A cognitively plausible model for grammar induction*

Roni Katzir
rkatzir@post.tau.ac.il

September 17, 2014

Abstract

This paper aims to bring theoretical linguistics and cognition-general theories of learning into closer contact. I argue that linguists' notions of rich UGs are well-founded, but that cognition-general learning approaches are viable as well and that the two can and should co-exist and support each other. Specifically, 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 according to the principle of Minimum Description-Length. This mapping from UGs to learners maintains a minimal ontological commitment: the learner for a particular UG uses only what is already required to account for linguistic competence in adults. I suggest that such learners should be our null hypothesis regarding the child's learning mechanism, and that furthermore, the mapping from theories of UG to learners provides a framework for comparing theories of UG.

1 Introduction

A central task in theoretical linguistics (TL) 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 TL is characterizing the range of possible grammars that adult speakers can have. Thus, 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 characterization of the range of allowable grammars – which can be thought of as a reference machine into which individual grammars are written – is

*I thank Adam Albright, Asaf Bachrach, Bob Berwick, Michael Collins, Danny Fox, Tova Friedman, Arnon Lotem, Alec Marantz, Andrew Nevins, Tim O'Donnell, Amy Perfors, Ezer Rasin, Raj Singh, Donca Steriade, Josh Tenenbaum, and the audiences at MIT, Tel Aviv University, Ben Gurion University, Hebrew University, and the Cornell Workshop on Grammar Induction, as well as the reviewers and editors of JLM. This work has been supported by ISF grant 187/11.

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. Moreover, UG can provide an evaluation metric that allows the child to compare potential grammars given the data. In its original formulation, this evaluation metric was stated in terms of simplicity, a notion that – though defined with respect to a concrete UG – is also often seen as a cognition-general (CG) principle.

One might hope, then, that TL theories of competence and CG theories of learning would have a close relationship: that theories of UG would map onto theories of learning through an evaluation metric, and that theories of learning would restrict the choice of UG. In practice, however, the evaluation metric has been largely abandoned, and the two domains have never succeeded in constraining one another. Worse, TL and CG approaches have grown to be considered mutually incompatible. There are various different aspects to this ostensible incompatibility, such as whether linguistic knowledge involves structured, rule-like representations or not, whether probabilities play a role, and so on. Perhaps most fundamental among the perceived differences is how the two approaches view learning. TL, following a more hopeful beginning, has adopted a deeply skeptical stance that rejects the possibility of any meaningful learning and relegates most of the linguistic ability of adults to the innate component, and often to UG itself (that is, to the reference machine). CG, on the other hand, tends to be confident of learning and skeptical of the innate component (and especially of UG). The perceived incompatibility between TL and CG has led over the years to a growing divide between the two disciplines.

Over the past decade or so, the Bayesian program for cognition and the closely related framework of Minimum Description Length (MDL) have brought the two disciplines closer by articulating CG views that can integrate probabilistic reasoning with structured, symbolic representations. In the other direction, proposals such as Marcus (2000) and Yang (2004, 2010) offer TL perspectives that connect with CG approaches to learning. But a sizable gap remains: even CG-oriented TL proposals such as those of Marcus and Yang still question the ability of general learning mechanisms to generalize correctly from the data, embracing instead restrictive theories of the innate component; and even TL-oriented CG proposals such as Goldsmith (2001), Dowman (2007), Foraker et al. (2009) and Perfors et al. (2011) still emphasize the power of general-purpose learning mechanisms and question whether the innate component should be quite as rich as TL would have it.

