A Cognitively Plausible Model for Grammar Induction

Roni Katzir
rkatzir@post.tau.ac.il

October 15, 2010

Abstract

This paper aims to bring theories of Universal Grammar (UG), studied by theoretical linguists, and theories of learning into closer contact. I use the observation that any theory of UG provides a learning criterion – the total memory space used to store a grammar and its encoding of the input – that supports learning. This mapping from theories of UG to learners, which I refer to as Bare-Minimum Learning (BML), maintains a minimal ontological commitment: the learner for a particular theory of UG uses only what that theory already requires to account for linguistic competence in adults. I argue that the learners obtained using BML are not only conceptually correct starting points for learning, but also manage to avoid some of the main empirical shortcomings of learning proposals in the literature. I provide a proof-of-concept implementation that will demonstrate the performance of BML using one theory of UG and several inputs.

1 Introduction

A central task in theoretical linguistics is constructing theories of competence – grammars (alternatively seen as computer programs) that have an opinion (a simple yes/no or a more fine-grained evaluation) about possible inputs. A broader goal of theoretical linguistics is characterizing the range of possible grammars that adult speakers can have. Thus, theoretical linguists agree that humans can mentally represent grammars from a set of possible candidates and use these grammars to analyze inputs. Of course, much disagreement remains about the correct competence theories and the characterization of the range of theories. The theory of the range of allowable grammars is often referred to as Universal Grammar (UG).[1] Starting with UG, the child reaches a particular grammar through exposure to a linguistic environment. As pointed out by Chomsky (1965), this view assigns a central role to learnability in investigating UG: a linguistic theory must specify a range of grammars that can be attained using the cognitive machinery and data available to the child.

[1]The term UG has sometimes been associated with approaches that assume a substantial innate component. Here I will use it neutrally. This paper makes no claims as to the correct theory of UG.

One might hope, then, that theories of competence and theories of learning would have a close relationship: that theories of UG would direct the learning process, and that theories of learning would restrict the choice of UG. This hope is evident in the evaluation metric of Chomsky (1955/1975) and Chomsky and Halle (1968), which suggested a way to use the representation provided by any theory of UG for learning. In practice, however, the evaluation metric has been largely abandoned, and the two domains have never succeeded in constraining one another. On the one hand, theories of UG have grown increasingly skeptical about learning. This has hindered the search for a general correspondence between theories of UG and theories of learning and has led instead to theories of learning that are specific to a particular theory of UG. Such UG-specific theories are often not principled, and they use more machinery than would seem necessary based on independent considerations. This makes it hard to choose between such theories, as will be discussed below. It also limits the flow in the opposite direction: if the mechanisms are arbitrary and hard to choose from, learning becomes less useful for deciding between theories of UG. If one particular learning algorithm for a constraint-based theory of phonology, for example, does better than a different learning algorithm for a rule-based theory of phonology, what does this show? What is missing is a tight, principled framework relating theories of UG to theories of learning.

My goal in this paper is to take a step toward restoring the collaboration originally envisioned in early generative grammar between UG and learning by providing exactly this kind of framework. This will allow the choice of UG to induce a learning solution, which, in turn, will allow empirical findings about learning to decide between theories of UG. The idea, discussed in detail below, relies on the trivial fact that any complete theory of UG provides us with a schema for storing certain grammars in memory and with a way to encode inputs with each of these grammars. Clearly, if this, or something very close to it, can account for learning, then it should be our default theory of learning: we need a complete theory of UG independently of learning, in order to account for linguistic competence in adults, so we should start our investigation of learning by checking what having a theory of UG gives us. Departures should only be made if the facts require them. I will try to show that adding two natural assumptions to these indeed provides us with such a principled and general learning mechanism. These additional assumptions are first, that during learning an additional grammar can be considered and second, that the two grammars can be compared with respect to the amount of memory space needed to store each grammar and its parse of the input. The result will be a general framework, which I refer to as Bare-Minimum Learning (BML), that maps any complete theory of UG to a learning mechanism based on the assumptions just mentioned. Given a theory of UG T, BML provides a learner that, given an input D, searches the space of grammars sanctioned by T for the grammar G that minimizes the amount of memory required to store both G (under the encoding provided by T) and G's shortest encoding of D.[2] I will argue that the learners obtained using BML are not only conceptually correct starting points for learning, but also manage to avoid some of the main empirical shortcomings of learning proposals in the literature. To establish the practical applicability of BML, I would need to demonstrate its performance across

[2]In effect, the learner obtained for any theory of UG will be a radical version of Minimum Description-Length (MDL) that uses only those representations sanctioned by that theory of UG.

a wide range of theories of UG and possible inputs. At present I cannot do this, but I will provide a proof-of-concept implementation that will demonstrate the performance of BML using one naive theory of UG and several inputs.

Before I present BML and the proof-of-concept implementation in detail, I will review some of the historical facts that have led to the current lack of collaboration between theories of UG and theories of learning.

