\text{orig}}}{F_{\text{total,orig}}} - f_{s,\text{orig}} \log_2 \frac{F_{\text{total,orig}}}{F_{\text{total,new}}} \right) \\ &- F_{\text{total,orig}} \log_2 \frac{F_{\text{total,orig}}}{F_{\text{total,new}}} + (\Delta N_{\text{rules}} + \Delta N_{\text{symbols}}) C_{\text{prob}}\end{aligned} \tag{4}$$

Here ΔN_{rules} and $\Delta N_{\text{symbols}}$ are the differences in the number of rules and symbols, respectively, between new versus original grammars. $f_{\text{s,orig}}$ and $f_{\text{s,new}}$ are occurrence frequencies of symbol s in the original and new grammars, respectively. $F_{\text{total,orig}} = \sum_s f_{s,\text{orig}}$ and $F_{\text{total,new}} = \sum_s f_{s,\text{new}}$ are the total occurrence frequencies of all symbols in the whole original and new grammars, respectively. C_{prob} is the constant length for encoding probabilities to fixed decimal accuracy for symbols in the grammar and rules in the language. Using Eq. 4, and an assumed value of $F_{\text{total,orig}}$, it is now straightforward to use the explicitly defined new versus original grammars, such as in Example 4, to obtain grammar encoding differences. (Note that in the grammar definition everything within brackets is considered a single symbol). In Appendix S2, we show a sample detailed calculation of grammar length differences using the linguistic restriction on *is* contraction.

3.2.4. Calculating language encoding cost differences under new versus original grammars

The complex grammar allows for a more efficient encoding of language: Under the complex grammar, every time *donate* occurs, we save $-\log_2(0.2) = 2.3$ bits. This is because, under the complex grammar, 0 bits are required to encode which form *donate* occurs in because we already know that *donate* can only occur in the prepositional form. In practice, the encoding length savings will also contain the weighted contributions from all other linguistic situations whose encoding costs change. In Example 4, this includes the two alternations of *give* as well (i.e., differences in encoding lengths of *give-dative* and *give-prepositional* under new vs. original grammars) as well as *donate*. In order to calculate total encoding savings, we assume that the relative occurrence frequencies estimated from corpora for each linguistic situation is representative of their average relative occurrence probabilities in language. Total language encoding savings under the new grammar can be reported with respect to the number of occurrences of a particular linguistic situation

(e.g., encoding savings per occurrence of *donate*). We can then calculate the total savings per occurrence of the *j*th relevant linguistic situation, TotalSavings_j, by summing the frequency-weighted encoding gains/losses over all situations and dividing by the frequency of occurrence for the *j*th specific situation.

$$\text{TotalSavings}_j = \frac{\sum_i \text{freq}_i \times \text{save}_i}{\text{freq}_j},$$

where freq_i is the frequency of the *i*th relevant linguistic situation and save_i is the difference in encoding costs under the new versus original grammars. Now, given the grammar length cost and the encoding length savings, we can estimate how much language exposure is necessary before the more complicated grammatical rule is learned. In MDL terms, this corresponds to the amount of data needed before the more complicated grammar is the most worthwhile investment, that is, yields a smaller overall description length.

3.2.5. Using corpora to approximate learner's input

The crucial step in assessing learnability of a specific grammatical rule is to make an assumption about the grammatical input to the child learner by choosing a representative corpus. It is from the chosen corpus that one estimates the occurrence frequencies for the construction that is being learned as well as the relative probabilities of occurrence for the other constructions relevant to the two grammars. For example, if we were evaluating the learnability of the restriction on the dative alternation of *donate*, as described in Example 4, we would need to evaluate the expected probability of the prepositional and direct form. For the original grammar, the probability of the prepositional form would be estimated using occurrences of both *donate* and *give*. For the new grammar, the probability of the prepositional form would be estimated using only occurrences of *give*. We then would use the corpora to estimate the occurrence frequency of *donate*.

The most appropriate type of corpus for our purposes is one of child-directed speech. Here, we use the Brown and Bates corpuses from the CHILDES database (Mac Whinney, 1995). However, many of the constructions do not occur often enough to allow for estimates of their occurrence frequencies to be statistically significant. Therefore, we will also provide analysis based on additional corpus sets using the spoken and full portions of the BNC (Davies, 2009) and the spoken and full portions of the Corpus of Contemporary American English (COCA) (Davies, 2008). The Bates corpus contains transcripts of child-directed speech for 27 children at ages 1 year 8 months and 2 years 4 months. The Brown corpus contains transcripts of child-directed speech for 3 children ranging from ages 1 year 6 months to 5 years 1 month. In the combined Bates and Brown corpora, there are a total of 342,202 words worth of child-directed speech. The BNC contains 100 million words (90% written, 10% spoken). Written excerpts include extracts from periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, and school and university essays. The spoken part consists of unscripted informal conversations (recorded by demographically balanced volunteers selected from different age, region, and social classes) and spoken language from wide-ranging contexts such

as formal meetings to radio shows and phone-ins. The COCA consists of 385 million words (79 million spoken) equally sampled from spoken, fiction, popular magazine, newspaper, and academic sources. It was updated with 20 million words each year starting in 1990. Children in working class families hear an average of 6 million words per year (Hart & Risley, 1995). Using these numbers, we can estimate the proportion with which a construction occurs within the corpus (by dividing the construction occurrence frequency by the corpus word count) and multiply this proportion by 6 million words to approximate the occurrence frequencies in a year of a 2- to 5-year-old child's experience.

4. Results

In the section, we show learnability results for 19 linguistic constructions that we examine using an example instantiation of our framework. We will examine restrictions on the contractions of *want to*, *going to*, *is*, *what is*, and *who is*. We will examine the optionality of *that* reduction. We will examine restrictions on the dative alternation for the following verbs: *donate*, *whisper*, *shout*, *suggest*, *create*, and *pour*. Finally, we will also examine fixed transitivity for the following verbs: *disappear*, *vanish*, *arrive*, *come*, *fall*, *hit*, and *strike*. For each of these constructions, we will provide an overview of the linguistic restriction to be learned, a description of our proposed new versus original grammars, and then the results of our quantitative analysis (see Appendix S3 for the full list of original vs. new grammatical descriptions). Some of our chosen constructions had clear learnability results agreed upon by all corpora, whereas others had results that varied much more among corpora. Again, we stress that our analysis is not tied to any particular learnability outcomes for the debate on PoNNE and POS, but simply provides a quantitative method of assessing an upper bound on learnability. The results of our analysis show that some restrictions are easily learnable by an ideal learner from language statistics alone, whereas others are not.

Only the contractions of *going to*, *what is*, *who is*, and *is* are analyzed using the CHILDES corpus because these were the only ones that occurred in CHILDES with high enough frequency. Additionally, for these and the rest of the constructions, we show results using both spoken and entire portions of the BNC and COCA. We will refer to these in the text as spoken BNC, all BNC, spoken COCA, and all COCA. Not all constructions occurred in the spoken portions of these corpora, but for convenience we will display them in the same table. Although BNC and COCA are not necessarily the most accurate representations of child-directed speech, they still may serve as a rough approximation as they are a reflection of the relative frequencies of the constructions in general language. More precise estimates can be made by creating/using large corpora that are carefully collected to represent gradual and accumulated language experience. This is an avenue for further research.