This paper has two goals. First, I wish to explain why the skepticism in both directions is misguided. In particular, I will explain why linguists believe in a complex innate component – including a nontrivial UG – even in the face of powerful statistical

[1] Elsewhere in the literature, UG is sometimes used to refer to the range of possible grammars (rather than to its intensional characterization as a reference machine), and sometimes it is used to refer to the combination of the range of possible grammars and the learning mechanism. Here UG will refer strictly to the reference machine. 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.

learners. I will do this by presenting two kinds of evidence that linguists rely on that have nothing to do with questions of learnability in principle. I will also explain why many cognitive scientists are confident that learning is a real possibility, despite the arguments against learning in the TL literature. My second goal is to offer a TL view that treats the learnable and the innate as mutually supportive rather than conflicting. The ability of CG mechanisms to learn, on this view, is interpreted not as a reason to reduce the innate component – though it will be a reason to bring back into consideration theories that leave much to be learned – but rather as a way to extract nuanced predictions from competing theories of that component.

I start, in section 2, by reviewing the history of the divide between TL and CG, focusing first on the roots of TL pessimism regarding learning (section 2.1) and then on CG optimism regarding the same (section 2.2). In section 3 I evaluate the two positions and argue that TL was wrong to dismiss learning but right to emphasize potentially restrictive UGs, while CG was right to emphasize learning but wrong to dismiss potentially restrictive UGs (section 3.1). In section 3.2 I explain how the co-existence of rich UGs and meaningful learning is not only possible but in fact a good state of affairs, one that allows us to revive the old hope of mutual collaboration from the early days of generative grammar. In section 3.3 I explain how any fully explicit theory of UG provides us with a CG learner – specifically, a Minimum Description-Length (MDL) learner – and that this provides both a starting point for the study of learning and a basis for comparing competing theories of UG. Section 4 illustrates this mapping from UG to MDL learner using a simple UG and a couple of toy examples. Section 5 concludes.

2 TL and CG: A brief history of the schism

2.1 TL: Skepticism about learning

2.1.1 Identification in the limit

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.[2] For theoretical linguists, this has confirmed a growing skepticism (already discussed explicitly in Chomsky, 1965, pp. 56–58) 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. 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 severely restricted. Osherson et al. (1984) formulate further assumptions about human learning that, if correct, would entail an even more restrictive version UG 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 UG 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 a restricted enough UG addresses the theoretical problem of *iitl*, even the 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 Clark and Roberts (1993), Gibson and Wexler (1994), Niyogi and Berwick (1996), and Yang (2002) for attempts to tackle the practical issues of acquisition within P&P and Tesar and Smolensky (1998), Boersma and Hayes (2001), and Magri (2013) for a similar discussion within OT.

2.1.2 Poverty of the stimulus

Much of the disagreement between TL and CG has centered on a form of argument known as the argument from the poverty of the stimulus (POS), involving some property P that humans demonstrate in their language in spite of apparently insufficient support for P in the data. To cite a well-known (and highly controversial) example, English-speaking children will form a yes/no question by fronting the structurally high-

[2] A full characterization of when a family of languages is *iitl* is provided by Angluin (1980). Algorithms that guarantee *iitl* for various classes of languages include Angluin (1982), Koshiba et al. (1997), Clark and Eyraud (2007), Heinz (2010), and Yoshinaka (2011). Note that arguments such as Gold's show that, under the relevant assumptions, *no* learner can succeed. This is a stronger result than showing that a particular learner cannot succeed (such as the problem identified by Braine, 1971, Baker, 1979, and Dell, 1981 for the specific evaluation metric of Chomsky and Halle, 1968).

est auxiliary rather than the leftmost one, thus forming the yes/no interrogative version of *The monkey that is jumping can sing* by asking *Can the monkey that is jumping sing?* rather than **Is the monkey that jumping can sing?* (where $*$ marks ungrammaticality). They do so, it appears, despite hearing only simpler yes/no questions such as *Is the monkey jumping?* (from *The monkey is jumping*) and *Can the monkey sing?* (from *The monkey can sing*), where structurally highest and leftmost amount to the same thing. This has been taken to show that the innate component ensures this choice by making available structure-dependent generalizations but not rules that depend on linear order. See Berwick et al. (2011) and Clark and Lappin (2011), as well as references therein, for discussion.