2 Background

2.1 The pessimistic view

In an influential paper, Gold (1967) introduced a learning paradigm, *identification in the limit (iitl)*, and proved that learning of this kind is impossible even in seemingly simple cases. In *iitl*, a learner g is presented with a sequence (or *text*) T of elements from a language L, where L is known to be taken from a set C of candidate languages. After each new element in T is presented, g guesses a language in C. If after a certain point all of g's guesses are the same correct guess (in this case, L), we will say that g has identified L in the limit from T. If g can identify in the limit any $L \in C$ based on any *fair* text in L (that is, a text in L in which every $w \in L$ appears at some point, and in which nothing appears that is not in L), we will say that g identifies C in the limit. If such a g exists, we will say that C is identifiable in the limit.

Certain simple families of languages are *iitl*. For example, the set of all finite languages over a finite alphabet Σ is *iitl*: if g guesses at each point the language that is the union of all the elements in T that have been encountered so far, it will always identify the source language in the limit. Similarly, any C that can be written as $\{L_i | i \geq 1\}$, where $L_i \subset L_{i+1}$ for all i, is *iitl*: g can identify C in the limit by always guessing the minimal L_i that contains all the elements in T that have been encountered so far. Changing these families of languages only slightly makes them not *iitl*. For example, adding a single infinite language to the set of all finite languages makes the set not *iitl*. In the second, more general example, adding $L_\infty = \bigcup L_i$ to C makes the result (as well as any set that contains it) not *iitl*. To see why, assume to the contrary that $C' = C \cup \{L_\infty\}$ is *iitl*. Let g be a learner that identifies C' in the limit. We can construct a text T that starts as a text in L_1 up until the first point where g guesses L_1 (such a point exists by assumption), continues as a text in L_2 up until the first following point where g guesses L_2, then continues as a text in L_3 until g guesses L_3, and so on. The result is a text in L_∞, but g makes infinitely many different guesses and so never converges on a correct answer, contrary to assumption.

Gold's setting rules out learning even in intuitively very simple families of languages, like the set of all regular languages. For theoretical linguists, this has confirmed a growing skepticism about the role of learning in linguistic competence. The skepticism was grounded in a general sense that learning is hard and that the data available to the child are insufficient. It was further supported by the discovery of a variety of linguistic universals, as well as by empirical evidence suggesting that learning in animals and humans tends to be restricted and selective. See in particular the experiments of Peña et al. (2002), Wilson (2006), Moreton (2008), and Endress and Mehler (2009),

in which humans have failed to notice various phonological patterns in the input.

Gold's results can be seen as providing formal justification for this skepticism: assuming *iitl* is an appropriate model for language learning in humans, the set of possible languages must be a severely restricted set. I will refer to this conclusion as Radical Universal Grammar (RUG). Osherson et al. (1984) formulate further assumptions about human learning that, if correct, would entail an even more restrictive version of RUG, in which the task of the learner is reduced to choosing from a finite set of candidate languages. Examples of linguistic approaches that adopt the finite version of RUG are the Principles and Parameters framework of Generative Grammar (P&P; Chomsky, 1981) and Optimality Theory (OT; Prince and Smolensky, 1993). It is worth noting that, while RUG addresses the theoretical problem of *iitl*, even its finite version does not guarantee an easy task in practice, since a finite space can still be dauntingly large. In the P&P framework, for example, there are 2^n settings, where n is the number of parameters (on the standard assumption that parameters are binary), and in OT there are $n!$ different constraint rankings, where n is the number of constraints. Noise and cognitive limitations further complicate the task. See Gibson and Wexler (1994) and Niyogi and Berwick (1996) for attempts to tackle the practical issues of acquisition within P&P and Tesar and Smolensky (1998) and Boersma and Hayes (2001) for a similar discussion within OT.

2.2 The probabilistic turn

Other work, both theoretical and experimental, supports a less restrictive view on learning. First, as has often been observed, some of Gold's assumptions do not seem to match the situation of the human language learner. In particular, the learner is expected to guess perfectly based on any fair text in the target language. No provision is made for discounting (or excluding completely) texts that are in some sense deviant, and no guess that is less than perfect counts. In acquisition, on the other hand, it is far from obvious that all sequences of inputs are equally good, and learning may well count as successful even if the child ends up having somewhat different judgments from its parents' about various sentences.[3] Relaxing this requirement, as has been done in the probabilistic settings of Horning (1969) and others, yields notions of learning that are often much more inclusive than *iitl*. Horning's setting involves the same form of text presentation as Gold's, but the texts are generated by taking independent, identically distributed samples from the strings generated by a probabilistic context-free grammar (PCFG), and the criterion for learning is modified. On these assumptions, the set of languages generated by PCFGs is learnable, even though the set of languages generated by Context-Free Grammars (CFGs) is not *iitl*.[4]

[3] A different aspect of *iitl* that could be changed with significant consequences for learnability is the assumption that the learner is only exposed to positive evidence. If the learner is exposed both to positive and to negative evidence (for example, as a sequence of strings paired with a grammaticality judgment), many more families of languages become learnable, including families that might be of potential linguistic interest. (Intuitively, the reason negative evidence helps is that it breaks all the subset relations between the languages in C – see Gold (1967) for discussion.) Unfortunately, infants do not seem to have access to anything like systematic negative evidence (Brown and Hanlon, 1970; Marcus, 1993).