As mentioned above, the approximation of $F_{\text{total,orig}}$ (the estimated total number of symbols used in a child's original grammar) depends on whether vocabulary words and syntax rules are encoded separately or together. However, because of the behavior of logarithmic growth, the relative results remain similar for a large range of $F_{\text{total,orig}}$ estimates. Here, we report results for estimates of $F_{\text{total,orig}} = 200$ and 100,000 for the following: encoding

length differences in original versus new grammars, encoding length savings for each occurrence of the construction being examined, the estimated number of occurrences needed for the construction to be learnable under MDL, the estimated occurrence frequency in 1 year of a child's language experience based on the frequencies found in the corpus, and the estimated number of years a child would need to learn the construction based on MDL. We use the approximation that a child hears about 6 million words per year (Hart & Risley, 1995).

4.1. Specific constructions and learnability results

We present our results as follows: We group the 19 constructions we have analyzed into similar classes (e.g., alternating verbs, transitive verbs, different contractions). For each class, we first explain the rules to be learned and describe the original and new grammar being considered for learning. Then for each construction class, we present a table summarizing results from each corpus. First, we report encoding length savings, and projected occurrences in a year, the two values that do not depend on estimated values of $F_{\text{total,orig}}$. We then report other values that do depend on estimated values of $F_{\text{total,orig}}$. The columns in our results tables are as follows: (1) ΔL_{d}, encoding length savings per construction occurrence (MDL Part 2). (2) $O_{1\text{yr}}$, the number of projected occurrences in 1 year of a child's language experience. This is calculated as follows: $O_{1\text{yr}} = (T_{\text{year}}/T_{\text{corpus}}) \times O_{\text{corpus}}$, where T_{year} is the total number of words per year a child hears on average (approximated as 6 million), T_{corpus} is the total number of words in the combined corpus (342,202 for CHILDES, 10 million for spoken BNC, 100 million for all of BNC, 79 million for spoken COCA and 385 million for all of COCA), O_{corpus} is the total number of occurrences in the corpus. (3) $\Delta L_{\text{g,200}}$, differences in grammar encoding costs (MDL Part 1), assuming $F_{\text{total,orig}} = 200$. (4) N_{200}, the number of construction occurrences needed to make learning the more complicated grammar ''worthwhile.'' (5) Y_{200}, the estimated amount of time in years necessary for acquisition of the language construction by an ideal learner assuming $F_{\text{total,orig}} = 200$. This is calculated as follows: $Y_{200} = N_{200}/O_{1\text{yr}}$, where N_{200} is the number of occurrences needed to learn the construction. (6) $\Delta L_{\text{g,100000}}$. (7) N_{100000}. (8) Y_{100000}. Columns 6–8 are the same as Columns 3–5 but for assumed value of $F_{\text{total,orig}} = 100{,}000$. These results assume that the child learner receives evenly distributed language input throughout all years, which is obviously untrue. However, these numbers provide a rough approximation of relative learnability for the different constructions by an ideal learner. In summary, the symbols for our results table are as follows:

ΔL_{d}: Encoding length savings per construction occurrence (bits)
$O_{1\text{yr}}$: Estimated occurrence frequency in a year of language experience
$\Delta L_{\text{g,200}}$ and $\Delta L_{\text{g,100000}}$: Grammar cost differences (bits)
N_{200} and N_{100000}: Number of occurrences needed for learning
Y_{200} and Y_{100000}: Number of projected years needed to learn
BNC_{sp}: British National Corpus, spoken
BNC_{all}: British National Corpus corpus, all
COCA_{sp}: Corpus of Contemporary American English, spoken

$COCA_{all}$: Corpus of Contemporary American English, all
CHIL: CHILDES corpus (if available)

Contraction of want to:

a. Which team do you want to beat?
b. Which team do you wanna beat?
c. Which team do you want to win?
d. *Which team do you wanna win?

Want to can be contracted to *wanna* under most linguistic circumstances. However, the linguistic restriction arises in cases exemplified by the following: In (a,b) contraction is allowed, whereas in (c,d) contraction is not allowed. In general, contraction is not allowed in a wh-question when the wh-word refers to a subject of the infinitive verb. The difference between the two example sentences above is the difference between the implied object of the word *want*. In the first sentence, *you* want *you* to beat some team and *you* is both the implied object and the subject of *want*. In the second sentence, *you* want a *team* to win. Here, the implied object is *team*, not *you*. Crain and Lillo-Martin (1999) have stated, ''this restriction is a prime candidate for universal, innate knowledge.''

Under the original grammar, we assume that contraction of *want to* is always allowed. Under the new grammar, contraction of *want to* is only allowed when the implied object and subject of *want* are the same. See Table 2 for analysis results.

Contraction of going to:

a. I'm going to help her.
b. I'm gonna help her.
c. I'm going to the store.
d. *I'm gonna the store.

When *going* introduces an infinitive verb such as in (a,b), contraction is allowed. When *going* is used to introduce a prepositional phase such as in (c,d), contraction is not allowed. There has been significant discussion as to the exact explanation for the restriction on *going to* contraction, which includes morpholexical and intonational arguments (Park, 1989; Pullum, 1997). Under our original grammar, we assume contraction of *going to* is always allowed. Under our new grammar, contraction of *going to* is allowed when *going* introduces an infinitive verb (i.e., *going to help*) and not allowed when *going* introduces a prepositional phrase (i.e., *going to the store*). See Table 3 for analysis results.

Table 2
Learnability results for *want to*

want to	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	0.4	1.2	158	386	321.8	283	691	575.6
BNC_{all}	0.13	0.2	158	1,265	5,271	283	2,263	9,427
$COCA_{sp}$	0	0	158	NA	NA	283	NA	NA
$COCA_{all}$	0.03	0.3	158	5,694	21,493	283	10,185	38,442

Table 3
Learnability results for *going to*

going to	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	1.0	108.6	112.3	108	1.0	201.6	194	1.8
BNC_{all}	0.5	36.9	112.3	241	6.5	201.6	432	11.7
$COCA_{sp}$	0.1	55.9	112.3	1,837	32.9	201.6	3,298	59.0
$COCA_{all}$	0.1	37.0	112.3	1,187	32.1	201.6	2,132	57.6
CHIL	0.80	438	112.3	140	0.32	201.6	252	0.58

Contraction of is:

a. Jane is taller than John.
b. Jane's taller than John.
c. Jimmy is shorter than she is.
d. *Jimmy is shorter than she's.

Is can be contracted if it occurs in the middle of the sentence (a,b), but not at the end (c,d) (Crain, 1991; Crain & Lillo-Martin, 1999; Lightfoot, 1998a; Pullum & Scholtz, 2002). There are also several other restrictions on the contraction of *is* (Lightfoot, 1998a). Thus, restrictions on *is* contraction could also be analyzed including other more general restriction rules. For our current analysis, we will only explore restrictions on *is* contraction in individual contexts such as at the end of a sentence, as described here, and in the two other contexts below. The fact that children never apparently make incorrect contractions of *is* has been used as an argument for evidence of Universal Grammar (Crain & Lillo-Martin, 1999; Lightfoot, 1998a). Others have argued that *is* contraction can be learned under the rule that contraction only occurs when *is* is not stressed in the sentence (Pullum & Scholtz, 2002). Our results show that *is* contraction in this context is very easily learnable from language statistics under MDL.