While the form of POS arguments is clear enough, it is often difficult to establish any particular POS argument for humans in practice, even in a simple case such as the one just mentioned.[3] For example, how can we determine just what kind of evidence would suffice to make the relevant choice empirically? Could there be indirect sources of information that would predispose the child against forming ordering-based generalizations? And how sure are we that we know exactly what data the subjects have encountered over those few years prior to the experiment? Some progress has been made on these questions (see Legate and Yang (2002), Lidz et al. (2003), Yang (2010), and Hsu and Chater (2010) for thoughts on quantifying the information available to the the child; see Crain and Pietroski (2002) for how POS can be constructed from developmental stages in which children exhibit very specific linguistic knowledge that is incompatible with their ambient language but compatible with other natural languages; and see Wilson (2006) for an experimental paradigm designed to test the child's generalization beyond the data in POS situations), but the core weakness of relying on what we think can be learned and what we think the child hears – two questions that can be prohibitively difficult to answer – remains.[4]

2.1.3 Richness of the stimulus

If children can be shown to systematically *not* demonstrate a property P in their language despite an adequate amount of evidence supporting P in the input, we can conclude that this failure is due to the innate component. We can term such evidence an argument from the richness of the stimulus (ROS).[5] For example, Peña et al. (2002) have shown that, while humans are capable of extracting abstract dependencies within words, they fail on this task when combined with a segmentation task (a task that subjects perform well on, both on its own and when combined with the task of extracting word-internal dependencies). Similarly, Moreton (2008) has shown that humans are significantly better at learning certain phonological dependencies – specifically, dependencies relating the height of the vowels in two adjacent syllables – than other phonological dependencies – dependencies relating the height of a vowel to the voicedness of

[3] In organisms for which it is possible to conduct controlled POS experiments, the situation is different, as Dyer and Dickinson (1994)'s work on honeybees shows.

[4] This is not to say that the POS argument above has been shown to be incorrect. Despite multiple attempts to do so in the CG literature, the POS argument using subject-auxiliary inversion remains an open question. See Berwick et al. (2011) for relevant discussion.

[5] See Smith (1966) for an early example of this kind of argument in humans, and see Garcia et al. (1974) for a particularly clear example of the argument in rats.

the following consonant and dependencies relating the voicedness of consonants in two adjacent syllables – even though the two patterns are equally prominent perceptually and are both abundantly represented in the input.

One must ensure, of course, that prior exposure has not biased the subjects against observing the relevant patterns. This, however, is considerably easier in practice than the reverse task, essential to POS, of ensuring that a certain pattern is never attested in the data. And as the above examples show – see Bonatti et al. (2005), Endress et al. (2007), Endress and Mehler (2009a), Becker et al. (2011), and Hunter and Lidz (2012), among others, for further evidence of this kind – ROS lends itself to the design of controlled experiments that can inform us about what humans fail to learn.

2.1.4 Typology

Perhaps the most common source for enrichments of the innate component come from the routine TL task of examining individual languages and comparing the results across a range of languages. If language after language shows the same property P (which can be an absolute universal, such as "Has nouns" or an implicational universal, such as "If demonstratives and adjectives precede the noun, then demonstratives precede adjectives"), we can sometimes conclude that P is due to the innate component.

As usual, caution is needed: for some properties, other sources, such as communication pressure, might be responsible rather than the innate component. For example, $P =$ "Verbs have a small number of arguments" or $P =$ "Has vowels". More interestingly, P may arise not through any direct benefit to the speakers but as properties that enhance the transmission of language between generations of speakers. See Kirby (2000, 2002); Kirby et al. (2004); Smith et al. (2003) as well as Niyogi and Berwick (1997, 2009). Less frequently, P can be explained away by appealing to historical accident.[6]