[4] Other classes of languages also become learnable in Horning's setting, and the conditions on the generation of texts have been relaxed to any stationary ergodic process by Clark (2001). For further discussion

Horning's results teach us that a probabilistic approach is both more natural and more successful than *iitl*. Experimental data about specific learning tasks has provided empirical evidence for the role of statistics in learning, as well as further clarification of the requirements for a successful theory of learning in humans. One example is the segmentation experiments of Saffran et al. (1996), who showed that infants can reliably segment an artificially-generated input after a short exposure.[5] Since the only cues for segmentation in these experiments are statistical, we can conclude that a learner must be able to make use of statistical regularities in the input. In addition, these results show that a model for human learning should succeed even with unsegmented input.[6] Finally, the success of the babies in learning after such a brief exposure provides a preliminary quantitative measure of the performance of the learner.

Experimental results about learning tasks, of the kind mentioned above, have sometimes inspired task-specific learning models. For example, the results of Saffran et al., as well as those of subsequent experiments within the paradigm, have been taken to show that humans can employ certain segmentation techniques. One mechanism, based on Harris (1955) and suggested as the mechanism behind the infant segmentation data by Aslin et al. (1998), involves the tracking of transitional probabilities between syllables. Transitions tend to be more restrictive within words than across words, so segmentation can proceed by finding drops in transitional probability. Different task-specific models of segmentation have been offered by Brent and Cartwright (1996), Christiansen et al. (1998), Brent (1999), Mattys et al. (1999), Johnson and Jusczyk (2001), Venkataraman (2001), and Batchelder (2002), among others.

While sometimes helpful for the tasks for which they were created, task-specific mechanisms give rise to certain concerns. First, they tend to be unprincipled: the machinery introduced to handle the task is often unrelated to any independently justified cognitive machinery, and different proposals are hard to choose from in terms of parsimony. Second, they often do not generalize to other tasks: a segmentation-specific mechanism will not help with, say, the learning of part-of-speech tags, and a learner of part-of-speech tags will not help with segmentation. Although these models invest in learning machinery that goes beyond what would be motivated by considerations of competence, they only seem to apply to the task for which they were created. If we wish to learn more, the task-specific approach does not suggest a clear way to proceed.[7] Finally, as argued by de Marcken (1996), Clark (2001), and Goldwater et al. (2009), task-specific learners tend to fail in the face of patterns in the data that go beyond their particular task. For example, the tendency of *of* to be followed by *the* can lead a unigram learner to add *of the* to the lexicon, since that is the only way in which the model can represent such a dependency. This results in undersegmentation. Goldwater

see Osherson et al. (1986), Angluin (1988), and Clark (2001). The most general statement of the role of probabilities in induction is due to Solomonoff (1964a,b), to which I return below. A different paradigm of learning which relies on the probabilistic relaxation of the requirements of complete identification and identification from any text is the framework of PAC learning (Valiant, 1984).

[5]Other examples include the tasks of categorization, the learning of phonotactics, and the induction of grammatical rules.

[6]Removing the segmentation marks in the text makes the learning problem harder. For example, the family $C = \{\{a\}, \{aa\}\}$ is trivial to learn from a segmented text but impossible to learn from an unsegmented text. Both Gold and Horning require the input to be segmented.

[7]An exception is de Marcken (1996)'s learner, which is close in spirit to the learner proposed below.

et al. suggest that higher-order *n*-gram models, such as bigram models, can alleviate the problem of undersegmentation. These higher-order models, however, are still less than fully general. They fail to capture dependencies in syntax and semantics that go beyond the degree of the *n*-grams that are used, leading to the potential replication of the problems discussed by Goldwater et al.[8] In other words, trying to learn less than everything at once, which might seem to make the learning task easier, actually hurts the learning process.

2.3 Prediction and description length

An approach to learning that is radically different from the task-specific one is the idea of learning everything at once, with particular learning tasks (such as segmentation, categorization, syntactic learning, and so on) arising as by-products of a very general learning process. Here a principled approach is provided by the theory of prediction developed by Solomonoff (1964a,b).[9] Simplifying, we consider all the different hypotheses about the data, each treated as a computer program that outputs the data, and we evaluate each hypothesis according to its length. The learner bases its guesses about the continuation of the input based on a weighted sum of all the hypotheses compatible with the observations so far, with shorter hypotheses receiving higher weights. Recently, this approach has been proposed by Chater and Vitányi (2007), who refer to it as *ideal learning*, as a useful abstraction for evaluating certain claims about the learnability of natural language.