Under the original grammar, we assume contraction of *is* is always allowed. Under our new grammar, we assume contraction of *is* is not allowed when is occurs at the end of the sentence. See Table 4 for analysis results.

Contraction of what is/who is:

a. What is your name?
   Who is here?
b. What's your name?

Table 4
Learnability results for *is*

is	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	1.6	4,655.4	112.3	71	0.0	201.6	128	0.0
BNC_{all}	0.9	1,352.5	112.3	132	0.1	201.6	237	0.2
$COCA_{sp}$	3.7	3,265	112.3	30.3	0.01	201.6	54.4	0.017
$COCA_{all}$	1.1	1,723.3	112.3	103	0.1	201.6	185	0.1
CHIL	2.9	6,698	112.3	39	0.006	201.6	71	0.01

Who's here?
c. What is it?
Who is it?
d. *What's it?
*Who's it?

What is and *who is* can be contracted to *what's* and *who's*, only if the phrase is not followed by an *it*, which then terminates the sentence or phrase. In (a,b) contraction is allowed, whereas in (c,d) contraction is not allowed.

Under the original grammar, we assume that contraction of *what is/who is* is always allowed. Under the new grammar, if *what is/who is* is followed by *it* and a punctuation, contraction is not allowed. See Tables 5 and 6 for analysis results.

Optionality of ''that'':

a. Who do you think mom called?
b. Who do you think that mom called?
c. Who do you think called mom?
d. *Who do you think that called mom?

Wh-questions (e.g., questions beginning with *Who* and *What*) generally may have complement clauses that begin with or without *that*. The linguistic restriction rules depend on the wh-trace (trace of the wh-word). The wh-trace is the empty position where the noun replacing the wh-word would occur in the answer to the Wh-question. For example, sentence (a) could be answered with *I think mom called dad*. Here, the wh-trace (i.e., position of replacement noun *dad*) is at the end of the sentence. In this situation, *that* can be either present or

Table 5
Learnability results for *what is*

what is	ΔL_{d}	$O_{1\mathrm{yr}}$	$\Delta L_{\mathrm{g},200}$	N_{200}	Y_{200}	$\Delta L_{\mathrm{g},100000}$	N_{100000}	Y_{100000}
$\mathrm{BNC_{sp}}$	1.6	240.0	112.3	71	0.3	201.6	127	0.5
$\mathrm{BNC_{all}}$	1.3	60.2	112.3	85	1.4	201.6	153	2.5
$\mathrm{COCA_{sp}}$	1.6	68.4	112.3	70	1.0	201.6	126	1.8
$\mathrm{COCA_{all}}$	1.3	50.2	112.3	84	1.7	201.6	151	3.0
CHIL	2.1	1,2589	112.3	54	0.004	201.6	97	0.008

Table 6
Learnability results for *who is*

who is	ΔL_{d}	$O_{1\mathrm{yr}}$	$\Delta L_{\mathrm{g},200}$	N_{200}	Y_{200}	$\Delta L_{\mathrm{g},100000}$	N_{100000}	Y_{100000}
$\mathrm{BNC_{sp}}$	2.4	40.2	112.3	46	1.1	201.6	83	2.1
$\mathrm{BNC_{all}}$	0.8	10.3	112.3	149	14.5	201.6	267	26.0
$\mathrm{COCA_{sp}}$	1.2	6.5	112.3	94	14.4	201.6	169	25.9
$\mathrm{COCA_{all}}$	0.8	6.4	112.3	140	21.7	201.6	251	39.0
CHIL	2.0	351	112.3	55	0.16	201.6	99	0.28

absent when the wh-trace is at the end of a sentence (a,b). Sentence (b) could be answered with *I think dad called mom.* Here, the wh-trace (i.e., position of replacement noun *dad*) is at the beginning of the complement clause and the reduction of *that* is mandatory for the interrogative. For a more explicit description of the restriction on *that* reduction, see the grammatical descriptions in Appendix S3. The restriction regarding the reduction of *that* has been a classic central example in PoNNE arguments. Many have maintained that children do not receive the language experience needed to learn this restriction (Crain, 1991; Crain & Lillo-Martin, 1999; Haegeman, 1994).

Under the original grammar, we assume all complement clauses in wh-questions may begin with or without *that*. Under the new grammar, *that* must be omitted (i.e., mandatory *that* reduction) before the complement clause when the wh-trace is at the beginning of the complement clause. See Table 7 for analysis results.

Dative alternations:

a. I gave a book to the library.
   I told the idea to her.
   I made a sculpture for her.
   I loaded the pebbles into the tank.
b. I gave the library a book.
   I told her the idea.
   I made her a sculpture.
   I loaded the tank with pebbles.
c. I donated a book to the library.
   I shouted/whispered/suggested the idea to her.
   I created a sculpture for her.
   I poured the pebbles into the tank.
d. *I donated the library a book.
   *I shouted/whispered/suggested her the idea.
   *I created her a sculpture.
   *I poured the tank with pebbles.

The verbs *give*, *tell*, *make*, and *load* can undergo the dative alternation, which means that *give* can be used in both the prepositional and direct construction (a,b). A language learner may mistakenly also assume that the respective semantically similar verbs *donate*, *shout*, *whisper*, *suggest*, *create*, and *pour* can also undergo the dative alternation (c,d) when in

Table 7
Learnability results for *that*

that	ΔL_{d}	$O_{1\mathrm{yr}}$	$\Delta L_{\mathrm{g},200}$	N_{200}	Y_{200}	$\Delta L_{\mathrm{g},100000}$	N_{100000}	Y_{100000}
$\mathrm{BNC_{sp}}$	0.1	32.4	247.6	2,643	81.6	416.7	4,448	137.3
$\mathrm{BNC_{all}}$	0.05	10.4	247.6	4,694	449.6	416.7	7,900	756.7
$\mathrm{COCA_{sp}}$	0.3	86.4	247.6	984	11.4	416.7	1,656	19.2
$\mathrm{COCA_{all}}$	0.1	24.3	247.6	1,703	70.1	416.7	2,866	118.0

fact, these verbs are only allowed in the prepositional construction. The dative alternation restriction was one of the original examples cited by Baker as being puzzling in light of the PoNNE (also known as Baker's paradox). These restrictions on the dative alternation have been one of the primary examples of the learnability paradox under PoNNE (Baker, 1979; Bowerman, 1988; Fodor & Crain, 1987; Gropen, Pinker, Hollander, Goldberg, & Wilson, 1989; Mac Whinney, 1987; Mazurkewich & White, 1984; Pinker, 1989; White, 1987). A notable point concerning dative alternations, is that children *do* overgeneralize this restriction in speech production (Bowerman, 1988; Gropen et al., 1989). This is in contrast to other restrictions that are rarely uttered and for which the overgeneral forms are usually not explicitly produced in children's speech. Our ability to eventually learn this dative restriction as adults has been widely used to support the importance of innate linguistic knowledge (Baker, 1979; Gropen et al., 1989; Pinker, 1989). The knowledge of this restriction has been argued to require innate knowledge of morphophonological and semantic criteria (Gropen et al., 1989; Mazurkewich & White, 1984). Several studies have examined children's knowledge of the restrictions on the dative alternation (Gropen et al., 1989; Mazurkewich & White, 1984; Theakston, 2004).These studies found that children acquired grammatical knowledge for low-frequency verbs at later ages than for high-frequency verbs.