But in many cases, P has little if anything to recommend it in terms of communication efficiency and other functionalist criteria. Suppose, to take a syntactic example discovered by Ross (1967), that I heard you say that Max and some lady left the party together last night, but I don't know the identity of the lady in question. I could use a roundabout inquiry such as *I heard that Max and some lady left the party together; can you tell me which lady?*, or I could use a paraphrase such as *Which lady did Max leave the party with _?*, where the conjunction *and* in the original sentence is replaced with the preposition *with*. But what I cannot do, in English or in any other known language, is use the standard way to form a question and say **Which lady did Max and _ leave the party last night?*, despite its obvious usefulness for the conversation (P in this case could be "Does not allow a question to target a single conjunct"). To cite a different example, discovered by Horn (1972), no natural language has a connective corresponding to *NAND* (= not and) or a quantificational determiner corresponding to

[6]Controlling completely for historical accident is quite challenging in practice, but the emergence of the Nicaraguan Sign Language (Senghas et al., 2004) and of the Al-Sayyid Bedouin Sign Language (Sandler et al., 2005) provide an approximation. In non-human species it is sometimes possible to explore typological questions in lab settings that control in full for historical accident, as shown by the work of Feher et al. (2009) on the emergence of typical song patterns in zebra finches over several generations, starting from birds grown in isolation.

NALL (= not all), despite the usefulness of these concepts in everyday life (as well as in artificial settings).[7] In such cases, it seems reasonable to ensure P through the innate component.[8,9]

2.2 CG: Optimism about learning

2.2.1 The probabilistic turn

Other work, both theoretical and experimental, supports a less restrictive view on learning than the TL view. 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 in *iitl* 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.[10] 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*.

Horning's results – and those of later probabilistic developments such as Wexler and Culicover (1980), Osherson et al. (1986), Angluin (1988), Kapur (1991), and Chater and Vitányi (2007) – can be seen as evidence that a probabilistic approach is both more natural and more successful than *iitl*.[11] Experimental data about specific

[7] See Horn (2011) and Katzir and Singh (2013) for discussion of the general context of this typological fact.

[8] Evans and Levinson (2009) and Levinson and Evans (2010) have made the remarkable claim that language universals do not exist. They do not discuss the Ross (1967)'s and Horn (1972)'s cases discussed above. See the commentaries following Evans and Levinson (2009), as well as Abels and Neeleman (2010), Crain et al. (2010), Reuland and Everaert (2010), Harbour (2011), and Matthewson (2012), among others, for additional problems with Evans and Levinson's claim.

[9] The discussion in this subsection is framed as one about absolute properties. See Tily and Jaeger (2011) and Piantadosi and Gibson (2013) for discussion of the challenges of obtaining a large enough sample to establish such universals statistically. In addition to absolute universals, quantitative typological evidence offers a rich source of information for TL, though using this information is still difficult at present. See Sauerland and Bobaljik (2013) for an interesting example.

[10] 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).

[11] Care must be taken, however, in interpreting positive results about such models from the perspective of language acquisition. Horning (1969)'s original result applies to (unambiguous) PCFGs, a class of grammars that is not a realistic model of natural languages. Osherson et al. (1986) prove that a much broader class of

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.[12] 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.[13] Finally, the success of the babies in learning after such a brief exposure provides a preliminary quantitative measure of the performance of the learner. Further evidence that humans are skillful statistical learners come from Sobel et al. (2004) and Griffiths and Tenenbaum (2006), among others, who demonstrate the sensitivity of humans (both children and adults) to statistical information.

2.2.2 Task-specific approaches

Experimental results about learning tasks, of the kind mentioned above, have sometimes inspired task-specific (but domain-general) learning models: relatively simple mechanisms, usually sensitive to statistics, that form part of a CG toolkit. 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. Other task-specific (but potentially domain-general) learning mechanisms that have been proposed in the literature include mechanisms for processing identity relations (Endress et al., 2007) and positional relations (Endress and Mehler, 2009b).[14]

languages can be identified with probability one from a similar form of text presentation (that is, through independent identically distributed draws from the language; see Clark, 2001 for further extension). However, this result requires knowing the possible distributions. If this assumption is replaced by more realistic requirements, the classes of languages that can be identified become considerably more limited, as shown by Angluin (1988) and Pitt (1989). In fact, if the child is required to perform distribution-free learning with probability one, the classes of languages that are identifiable revert to those that are Gold-identifiable. See Niyogi (2006) and Clark and Lappin (2011) for further discussion.