While fully general and mathematically sound, ideal learning is not cognitively plausible in its original form. It requires considerable computational resources, making the learning task much more costly than seems reasonable.[10] Moreover, it relies on the ability to compute infinite sums (or approximations thereof), exponents, and fractions, all of which go well beyond what are normally assumed to be basic components of human cognitive machinery. A final concern is that most linguistic theories maintain that speakers have an actual grammar, rather than a weighted average of grammars. From a linguist's perspective, if a speaker uses a grammar G to evaluate an input x, then the grammaticality of x will be determined by G's ability to generate x: if G generates x then x is grammatical, otherwise it is not. Whether some other grammar G' can generate x does not affect the linguist's notion of grammaticality.[11]

[8] An alternative direction is to incorporate additional linguistic knowledge into the learning model, as suggested by Yang (2004) and Gambell and Yang (2006), who advocate the use of a constraint on stress assignment for lexical induction. Gambell and Yang's constraint is that each word must have exactly one stress. This condition, however, seems inappropriate for function words, such as *the* and *in*, and it requires the input to be annotated for stress assignment, a requirement that seems unrealistic in view of the fact that stress lacks a systematic, cross-linguistic expression.

[9] Related notions were developed by Kolmogorov (1965) and Chaitin (1966). See Li and Vitányi (1997) for discussion. Learning of this kind is guaranteed to minimize errors in a certain sense, as shown by Solomonoff (1978).

[10] In its pure form, ideal learning is not even computable (though see Solomonoff, 2008 for thoughts on how to address this concern). Restricting the set of hypotheses can ensure computability, but the computations can still be prohibitively complex.

[11] As pointed out to me by Tim O'Donnell and by Bob Frank, evidence such as that in Griffiths and Tenenbaum (2006) suggests that weighted averages might be an appropriate model for other cognitive tasks.

The approximation to ideal learning known as Minimum Description Length (MDL; Rissanen, 1978) offers a way to overcome the difficulties of ideal learning while maintaining both the weighting of hypotheses according to their length and the idea of general learning, with particular tasks falling out as by-products.[12] In MDL the hypothesis space is restricted, and the search aims at finding a single hypothesis that minimizes the total description length. MDL has been used for grammar induction in the works of Stolcke (1994), Chen (1996), Grünwald (1996), de Marcken (1996), Osborne and Briscoe (1997), Brent (1999), Clark (2001), Goldsmith (2001), and Zuidema (2003) among others. I suggest that MDL – indeed, radical MDL, without any additional machinery or heuristics – is the only approach compatible with the idea, mentioned in the introduction, that each theory of UG induces a theory of learning that uses only what is already required to account for linguistic competence. The framework for mapping theories of UG to theories of learning that I describe immediately below makes direct use of radical MDL of this kind.

3 Architecture

If theories of UG define theories of learning, our starting point for learning given a theory of UG should be its corresponding theory of learning. A theory of UG provides a set of possible grammars. Any of these can be the grammar of a competent speaker, who stores that grammar in memory and uses it to obtain an opinion about data. At the very least, then, assuming a theory of UG T with a set $\mathbb{G}$ of possible grammars commits us to the following assumptions:

(1) a. A competent adult speaker has a grammar, $G \in \mathbb{G}$
    b. G is stored in memory
    c. G is used to parse inputs

In order to make learning possible, we must allow a learner who currently represents G to also consider some other grammar G' and to switch from G to G' under certain conditions. Of the very few properties that we can rely on to compare the two grammars in the general case, total storage space is a natural candidate, and one that accords well with the intuition behind MDL, which equates learning with compression. I therefore add the following two assumptions:

(2) a. During language learning, a second grammar, $G' \in \mathbb{G}$ can be stored in memory and used to parse the input
    b. The memory size used to store G and its parse of the input can be compared to the memory size used to store G' and its parse of the input

These assumptions amount to little more than saying that grammars can be used for parsing and that the overall description length of two grammars can be compared. My claim is that these assumptions already provide the language learner with an inherent learning mechanism: given an input D, the language learner searches through $\mathbb{G}$ for

[12] See also the closely related approach known as Minimum Message Length (MML; Wallace and Boulton, 1968).

the grammar G for which the encoding of G (as defined by T) and of D (using G) is the shortest. I will refer to this mapping of theories of UG into minimally-committed learners as Bare-Minimum Learning (BML). By relying only on what the theory of UG under consideration is already committed to, BML offers a natural starting point for the study of learnability. Moreover, as suggested in the introduction, it provides a framework in which theories of UG can be compared with respect to their predictions about learning.

BML's conceptual advantages will, of course, only be relevant if the learners provided by BML are also practical. The success of a variety of MDL-based approaches in the literature is encouraging in this respect. Unfortunately, though, these approaches tend to treat MDL as a heuristic, often combined with others. For Stolcke (1994), for example, MDL is used to evaluate the rules in a PCFG, while a Dirichlet prior is used for the probabilities. For Clark (2001), MDL is combined with a particular condition of mutual information to determine constituency. Consequently, their results do not prove that the radical MDL learner induced by a theory of UG is a viable learner.[13] To establish that BML is a practical option, I must therefore provide my own implementation and demonstrate the learning it achieves across a range of theories of UG and of inputs. While I do not yet have this kind of broad-range demonstration, I will provide the initial results of a proof-of-concept implementation using one particular theory of UG – a naive representation for the set of all CFGs – and across a small range of inputs. This is meant as a starting point: suggestive evidence that BML is viable, with more conclusive demonstrations awaiting future work.