For our current learnability analysis, we treat each semantically similar category of verbs uniquely. Under the original grammar, we assume that *give* and *donate* are in the same category of ''giving verbs,'' *tell*, *shout*, *whisper*, and *suggest* are all in the same category of ''telling verbs,'' *make* and *create* are in the same category of ''making verbs,'' *fill* and *pour* are in the same category of ''filling verbs,'' and hence all can undergo dative alternation. Under the new grammar, the relevant individual verbs in each of the categories can only appear in the prepositional construction. See Tables 8–13 for analysis results.

Table 8
Learnability results for *donate*

donate	ΔL_{d}	$O_{1\mathrm{yr}}$	$\Delta L_{\mathrm{g},200}$	N_{200}	Y_{200}	$\Delta L_{\mathrm{g},100000}$	N_{100000}	Y_{100000}
$\mathrm{BNC_{sp}}$	4.9	7.2	44.9	9	1.3	89.7	18	2.5
$\mathrm{BNC_{all}}$	2.6	15.7	44.9	17	1.1	89.7	34	2.2
$\mathrm{COCA_{sp}}$	4.3	14.8	44.9	10	0.7	89.7	21	1.4
$\mathrm{COCA_{all}}$	3.4	15.0	44.9	13	0.9	89.7	26	1.7

Table 9
Learnability results for *shout*

shout	ΔL_{d}	$O_{1\mathrm{yr}}$	$\Delta L_{\mathrm{g},200}$	N_{200}	Y_{200}	$\Delta L_{\mathrm{g},100000}$	N_{100000}	Y_{100000}
$\mathrm{BNC_{sp}}$	5.0	1.2	44.9	9	7.5	89.7	18	15.0
$\mathrm{BNC_{all}}$	4.6	3.5	44.9	10	2.8	89.7	20	5.5
$\mathrm{COCA_{sp}}$	6.7	0.7	44.9	7	9.8	89.7	13	19.5
$\mathrm{COCA_{all}}$	6.7	2.3	44.9	8	3.4	89.7	16	6.8

Table 10
Learnability results for *whisper*

whisper	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	NA	0.0	44.9	NA	NA	89.7	NA	NA
BNC_{all}	5.3	0.5	44.9	9	17.8	89.7	17	35.6
$COCA_{sp}$	6.9	0.2	44.9	7	28.6	89.7	13	57.2
$COCA_{all}$	5.9	1.2	44.9	8	6.5	89.7	15	12.9

Table 11
Learnability results for *suggest*

suggest	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	NA	0.0	44.9	NA	NA	89.7	NA	NA
BNC_{all}	5.3	0.5	44.9	9	17.8	89.7	17	35.6
$COCA_{sp}$	NA	0.0	44.9	NA	NA	89.7	NA	NA
$COCA_{all}$	6.2	0.2	44.9	7	35.7	89.7	14	71.3

Table 12
Learnability results for *create*

create	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	0.4	6.6	44.9	127	19.2	89.7	253	38.4
BNC_{all}	0.2	6.9	44.9	187	27.0	89.7	373	54.1
$COCA_{sp}$	0.3	12.1	44.9	129	10.7	89.7	259	21.4
$COCA_{all}$	0.3	12.9	44.9	132	10.2	89.7	263	20.4

Table 13
Learnability results for *pour*

pour	ΔL_d	O_{1yr}	$\Delta L_{g,200}$	N_{200}	Y_{200}	$\Delta L_{g,100000}$	N_{100000}	Y_{100000}
BNC_{sp}	NA	11.4	44.9	NA	NA	89.7	NA	NA
BNC_{all}	0.4	12.8	44.9	124	9.7	89.7	247	19.4
$COCA_{sp}$	0.4	7.3	44.9	115	15.8	89.7	230	31.5
$COCA_{all}$	0.4	23.7	44.9	110	4.6	89.7	220	9.3

Fixed transitivity:

a. I hid the rabbit.
   I landed the plane.
   I dropped the ball.
   I pushed him.

b. The rabbit hid.
   The plane landed.
   The ball dropped.
   I pushed.

c. I disappeared/vanished the rabbit
   *I came/arrived the train.
   *I fell the ball.
   I struck/hit him.
d. I disappeared/vanished.
   The train came/arrived.
   The ball fell.
   *I struck/hit.

The verbs *hide*, *land*, *drop*, and *push* can occur both transitively and intransitively. This means they can appear both with and without an object (a,b). A language speaker may mistakenly also assume that the respective semantically similar verbs *disappear* (or *vanish*), *come* (or *arrive*), *fall*, and *hit* (or *strike*) can also occur both transitively and intransitively (c,d) when in fact, *disappear*, *vanish*, *come*, *arrive*, and *fall* are intransitive only (cannot occur with an object), whereas *hit* and *strike* are transitive only (cannot occur without an object). The restrictions on verb transitivity are also among one of the most cited examples investigated in light of the PoNNE (Ambridge, Pine, Rowland, & Young, 2008; Brooks & Tomasello, 1999; Brooks, Tomasello, Dodson, & Lewis, 1999; Brooks & Zizak, 2002; Theakston, 2004). As with the dative alternations, children have been noted to make overgeneralization errors of transitivity in their speech production (Bowerman, 1988; Gropen et al., 1989).

Here we again treat each semantically similar category of verbs uniquely. Under the original grammar, we assume that *hide*, *disappear*, and *vanish* are in the same category of ''disappearing verbs,'' *land*, *come*, and *arrive* are all in the same category of ''arriving verbs,'' *drop* and *fall* are in the same category of ''falling verbs'' and *push*, *hit*, and *strike* are in the same category of ''violent action verbs'' and hence all appear both transitively and intransitively. Under the new grammar, the relevant individual verbs in each of the categories can only appear in the correct type of transitivity. See Tables 14–20 for analysis results.

4.2. Summary of learnability results

The grammatical complexity of learning the new rule is represented by the differences in encoding length between grammatical descriptions (i.e., encoding investment). The restrictions on verb alternation and verb transitivity were the least costly investments because here learning only involved recognizing that particular verbs should be part of a separate category. Contractions were the next least costly rule, requiring about twice as many bits as verb alternation and transitivity. This is because learning of contractions requires knowledge of the surrounding syntactic context in which the contraction is not allowed. In particular, *want to* is the most complicated contraction to learn because the relevant syntactic context required identifying the referent implied object. The most complicated/costly construction of all was the restriction on *that* reduction, which requires about five times as many bits as restrictions on verb alternation and transitivity. How difficult we predict a construction to

Table 14
Learnability results for *disappear*

disappear	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	0.2	61.8	44.9	206	3.3	89.7	412	6.7
BNC_{all}	0.3	117.2	44.9	129	1.1	89.7	258	2.2
$COCA_{sp}$	0.1	233.4	44.9	378	1.6	89.7	755	3.2
$COCA_{all}$	0.1	287.0	44.9	425	1.5	89.7	849	3.0