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

[13] 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.

[14] See Endress et al. (2009) and Endress and Bonatti (2013) for further discussion of such mechanisms and qualifications of their generality.

2.2.3 Prediction and description length

Another CG approach, one that is radically different from the task-specific approach – and the one I will try to support in this paper – 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 (1964b,a).[15] 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) and Hsu et al. (2011) as a useful abstraction – a form of *ideal learning* – for evaluating certain claims about the learnability of natural language.

While fully general and mathematically sound, ideal learning as originally formalized is not cognitively plausible, nor is it meant to be. In its pure form, ideal learning is not even computable (though see Solomonoff, 2008 for thoughts on how to address this concern). Another challenge to making Chater and Vitányi's model cognitively plausible is that it is stated with respect to a very broad UG – in its original form, a Turing-complete UG (which is the source of the non-computability). If we wish to take into account arguments for a more restrictive innate component, such as the arguments from ROS and from the typology, we should re-state Chater and Vitányi's model in terms of more limited UGs. Restricting the set of hypotheses can both ensure computability and make the model work with linguistically realistic UGs, but the computations required to derive the predictions in a Solomonoff-based ideal learner such as Chater and Vitányi's can still be prohibitively complex.

The approximation to Kolmogorov Complexity 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.[16] In MDL – and in the closely related Bayesian framework – the hypothesis space is restricted, and the search aims at finding a single hypothesis that minimizes the total description length (or, in the Bayesian framework, a hypothesis that maximizes the posterior probability). MDL has been used for grammar induction in the works of Berwick (1982), Rissanen and Ristad (1994), Stolcke (1994), Brent and Cartwright (1996), Chen (1996), Grünwald (1996), de Marcken (1996), Osborne and Briscoe (1997), Brent (1999), Clark (2001), Goldsmith (2001), Onnis et al. (2002), Zuidema (2003), Dow-

[15] 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) and Chater and Vitányi (2007).

[16] See also the closely related approach known as Minimum Message Length (MML; Wallace and Boulton, 1968). An approach related to MDL and MML is the search for a grammar (usually a context-free grammar) that generates the input data as its only possible output. The problem of finding such a grammar – the so-called *shortest grammar problem* – has its roots in Lempel and Ziv (1976) and has been studied by Nevill-Manning and Witten (1997), Kieffer and Yang (2000), Charikar et al. (2005), and Debowski (2011), among others.

man (2007), Chang (2008), and Rasin and Katzir (2013) among others. In section 3.3 I will suggest that MDL arises as a natural criterion for the evaluation of grammars given the data – and thus as a natural CG learning mechanism – from the commitment to an explicit UG made in TL.

3 Reassessment

3.1 A rich UG and the possibility of learning both exist

As we saw, TL has good reasons to assume a nontrivial UG: while *iitl* seems inapplicable to the condition of the child, and while POS arguments are susceptible to successful learning models, ROS and typological arguments do not depend on learnability in principle. Indeed, the better the general-purpose mechanisms that one can assume, the more surprising both failures to learn and systematic typological patterns become. At the same time, the CG models of learning are clearly very much an option. None of the arguments against learning in principle holds, and it seems that humans are quite good at learning statistical distributions (as shown by Sobel et al. (2004), and Griffiths and Tenenbaum (2006), among others).

Assuming that (almost) everything is innate or that (almost) everything is learned was perhaps convenient at one point as a working hypothesis: if we already have an elaborate innate component, we might hope that we could do without a sophisticated learning mechanism, and vice versa. But a rich innate component and a powerful CG mechanism are not logically incompatible, and it is worth noting that the state of the art in each project still leaves a significant amount of work for the other. At the very least, then, the two respective research projects should continue to co-exist: TL should keep studying the innate component focusing on ROS and typological evidence, perhaps showing more caution with POS arguments than it did before; and CG should keep studying what humans can learn and how, perhaps showing a better appreciation for the role of innateness in shaping adult linguistic abilities.