3.1 Encoding

As just mentioned, I will demonstrate BML using a naive theory of UG. This theory, which I will refer to as T_1, allows any CFG to be represented by listing all the rules in some order, with a category #, which is not one of the terminals or nonterminals in the grammar, serving as a separator. Since T_1 only allows CFGs, it can list each rule unambiguously as the left-hand side followed by the list of the categories on the right-hand side.[14] T_1 marks the end of the grammar with an additional separator. For example, the grammar in (1) will be listed as `ABA#ABC#A#BCD#...#EFG##`.

$$G \quad := \quad \begin{cases} A \rightarrow B\,A \\ A \rightarrow B\,C \\ A \rightarrow \epsilon \\ B \rightarrow C\,D \\ \vdots \\ E \rightarrow F\,G \end{cases} \tag{1}$$

[13] A notable exception is de Marcken (1996), whose sole criterion in evaluating hypotheses is MDL (which he uses as a proxy for Structural Risk Minimization). However, de Marcken's main focus is a sophisticated representation scheme that he develops, and it remains unclear from his work whether pure MDL could work elsewhere.

[14] This particular choice of encoding individual rules would change in extensions of the learner beyond CFG, but the general point will not be affected. As long as the grammar can be stored and used for parsing, it can be encoded, and the encoding can be used in a BML learner.

We still need to specify how T_1 encodes the categories in the list. Sticking to simple-minded (and deliberately suboptimal) choices, we will use a fixed code-length scheme for the different categories, where each category will be encoded using $k = \lceil \lg(|Categories| + 1) \rceil$ bits:

#	000
A	001
⋮	⋮
G	111

The number of bits per category, k, will have to be represented as well. We can do this by starting the code with a sequence of k 0's followed by a single 1, and by agreeing to treat $\underbrace{000}_{k}$ as #. Encoding the grammar in (1) then, will be $\underbrace{000}_{k}\,1\,\underbrace{001}_{k}\,\underbrace{010}_{k}\,\underbrace{001}_{k}\,\underbrace{000}_{k}\ldots\underbrace{000}_{k}$, and the total length of encoding G will be $|G| \approx k \cdot [\sum_{r \in G} |r| + 1]$.

As for determining the encoding of the data, D, given G, T_1 first groups rules by their left-hand side, and then enumerates the expansions:

Rule	Code
$A \rightarrow BA$	00
$A \rightarrow BC$	01
$A \rightarrow \epsilon$	10
$B \rightarrow CD$	0
$B \rightarrow b$	1
$C \rightarrow c$	ϵ
⋮	⋮

Suppose now that G provides the following parse for D: $T = [_A[_B \ldots]\,[_C \ldots]]$. T_1 encodes this parse by traversing the tree in pre-order, concatenating the code for each expansion choice given the left-hand side: $C(T) = C(A)C(A \rightarrow BC\,|A)C(\ldots|B)\ldots C(\ldots|C)\ldots$. In cases of ambiguity, T_1 takes the shortest encoding.

3.2 Search

Using the UG specified above as T_1, we can now take some input D and search for the grammar that minimizes the total description length of G and of the encoding of D given G. Any grammar G_0 that parses the input can serve as an initial hypothesis for the search. Moreover, G_0 provides a trivial upper bound on the size of the search, since the total description length provided by the target grammar is at most as large as that provided by G_0.

For the T_1, there is a very simple grammar that is guaranteed to parse D and can serve as G_0. This grammar is what I will refer to as the *concatenation grammar for* Σ, where Σ is the alphabet in which D is written. If $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$, the concatenation grammar for Σ is defined as follows:

$$G \;:=\; \begin{cases} \gamma \rightarrow\ \sigma_1\,\gamma \\ \vdots \\ \gamma \rightarrow\ \sigma_n\,\gamma \end{cases} \tag{2}$$

The concatenation grammar for Σ makes all texts of a certain length written in Σ equally easy to describe. It treats all symbols in all positions in D as equally good and therefore fails to capture any regularity other than the alphabet in which D is written. Consequently, it is only a good hypothesis for a random or near-random text. However, since it parses D it can serve as an initial hypothesis, and it provides an initial upper bound on the total description length using the target grammar.

Still, the bound provided by the concatenation grammar is huge, ruling out an exhaustive search. A greedy search is not likely to succeed, due to various local optima along the way. To address this problem, I chose to implement the search using Simulated Annealing (Kirkpatrick et al., 1983), though I wish to emphasize that I am not trying to model the search procedure in humans, and my only claims concern the definition of the objective function, stated in terms of total description length.