Table 15
Learnability results for *vanish*

vanish	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	0.3	4.2	44.9	128	30.6	89.7	257	61.2
BNC_{all}	0.5	37.7	44.9	92	2.4	89.7	185	4.9
$COCA_{sp}$	0.2	21.8	44.9	200	9.2	89.7	400	18.4
$COCA_{all}$	0.2	44.7	44.9	220	4.9	89.7	440	9.8

Table 16
Learnability results for *come*

come	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	0.0	957.0	44.9	1,583	1.7	89.7	3,165	3.3
BNC_{all}	0.1	487.5	44.9	422	0.9	89.7	844	1.7
$COCA_{sp}$	0.1	654.8	44.9	716	1.1	89.7	1,431	2.2
$COCA_{all}$	0.1	533.9	44.9	416	0.8	89.7	831	1.6

Table 17
Learnability results for *arrive*

arrive	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	0.2	58.8	44.9	236	4.0	89.7	471	8.0
BNC_{all}	0.2	151.1	44.9	205	1.4	89.7	411	2.7
$COCA_{sp}$	0.2	105.9	44.9	245	2.3	89.7	491	4.6
$COCA_{all}$	0.1	1,561.0	44.9	884	0.6	89.7	1,766	1.1

Table 18
Learnability results for *fall*

fall	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	1.2	70.2	44.9	39	0.6	89.7	77	1.1
BNC_{all}	0.5	87.8	44.9	90	1.0	89.7	181	2.1
$COCA_{sp}$	0.4	89.0	44.9	107	1.2	89.7	213	2.4
$COCA_{all}$	0.3	144.1	44.9	141	1.0	89.7	283	2.0

Table 19
Learnability results for *strike*

strike	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	0.9	48.6	44.9	51	1.0	89.7	102	2.1
BNC_{all}	1.4	61.7	44.9	31	0.5	89.7	63	1.0
$COCA_{sp}$	0.7	153.2	44.9	62	0.4	89.7	124	0.8
$COCA_{all}$	0.8	157.3	44.9	58	0.4	89.7	116	0.7

Table 20
Learnability results for *hit*

hit	ΔL_d	N_{1yr}	$\Delta L_{g,200}$	$N_{needed,200}$	Y_{200}	$\Delta L_{g,100000}$	$N_{needed,100000}$	Y_{100000}
BNC_{sp}	1.1	12.6	44.9	41	3.2	89.7	81	6.4
BNC_{all}	1.5	50.4	44.9	29	0.6	89.7	59	1.2
$COCA_{sp}$	1.6	39.5	44.9	28	0.7	89.7	57	1.4
$COCA_{all}$	1.2	44.9	44.9	37	0.8	89.7	74	1.7

learn depends on both the grammatical complexity of the rule (i.e., encoding investment cost) as well as how often the linguistic restriction occurs, both relative to alternative forms of the construction (i.e., encoding savings) and in absolute occurrence frequency (i.e., projected number of occurrences in a year of language experience).

Our results show that there is a significant split in learnability among the different constructions. Below we discuss general trends, focusing on the estimated number of years required for learning based on F_{total} = 100,000. Note that results for F_{total} = 200 will require roughly half the estimated time required for learning. Regardless of the value of F_{total} used, results in trends and implications on learnability are similar.

The restrictions on the contractions of *is* and *what is*, the dative alternation restriction on *donate* and the fixed transitivities of *come*, *fall*, and *strike* appear clearly highly learnable from language statistics alone under MDL. The restrictions on *is* seem to be the most easily learnable due to its extremely high occurrence frequency. Restrictions on *come*, *fall*, *hit*, and *what is*, all of which appear learnable in less than 4 years of language experience, also are highly learnable due to their relatively high occurrence frequencies. On the contrary, *donate* did not have particularly high occurrence frequencies. Instead, its learnability stems from the high encoding savings per appearance. Intuitively, this means that under the original grammar, the direct form of *donate* was expected to appear frequently relative to the prepositional form. Thus, the fact that *donate* does not appear in the direct form is easily noticed. Other constructions that had restrictions that appeared mostly learnable are *disappear*, *arrive*, *hit* and *shout*. Restrictions on *disappear*, *arrive*, and *hit* are estimated to take about 1–3 years from all the corpora except spoken BNC, which estimate these to take about 7–8 years. *Shout* is more easily learnable from the full corpora requiring an estimated 6–7 years for both all BNC and all COCA versus 15 and 20 years from spoken BNC and spoken COCA, respectively.

Among the most difficult to learn from language statistics under our chosen representations were restrictions on contraction of *want to*, optionality of *that* reduction, and dative

alternation of *suggest* and *create*. The restrictions on *want to* seem most clearly difficult to learn from language statistics alone, with the estimated years needed being thousands of years. This is due mostly to the extremely low occurrence frequency of *wanna* in the relevant context, although the high encoding investment costs and low encoding savings also contributed. Note that the low reported occurrence frequency of *want to* does not mean that appearances of *want to* and *wanna* were rare, but that appearance of *want to* in the form relevant for learning the linguistic restriction was rare (i.e., in wh-questions).[4]

The restriction on *that* reduction also required far longer than the length of a human childhood to acquire. Although the linguistic contexts relevant to learning restrictions on *that* insertion occurred reasonably often, this construction was difficult to learn due to its high grammatical complexity (extremely high encoding investment cost) and its very low encoding savings per occurrence. The low encoding savings reflects the fact that the reduced form of *that* (a clause not beginning with *that*) is much more common, occurring ~95% of the time, compared with the nonreduced form (a clause beginning with *that)*, which occurred ~5% of the time. This means *that* is often not present, even when its presence is allowed. Thus, a learner will have a harder time noticing the suspicious fact that it is always dropped in the cases when it is never allowed. Such low encoding savings also explains the difficulties of learning *suggest* and *create*. In general, experiments have shown that ''difficult'' constructions do take longer to learn. For example, the rules governing *that* insertion are not mastered in older children. A study by Gathercole showed that most fifth grade children could not identify ungrammatical insertions of *that* (Gathercole, 2002). However, most young adults are aware of the restrictions on *that* reduction (as well as the rest of our examined constructions) so learning does actually occur much earlier than our analysis would suggest. It is possible that the restrictions on these constructions can be represented as part of a different syntactic category or more general regularity, which would make it learnable from the available data. Alternatively, learning for these constructions could be achieved from numerous other sources previously mentioned, such as situational and communicational contexts, phonological cues, prosody, gestures, or innate language biases.