4 Results

4.1 Segmentation I

I presented the BML learner instantiated with the theory of UG specified as T_1 in section 3.1 above with an input similar to the one described by Saffran et al. (1996). In Saffran et al.'s experiment, a text was generated by the random concatenation of elements from a vocabulary consisting of the items `pabiku, golatu, daropi, tibudo`. This text was turned into speech using a synthesizer that produced a stream of speech with flat intonation and no word breaks. Eight-month old infants were exposed to this stream, and after two minutes (= 180 words = 1080 segments) they were able to distinguish between words (e.g. `pabiku`) and non-word sequences that appear in the text (e.g. `bikuda`).[15] Here are sample snapshots from the learning process using an input that is only 300 segments long (compared to 1080 in the original experiment), using an initial temperature of 15 and a maximum grammar-length of 200 bits. The first step, as explained above, is a concatenation grammar, which captures no regularities:[16]

G_0 :

$$\begin{array}{ll} \gamma \rightarrow k\,\gamma & \gamma \rightarrow i\,\gamma \\ \gamma \rightarrow o\,\gamma & \gamma \rightarrow u\,\gamma \\ \gamma \rightarrow d\,\gamma & \gamma \rightarrow p\,\gamma \\ \gamma \rightarrow a\,\gamma & \gamma \rightarrow g\,\gamma \\ \gamma \rightarrow r\,\gamma & \gamma \rightarrow b\,\gamma \\ \gamma \rightarrow l\,\gamma & \gamma \rightarrow t\,\gamma \end{array}$$

Grammar length: 126 Encoding length: 1200 Energy: 1326.0

[15] The text used by Saffran et al. (1996) was subject to the additional requirement that no word can repeat itself. In the text that I used for BML, repetitions are not prohibited. As far as I can tell, this does not affect the point made here.

[16] In the results reported here, the step in the search appears as the subscript of G; γ is the seed category; and numbered categories are non-terminal categories that are hypothesized by the learner during the search.

After a thousand steps, we already have `ro` from `daropi`, `la` and `go` from `golatu`, and `ku` from `pabiku`:

G_{1000} :

$d \rightarrow o$	$\gamma \rightarrow d\ \gamma$
$\gamma \rightarrow \gamma$	$\gamma \rightarrow u\ \gamma\ d$
$a \rightarrow$	$\gamma \rightarrow o\ \gamma\ g$
$\gamma \rightarrow t\ \gamma$	$\gamma \rightarrow l\ a\ \gamma\ i$
$\gamma \rightarrow r\ o\ \gamma$	$\gamma \rightarrow g\ o\ \gamma\ p$
$\gamma \rightarrow i\ \gamma\ t$	$\gamma \rightarrow p\ \gamma\ d$
$l \rightarrow u\ i$	$\gamma \rightarrow k\ u\ \gamma\ b$
$\gamma \rightarrow a\ \gamma$	$r \rightarrow$
$\gamma \rightarrow b\ \gamma$	

Grammar length: 192 Encoding length: 1023 Energy: 1215.0

As we proceed, more and more parts of the underlying vocabulary are discovered. Here, at the final step, we have all the words:

G_{100000} :

$5144 \rightarrow t\ i\ b\ u\ d\ o\ 5144$	$5144 \rightarrow p\ a\ b\ i\ k\ u\ 5144$
$5144 \rightarrow g\ o\ l\ a\ t\ u\ 5144\ r$	$5144 \rightarrow d\ a\ r\ o\ p\ i\ 5144$

Grammar length: 97 Encoding length: 100 Energy: 197.0

The results presented above show rules that correspond straightforwardly to the lexicon that was used to generate the input and thus reflect the correct segmentation of the input, based on its statistical regularities. Crucially, though, the theory of UG presented as T_1 in section 3.1 is not aware of the tasks of segmentation and lexicon induction, and it does not represent probabilities in its rules. Consequently, the instantiation of BML with T_1 is not aware of these notions either. It arrives at the correct segmentation as a by-product of its general search for the best grammar given the input.

4.2 Segmentation II

Below are some sample snapshots from a similar text in which the underlying words have varying lengths: `pabiku, tibudo, gola, tudaropi`. As was shown by Johnson and Jusczyk (2003), this makes the learning task harder for humans. The BML learner for the UG in section 3.1 was able to learn the grammar correctly (using the same parameters as before):

G_0 :

$$\begin{array}{ll}
\gamma \to i\,\gamma & \gamma \to o\,\gamma \\
\gamma \to a\,\gamma & \gamma \to b\,\gamma \\
\gamma \to t\,\gamma & \gamma \to r\,\gamma \\
\gamma \to u\,\gamma & \gamma \to g\,\gamma \\
\gamma \to d\,\gamma & \gamma \to p\,\gamma \\
\gamma \to l\,\gamma & \gamma \to k\,\gamma
\end{array}$$

Grammar length: 126 Encoding length: 1200 Energy: 1326.0

G_{6000} :

$$\begin{array}{ll}
418 \to t\,i\,b\,u\,418\,g & 418 \to k\,u\,418\,d\,d \\
418 \to t\,u\,d\,418\,418 & 418 \to 418\,g \\
418 \to p\,a\,b\,i\,418 & l \to \\
418 \to p\,b & l \to l\,k\,t \\
418 \to a\,r\,o\,p\,i\,418\,t & 418 \to p\,g \\
418 \to & 418 \to g\,o\,l\,a\,418\,418\,418 \\
418 \to b & 418 \to d\,o\,418
\end{array}$$

Grammar length: 200 Encoding length: 349 Energy: 549.0

G_{100000} :

$$\begin{array}{ll}
5288 \to t\,i\,b\,u\,d\,o & 5288 \to g\,o\,l\,a \\
5288 \to p\,a\,b\,i\,k\,u\,5288 & 5288 \to t\,u\,d\,a\,r\,o\,p\,i\,5288\,5288\,5288
\end{array}$$

Grammar length: 95 Encoding length: 94 Energy: 189.0

4.3 PCFGs

As discussed above, BML is more general than Horning (1969)'s learning approach in that it assumes neither a segmented input nor that the samples are independent and identically distributed. To allow the BML learner to take advantage of the more restrictive setting of Horning's paradigm, we can specify its goal when presented with a segmented input sequence to be the minimization of the sum of the grammar length and the sum of the encoding lengths for each element in the sequence.[17] At least in simple cases, the BML learner successfully identifies the generating grammar from an input presented in this way. Following are several snapshots from a run on an input that consists of 200 even-lengthed palindromes over the alphabet $\Sigma = \{a, b, c\}$ (the sequence reported here starts as `cccabaccabaccc, cbbc, bccccccb, aa, aabbaa, ...`; for performance purposes, no more than the first 25 characters of each element are presented):

[17] Note, however, that the BML learner treats its input as the prefix of a possibly infinite text rather than a complete element in the language. I will not discuss this issue.

G_0 :

$$\gamma \rightarrow a\ \gamma \qquad\qquad \gamma \rightarrow c\ \gamma$$
$$\gamma \rightarrow b\ \gamma$$

Grammar length: 19 Encoding length: 2314 Energy: 2333.0
G_{1400} :

$$\gamma \rightarrow c\ \gamma \qquad\qquad \gamma \rightarrow a\ \gamma\ b\ \gamma$$
$$\gamma \rightarrow c \qquad\qquad \gamma \rightarrow b\ \gamma\ c\ b\ \gamma$$

Grammar length: 32 Encoding length: 2122 Energy: 2154.0
G_{2800} :

$$209 \rightarrow c\ 209 \qquad\qquad 209 \rightarrow a\ 209$$
$$209 \rightarrow b\ 209\ b\ c\ c\ 209\ c\ b\ 209\ a \qquad\qquad 209 \rightarrow$$

Grammar length: 35 Encoding length: 2154 Energy: 2189.0
G_{4200} :

$$371 \rightarrow a\ 371\ a \qquad\qquad 371 \rightarrow$$
$$371 \rightarrow b\ 371\ b \qquad\qquad 371 \rightarrow c\ 371\ c$$

Grammar length: 27 Encoding length: 1480 Energy: 1507.0

G_{4200} is already the correct grammar (371 is the arbitrary category label of what would usually be written as S). Similar results were obtained with other simple CFGs, such as $a^n b^n$.

5 Discussion

I set out to bring theories of UG and theories of learning into closer contact. I noted that any theory of UG provides a learning criterion – the total memory space used to store a grammar and its encoding of the input – that supports learning. This mapping from theories of UG to learners, which I referred to as BML, maintains a minimal ontological commitment: the learner for a particular theory of UG uses only what that theory already requires to account for linguistic competence in adults. For example, the theory of UG used to demonstrate BML in this paper sanctioned non-probabilistic CFGs under a naive encoding, and it did not represent notions like segmentation or a lexicon. Consequently, the BML learner corresponding to this UG was not aware of these notions either. Nevertheless, as a by-product of its search for the best grammar licensed by its UG, it was able to acquire a 'lexicon' from an unsegmented input using statistical patterns in the data. It was also shown to be sufficiently powerful to learn simple formal languages in the Horning (1969) paradigm. These results suggest that BML is not only the conceptually correct starting point for the study of learnability but also a practical one.

Clearly, establishing the viability of BML will require going well beyond the suggestive results presented here and investigating the performance of BML across a wide range of theories of UG and of inputs. If successful, however, BML will be a framework that provides experimental results in learning with a clear linguistic interpretation and linguistic theories with a clear learnability interpretation. At present, experimental results regarding what humans can or cannot learn do not have an immediate impact on the choices linguists make between possible theories of UG. Conversely, linguistic theories are often accompanied by a variety of different ideas about learning. BML attempts to connect the two domains much more tightly: a competence theory induces a learning ability, and deviations from that baseline are only permitted if required by the empirical data. Beyond the inherent significance in improving current models of learning and bringing them in line with theoretical linguistics, success in this project will pave the way to a method for evaluating UG theories in terms of learnability. As mentioned in the introduction, such an evaluation has been perceived as a central goal since the early days of generative grammar, but in the absence of a general way to translate theories of UG into learning models, progress toward this goal has been limited. If BML is indeed viable, different linguistic theories can be compared directly in terms of the empirical appropriateness of the learning models they induce.