Some of the constructions, contractions of *going to* and *who is* and dative restrictions on *shout*, *whisper*, and *pour* and fixed transitivity of *vanish*, had estimated learnabilities that varied widely depending on the different corpus we used. These different results could arise for two reasons. First, the corpora may differ in the number of occurrences for the construction that has the linguistic restriction. Second, the encoding savings per occurrence may differ due to different ratios of occurrences for the constructions' alternate forms. Often, both reasons contribute to results differences because the greater frequency of occurrence of a given construction also often means it occurs more often relative to the alternative constructions. For example, *gonna* appears more often in CHILDES (O_{1yr} = 438) and spoken BNC (O_{1yr} = 109) relative to all BNC (O_{1yr} = 37), all COCA (37), and spoken COCA (56). Also, the relative frequency occurrences of *gonna* versus *going to* was much greater in the spoken BNC (~50% vs. 50%) and CHILDES (60% vs. 40%) than in COCA (5% vs. 95% for entire corpus as well as spoken only). This is reflected in the smaller savings per occurrence estimated from COCA (0.1 bits for spoken and entire corpus) versus BNC (1.0 bits for spoken and 0.5 bits for entire corpus), and CHILDES (0.8 bits). Hence, restrictions on *gonna* are more difficult according to

Spoken and all COCA (59 and 58 Years, respectively) and all BNC (12 Years) than according to CHILDES (0.6 Years) and spoken BNC (2 Years). The restriction on *who's it* occurred proportionally more often in CHILDES and next most often in spoken BNC (O_{lyr} = 351, 40, 10, 6, and 6 from CHILDES, spoken BNC, all BNC, spoken COCA, and all COCA, respectively). This is reflected in the estimated years required from the five corpora: 0.3, 2, 26, 26, and 39 years for CHILDES, spoken BNC, all BNC, spoken COCA, all COCA respectively. In these cases, child-directed speech appeared to have more contractions than adult corpora. The fact that *gonna* and *who's it* do appear so frequently based on estimates from the CHILDES corpus suggests that children do hear these contractions often enough to acquire it at an early age. The learnability of *whisper* varies significantly among the corpora, from an estimated 13 years using all COCA to infinite years using spoken BNC, which had no instance of *whisper*. Similarly spoken BNC had no occurrences of *load* in the direct form. Thus, it predicts that it is extremely difficult to learn the dative restriction on the similar verb *pour*, whereas COCA estimates learnability of this restriction in 9 years. The learnability of vanish ranges from 5 years based on all BNC to 61 years based on spoken BNC. The learnability of these constructions, based on the data and representations we have considered, remains unclear with our present analysis.

5. Discussion

The contribution of this paper is to provide a simple quantitative framework to assess the learnability of specific linguistic phenomena based on language statistics alone. Our framework allows for quantitative analysis of learnability under different assumptions regarding the formulation of the grammatical rule to be learned and the corpus that represents a learner's input. The purpose of our framework is to make these varying assumptions explicit and allow these assumptions to be varied and compared among one another in future work. By making these assumptions explicit, we can provide a common forum for quantifying and discussing the learnability of different linguistic constructions.

Although our framework does provide a method of estimating the amount of time an ideal learner would require for acquiring specific linguistic rules, this does not mean that our method would regularly predict concrete ages of acquisition in real children. Our methods only provide an upper bound on learnability for an idealized learner based on language statistics. However, measures of relative learnabity should give an indication for how relatively learnable constructions are in reality. Interestingly, even with our use of simple default assumptions, we are able to obtain some contrasting results on learnability: Some of the linguistic phenomena which have been viewed as raising puzzles for learnability appear to be learnable from a modest amount of data, whereas others seem to require vast quantities of data, which are not available to the child. Below we show comparison of our methods with existing data on grammar judgements in children. We also further discuss the possible implications results of our methods could have on child language learning.

5.1. Comparison with existing child data

It is difficult to pin down precisely the age at which children learn the various constructions. However, there have been a few studies that examine child grammatical knowledge of linguistic restrictions. These measures of grammatical knowledge can be compared with our results regarding the relative learnability of these constructions. Here we compare the results of learnability estimated from our method with experimental data from two studies. In both these cases, our predictions are in accordance with the experimental data. Furthermore, our assessment of learnability matches the experimental data better than entrenchment (Theakston, 2004; Tomasello, 2003), one of the other commonly used explanations for how children might retreat from overgeneralization errors. Entrenchment is the hypothesis that the likelihood of a child overgeneralizing a construction is related to the construction's input occurrence frequency.

We first compare our results with a study by Theakston (2004), which asks 5- and 8-year-olds whether ungrammatical, overgeneral uses of various constructions are acceptable. Here, we compare our results with the subset of their data that involves the verbs *shout*, *whisper*, *disappear*, *vanish*, *come*, *arrive*, *fall*, and *pour* appearing in the constructions we have analyzed. We predicted that the more learnable a construction was, the less likely the ungrammatical form would be accepted. Results for 5-years-olds show a very similar trend to that of the 8-year-olds so we will only show the 8-year-old data. Fig. 5a shows plots of the percentage of 8-year-old children who found an ungrammatical sentence acceptable versus MDL learnability, log(occurrences needed/projected occurrences in 1 year of child's experience), estimated using all BNC. Fig. 5b shows a similar plot but with occurrence frequency instead of learnability, representing the entrenchment hypothesis. Fig. 5a shows that there is a very linear trend between grammatical acceptability and our estimates of learnability with a noted outlier for the verb *pour*. Fig. 5b shows a less linear trend, but relatively monotonic relationship between grammatical acceptability with noted outlier of *pour* and an additional outlier of *fall.* Thus, it appears that our analysis of learnability using MDL matches the data better and more linearly than the entrenchment explanation. Results for spoken and all COCA (not shown) are very similar to the full BNC results we show. Spoken BNC results are less linear, perhaps due to the low occurrence frequency counts of many of the constructions in that corpus.

The second study is one by Ambridge et al. (2008), which elicits from 5- to 6-year-olds and 9- to 10-year-olds grammaticality judgements of grammatical and ungrammatical sentences for constructions susceptible to overgeneralization. This includes sentences involving three of the constructions we have analyzed: *disappear*, *vanish*, and *fall*. Here, we predicted that learnability would correlate with how much more grammatical the correct usage of the construction would be judged relative to the incorrect usage. In the data, for both groups of children, relative grammaticality was greatest for *fall*, less for *disappeared*, and least for *vanished*. Our assessments of learnability (occurrences needed/projected occurrences in 1 year of child's experience) scaled correctly with the relative perceived grammaticality. This was true for analysis based on both spoken and all parts of both BNC and COCA. Additionally, for all of our corpora except spoken BNC, the entrenchment hypothesis (i.e., pure frequency of occurrence) did not predict the right ordering of relative grammar judgements. This is

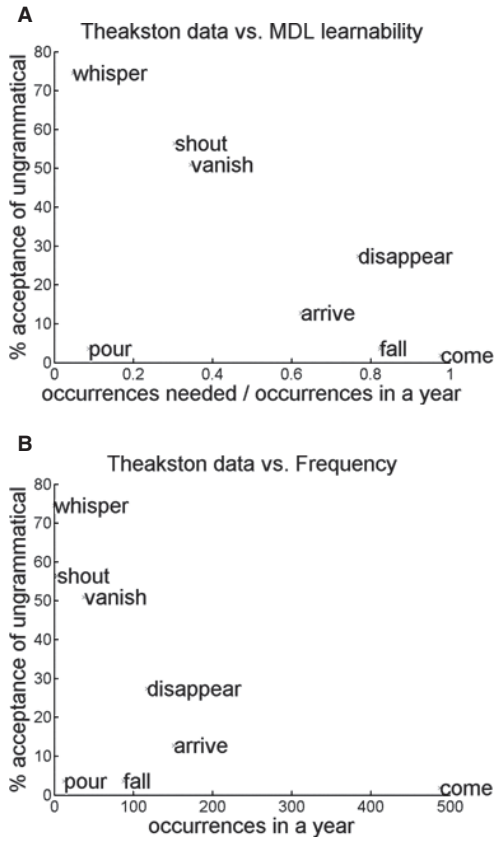

Fig. 5. Comparison of MDL analysis versus entrenchment hypothesis with data from Theakston (2004). The data are the percentage of 8-year-olds who found that ungrammatical, overgeneral uses of various constructions are acceptable. Here, we show results chosen from the subset of their data that involve the following constructions that we have analyzed: *shout*, *whisper*, *disappear*, *vanish*, *come*, *arrive*, *fall*, and *pour*. (A) Percentage of 8-year-old children who found an ungrammatical sentence acceptable versus MDL estimates of learnability (occurrences needed/projected occurrences in 1 year of child's experience) using the full BNC. (B) Similar plot but with occurrence frequency instead of learnability, representing the entrenchment hypothesis.

because *fall*, which has the greatest relative grammaticality judgment of the three constructions, does not have the highest projected 1 year occurrence frequency (except for in spoken BNC). However, despite its lower occurrence frequency, *fall* is estimated to be the most learnable of the three constructions due to its relatively high encoding savings and low new versus original grammatical description cost. Thus, our MDL analysis of learnability also appears to be more successful at matching this data than the entrenchment hypothesis.