References

Angluin, Dana. 1988. Identifying languages from stochastic examples. Technical Report 614, Yale University, March 1988.

Aslin, Richard N., Jenny R. Saffran, and Elissa L. Newport. 1998. Computation of conditional probability statistics by 8-month old infants. *Psychological Science* 9:321–324.

Batchelder, E. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* 83:167–206.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Brent, Michael. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Computational Linguistics* 34:71–105.

Brent, Michael, and T. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61:93–125.

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

Chaitin, Gregory J. 1966. On the length of programs for computing finite binary sequences. *Journal of the ACM* 13:547–569.

Chater, Nick, and Paul Vitányi. 2007. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51:135–163.

Chen, Stanley. 1996. Building probabilistic models for natural language. Doctoral Dissertation, Harvard University, Cambridge, MA.

Chomsky, Noam. 1955/1975. *The logical structure of linguistic theory*. Springer.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.
Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row Publishers.
Christiansen, Morten, Joseph Allen, and Mark Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes* 13:221–268.
Clark, Alexander. 2001. Unsupervised language acquisition: Theory and practice. Doctoral Dissertation, University of Sussex.
Endress, Ansgar, and Jacques Mehler. 2009. Perceptual constraints in phonotactic learning. Ms. To appear in Journal of Experimental Psychology: Human Perception and Performance.
Gambell, Timothy, and Charles Yang. 2006. Word segmentation: Quick but not dirty. Ms., Yale University.
Gibson, Edward, and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.
Gold, E. Mark. 1967. Language identification in the limit. *Information and Control* 10:447–474.
Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.
Goldwater, Sharon, Thomas Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112:21–54.
Griffiths, Thomas, and Joshua Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological Science* 17:767–773.
Grünwald, Peter. 1996. A minimum description length approach to grammar inference. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. G. S. S. Wermter and E. Riloff, Springer Lecture Notes in Artificial Intelligence, 203–216. Springer.
Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31:190–222.
Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.
Johnson, E., and Peter W. Jusczyk. 2001. Word segmentation by 8-month olds: When speech cues count more than statistics. *Journal of Memory and Language* 44:548–567.
Johnson, Elizabeth K., and Peter W. Jusczyk. 2003. Exploring statistical learning by 8-month olds: The role of complexity and variation. In *Jusczyk Lab final report*, ed. D. Houston, A. Seidl, G. Hollich, E. K. Johnson, and A. Jusczyk.
Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671–680.
Kolmogorov, Andrei Nikolaevic. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission (Problemy Peredachi Informatsii)* 1:1–7. Republished as Kolmogorov (1968).
Kolmogorov, Andrei Nikolaevic. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2:157–168.
Li, Ming, and Paul Vitányi. 1997. *An introduction to kolmogorov complexity and its applications*. Berlin: Springer Verlag, 2nd edition.
de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation,

MIT, Cambridge, Mass.
Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46:53–85.
Mattys, S., Peter W. Jusczyk, P. Luce, and J. Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* 38:465–494.
Moreton, Elliott. 2008. Analytic bias as a factor in phonological typology. In *Proceedings of WCCFL 26*, ed. Charles B. Chang and Hannah J. Haynie, 393–401. Somerville, MA: Cascadilla Proceedings Project.
Niyogi, Partha, and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61:161–193.
Osborne, Miles, and Ted Briscoe. 1997. Learning stochastic categorial grammars. In *Proceedings of CoNLL*, 80–87.
Osherson, Daniel N., Michael Stob, and Scott Weinstein. 1984. Learning theory and natural language. *Cognition* 17:1–28.
Osherson, Daniel N., Michael Stob, and Scott Weinstein. 1986. *Systems that learn*. Cambridge, Massachusetts: MIT Press.
Peña, Marcela, Luca Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science* 298:604–607.
Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.
Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.
Saffran, Jenny R., Elissa L. Newport, and Richard N. Aslin. 1996. Statistical learning by 8-month old infants. *Science* 274:1926–1928.
Solomonoff, Ray J. 1964a. A formal theory of inductive inference, part I. *Information and Control* 7:1–22.
Solomonoff, Ray J. 1964b. A formal theory of inductive inference, part II. *Information and Control* 7:224–254.
Solomonoff, Ray J. 1978. Complexity-based induction systems: Comparisons and convergence theorems. In *IEEE Transactions on Information Theory*, 4, 422–432.
Solomonoff, Ray J. 2008. Algorithmic probability: Theory and applications. In *Information theory and statistical learning*, ed. Frank Emmert-Streib and Matthias Dehmer, 1–23. Springer.
Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.
Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.
Valiant, Leslie G. 1984. A theory of the learnable. *Communications of the ACM* 27:1134–1142.
Venkataraman, A. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27:351–372.
Wallace, C.S., and D.M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.
Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in Cognitive*

Sciences 8:451–456.

Zuidema, Willem. 2003. How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)*, ed. Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, 51–58.