5.2. Possible extensions

Here we use our framework to provide learnability results using a particular set of grammatical assumptions. However, we do not suggest that these are definitive answers. We do

not aim to resolve the PoNNE debate in general, nor for the specific cases presented here. The intention of our framework is to provide provisional results for both easy and difficult-to-learn constructions that serve to challenge and provoke arguments from different sides of the debate. If a construction is estimated to be difficult under our analysis, and is easily acquired by children at an early age, then further research is needed to find either alternative representations of the rule that are easier to learn or other cues in language that could facilitate learning.

For example, it is possible that with a different representation that used wider language contexts, constructions that seem unlearnable under our instantiation of MDL will become learnable. For example, maybe *want to* contractions would become learnable if they are framed as part of a more general linguistic rule that occurs more frequently in language. Another example is that actual verb learning may not be as incremental as we have presented here. Our analysis assumed each verb construction was learned individually, which results in a conservative estimate on learnability. Alternatively, it is possible that once an intransitive verb is learned, a general category for intransitive verbs may be formed, with which other verbs may quickly be automatically associated. Such an alternative assumption can also be incorporated into our framework. If a construction does not appear learnable under any representation from language statistics alone, researchers will need to look elsewhere for learning cues. Real children obtain language cues from a wide variety of linguistic sources. Thus, a construction that appears unlearnable under our analysis such as contraction of *want to* maybe be learnable in the context of other external cues, gestures, prosody, intonation, or phonology, or even with assumed innate language biases. Alternatively, some form of positive evidence may be available for these constructions such as implicit social cues, or explicit instruction in school. It is also possible that other domain-general learning mechanisms such as those more similar to recurrent neural networks may be employed during learning. Directed experiments could then be conducted to test these hypotheses. Additionally, learnability is of course greatly affected by the corpus used. A more carefully assembled corpus that matches the cumulative child experience of language may provide more accurate estimates of learnability. Finally, for constructions that appear theoretically learnable, it remains to be shown whether the assumed representations actually are those used by children. For example, it could be possible that children actually use representations that result in the construction being unlearnable. Furthermore, our analysis reflects that of an ideal learner and further experiments are required to evaluate the degree to which this corresponds to the behavior of real learners.

Although we do not intend our method to be a precise model of language learning, we believe our framework can be used to provide clues into the process of learning. In the case that learnability results for certain constructions using our method should indeed be corroborated by child experimental data regarding age of acquisition, this could present further questions about the mechanisms children use to acquire language. Psychological studies have shown that MDL principles are indeed used for learning structures of categories (Chater & Vitányi, 2002; Feldman, 2000). Further research can be done to examine whether a similar computation is actually employed by children during language learning. Our method assumes that children have the capacity to choose from reasonable alternative

grammars. One reason complete MDL models of natural language learning are difficult to implement is that the space of all possible grammars is too large to compute. If real child learners did use an MDL-related algorithm for learning, they would have to be capable of coming up with a finite space of possible alternative grammars. For learning to proceed at a reasonable pace, these should be grammars that *do* decrease overall encoding description length. Otherwise, the learner would be spending his or her energies inefficiently considering the infinite space of all possible more costly grammars and never learn the correct (i.e., least costly) one. Future research could examine how children are able to form such reasonably constrained hypotheses of alternate grammars, which contain the correct grammars that they eventually adopt as the correct ones by adulthood. Furthermore, our method also implies that language learners have some notion of the relative complexity of various grammatical rules. With increased language experience, the contribution of the grammar encoding length to the overall MDL evaluation becomes small relative to the language encoding length. Given that grammar length differences become less important as language experience increases, one may hypothesize that older children may be less sensitive to differences in grammatical complexity than younger ones are, and test if this is true.

We invite researchers to use our method to compare and contrast the different language representations that may prove most amenable for learning. We hope this will help direct and focus directions of further research on the extent to which children are able to acquire language without negative evidence.

Notes

1. Practical MDL results depend on the chosen form of language representation. In contrast, theoretical MDL learnability results show that, in the limit of infinite data, encoding lengths do not depend significantly on the chosen representation: Grammars encoded by different programming languages will not differ in encoding length by more than a constant that does not depend on the data (Vitányi & Li, 2000).
2. Here we include the term $N_{\text{symbols}}C_f$, which encodes the probabilities with which each symbol is used in the grammar. These probabilities determine the code (i.e., distribution of code symbols ''001,'' ''01,'' etc.) that allows for most efficient representation of the grammatical description. From this code, one can then establish the correspondence between each grammar symbol and its encoded representation. This term was not included in the equation in Dowman (2007), where it was assumed that all grammars used the same set of symbols. By allowing for a specifically tailored set of symbols, our formulation allows for more efficient grammar representations.
3. It is also possible to encode all sentences by assuming that repeats of linguistic phrases are encoded differently from the first occurrences—for example, by using a code that refers back to the previous occurrence, rather than regenerating occurrences from scratch. This corresponds to learning *types* of utterances rather than tokens. This encoding is commonly used in compression methods. However, due to presumed limits on human memory capacity, such an encoding scheme is only conceivable for a

limited number of very common utterances and is unlikely to be applied to the bulk of language data (although it is possible that this type of encoding might be used at the level of individual constructions, rather than whole utterances). In any case, the type-based approach will typically require *more* data for learning, because, in essence, it strips out any repeated linguistic constructions. As our concern is to outline the behavior of an ''ideal'' learner, given specific representations of the linguistic structure to be learned, we do not consider such ''type-based'' encoding schemes here. Notice, in particular, that any linguistic structure that is unlearnable under our present analysis will also be unlearnable given a type-based encoding scheme.

4. The use of a corpus inevitably raises the question that the transcriber may not have heard a contraction correctly. We feel that while possible, this possibility is small. More important, this is unlikely to be the cause of the low-occurrence frequencies of *wanna* appearing in relevant contexts. This is because, while the contracted form *wanna* does occur frequently in general, it is only this contracted form in the context of a wh-question that does not occur in CHILDES, and indeed also rarely in the other corpora we examine. Other contracted forms also occur frequently throughout the corpora we examine. Thus, due to the high frequency in all corpora of contracted forms in general, we believe that the low frequency of *wanna* in the wh-question form is genuinely reflective of a low occurrence frequency in language usage. This argument also applies to other cases where occurrence frequencies were low.

Acknowledgments

This research was supported by ESRC grant RES-000-22-2768 and a Major Research Fellowship from the Leverhulme Trust to Nick Chater.

References

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

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

Baker, C. L., & McCarthy, J. J. (1981). *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Bates, E., Marchman, V. A., Thal, D., Fenson, L., & Dale, P. (1994). Development and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*, 85–123.

Beck, I. L., & McKeown, M. G. (1991). Social studies texts are hard to understand: Mediating some of the difficulties. *Language Arts*, *68*, 482–490.

Bowerman, M.. (1988). The 'No Negative Evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Oxford, England: Blackwell.

Braine, M. D. S.. (1971). On two types of models on the internalization of grammars. In D. I. Slobin (Ed.), *The ontogenesis of grammar* (pp. 153–186). New York: Academic Press.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*, 294–301.

Brooks, P., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, *75*, 720–738.

Brooks, P., Tomasello, M., Dodson, K., & Lewis, L. (1999). Young children's overgeneralizations with fixed transitivity verbs. *Child Development*, *70*, 1325–1337.

Brooks, P., & Zizak, O. (2002). Does preemption help children learn verb transitivity? *Journal of Child Language*, *29*, 759–781.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–581.

Chater, N. (2004). What can be learned from positive data? Insights from an 'ideal learner'. *Journal of Child Language*, *31*, 915–918.

Chater, N., & Vitányi, P. (2002). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.

Chater, N., & Vitányi, P. (2007). Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*, 135–163.

Chomsky, N. (1955). *The logical structure of linguistic theory*. Phd dissertation, University of Pennsylvania.

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

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Christiansen, M. H., & Chater, N. (2007). Generalization and connectionist language learning. *Mind & Language*, *9*, 273–287.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 597–612.

Crain, S., & Lillo-Martin, D. (1999). *An introduction to linguistic theory and language acquisition*. Oxford, England: Blackwell.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 385 million words, 1990-present. Corpus of Contemporary American English [On-line]. Available at: http://www.americancorpus.org [accessed on June 24, 2010]

Davies, M. (2009). BYU-BNC: The British National Corpus. British National Corpus [On-line]. Available at: http://corpus.byu.edu/bnc [accessed on June 29, 2010]

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 107–112). Mahwah, NJ: Erlbaum.

Dowman, M. (2007, June). Using minimum description length to make grammatical generalizations. Presented at Machine Learning and Cognitive Science of Language Acquisition meeting at UCL, London. Available at: http://videolectures.net/mlcs07_dowman_umd/

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *403*, 630–633.

Fodor, J. D., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar*. New York: McGraw-Hill.

Fodor, J. D., & Crain, S. (1987). Simplicity and generality of rules in language acquisition. In B. Mac Whinney (Ed.), *Mechanisms of language acquisition. Proceedings of the 20th Annual Carnegie-Mellon Conference on Cognition* (pp. 35–63). Hillsdale, NJ: Lawrence Erlbaum.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2007). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 275–281). Austin, TX: Cognitive Science Society.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*, 300.

Gathercole, V. C. M. (2002). Monolingual and bilingual acquisition: Learning different treatments of that-trace phenomena in English and Spanish. In D. K. Oller & R. E. Eiler (Eds.), *Language and literacy in bilingual children* (pp. 220–254). New York: Multilingual Matters.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*, 219–224.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*, 153–198.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, *65*, 203–207.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In S. Scheler, S. Wernter, & E. Rilof (Eds.), *Connectionist, statistical and symbolic approaches to learning for natural language* (pp. 203–216). Berlin: Springer Verlag.

Haegeman, L. (1994). *Introduction to government and binding theory*. Oxford, England: Blackwell.

Hart, B., & Risley, J. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes Publishing.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypothesis with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago, IL, University of Chicago Press.

Langacker, R. W. (1991). *Foundations of cognitive grammar* (Vols. 2). Stanford, CA: Stanford University Press.

Langley, P., & Stromsten, S. (2000). Learning context-free grammars with a simplicity bias. *Proceedings of the eleventh European conference on machine learning* (pp. 220–228). Barcelona: Springer-Verlag.

Lightfoot, D. (1998a). Promises, promises, general learning algorithms. *Mind & Language*, *13*, 582–587.

Lightfoot, D. (1998b). *The development of language: Acquisition, change, and evolution*. Oxford, England: Blackwell Publishers.

Mac Whinney, B. (1987). The competition model. In B. Mac Whinney (Ed.), *Mechanisms of language acquistion* (pp. 249–308). Hillsdale, NJ: Erlbaum.

Mac Whinney, B. (1995). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacDonald, M. C. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*, 35–54.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, England: Cambridge University Press.

de Marcken, C. (1996). *Unsupervised Language Acquisition*. PhD dissertation, MIT.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.

Mazurkewich, I., & White, L. (1984). The learnability and acquisition of the dative alternation in English. *Cognition*, *16*, 261–283.

McClelland, J. L., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6–19). Cambridge, England: Cambridge University Press.

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

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition? In *Proceedings of the 24th annual conference of the Cognitive Science Society* (pp. 720–725). Mahwah, NJ: Erlbaum.

Park, N. S. (1989). Weight as a linguistic variable with reference to English. *Language Research*, *28*, 803–894.

Perfors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the twenty-eighth annual conference of the Cognitive Science Society* (pp. 663–668). Mahwah, NJ: Erlbaum.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
Pinker, S. (1994). *The language instinct*. New York: Harper Collins.
Pullum, G. (1997). The morpholexical nature of English to-contraction. *Language*, *73*, 79–102.
Pullum, G., & Scholtz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50.
Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1000–1028.
Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.
Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, *93*, 147–155.
Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*, 241–254.
Seidenberg, M. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*, 1599–1603.
Spencer, J. P., Blumberg, M. S., McMurray, B., Robinson, S. R., Samuelson, L. K., & Tomblin, J. B. (2009). Short arms and talking eggs: Why we should no longer abide the nativist-empiricist debate. *Child Development Perspectives*, *3*, 79–87.
Stolcke, A. (1994). *Bayesian learning of probabilistic language models*. Berkeley: Department of Electrical Engineering and Computer Science, University of California.
Theakston, A. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development*, *19*, 15–34.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Tomasello, M. (2004). What kind of evidence could refute the UG hypothesis? *Studies in Language*, *28*, 642–644.
Vitányi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *IT-46*, 446–464.
White, L. (1987). Children's overgeneralizations of the dative alternation. In K. Nelson & A. Van Kleeck (Eds.), *Children's language* (6th ed., pp. 261–287). Hillsdale, NJ: Erlbaum.
Yang, C. (2004). Universal Grammar, statistics, or both? *Trends in Cognitive Sciences*, *8*, 451–456.

Supporting Information

Additional Supporting Information may be found in the online version of this article on Wiley InterScience:

Appendix S1. Approximation of speaker's grammar specific-situation rules allow for a more efficient encoding of the language data by specifying more accurate rule probabilities for specific situations.

Appendix S2. Example calculation of MDL applied to a corpus: In order to clearly describe our methodology we will explain our calculations using the restriction on *is* contraction as an example.

Appendix S3. Specific-situation grammars.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.