But there is also a more interesting option, one that allows a tighter collaboration between the two research projects and that enables discoveries in one to translate into tools for the other. This option, a hope from the early days of generative grammar, was made possible by the advent of the Bayesian program for cognition and of the closely related MDL framework, both of which allow the integration of structured representations and probabilistic reasoning. I will sketch an outline of this option immediately below.

3.2 Combining innateness with general learning

Practitioners of TL often find themselves with two different hypotheses, call them F_1 and F_2, that seem equally capable of explaining the observed linguistic phenomena. F_1 and F_2 might come from entirely different theoretical frameworks, such as Combinatory Categorial Grammar and Minimalism for syntax or Optimality Theory and SPE for phonology, or they may constitute two refinements of the same broad framework. This has led to what Steedman and Baldridge (2011) have called a crisis in syntactic

theory (though a similar problem arises in other subfields of TL, such as phonology and semantics): modern TL proposals are often meaningfully different in their essentials and yet comparably successful in accounting for the linguistic judgments of adult speakers. In order to choose between them, we need to look elsewhere.

One important source of evidence of this kind is the mapping from theories of competence to theories of processing, mediated by the competence hypothesis articulated by Miller and Chomsky (1963) and Chomsky (1965). This mapping has been used to argue for Lexical-Functional Grammar (over transformational grammars) by Bresnan and Kaplan (1982); for the flexible constituents endorsed by categorial grammars (over the rigid constituency of most other formalisms) by Steedman (1989); and for quantifier-raising (over *in situ* incorporation of quantifiers) by Hackl et al. (2012). I would like to suggest that combining CG with TL might provide another source of evidence of this kind, with a suitable mapping of UGs to CG learners (in section 3.3 below I will argue that such a mapping is available by default through the principle of MDL). The shape of possible experiments to distinguish between F_1 and F_2 is as follows. Suppose one finds two properties, P_1 and P_2, that some languages have but some do not – so that learning will be involved – and that can co-exist in the same language. To take a phonological example, P_1 might be that a voiceless consonant like $/p/$ is aspirated in the beginning of a syllable while a voiced consonant like $/b/$ is not (as in English: *[p^h]at* vs. *[b]at*; note that this is a choice of English: Hindi can aspirate both $/p/$ and $/b/$, while French aspirates neither), and P_2 might be that vowels are lengthened before a voiced consonant but not before a voiceless consonant (again as in English: *t[a:]b* vs. *t[a]p*; again, this is a choice of English: French, for example, shows no such lengthening). In a syntactic example, P_1 might be that a subject can be dropped (as in Italian, but not in English) and P_2 might be that questions are marked by overt dislocation (again, as in Italian, but not in Japanese).

Given a CG mechanism M that seems cognitively plausible, we can now obtain two combinations, $M + F_1$ and $M + F_2$, and each combination can be run on a realistic corpus of child-directed speech. While F_1 and F_2 might both be capable of representing both P_1 and P_2, there might be a significant difference in how well the combinations $M + F_1$ and $M + F_2$ can learn the two and the order in which they do so. If this is the case, we now have a criterion for choosing between F_1 and F_2: whichever provides a better match with data from actual child language acquisition will receive support. Since M was proposed as a general-purpose learning mechanism and was not tailor made to handle either F_1 or F_2, such evidence can be taken seriously.

Experiments of this kind require researchers in each project to pay closer attention to work done in the other than has usually been the case. Still, I think that they are a more productive – and, given current understanding, a more sensible – direction for future work on language and learning than further attempts to determine whether language is more innate than learned or vice versa.

3.3 An argument for MDL

I have tried to show why TL and CG can and should have a much closer relationship than they currently enjoy. In this section I will provide an argument that any explicit theory of UG already comes with the evaluation metric (or objective function) that

forms the central component of a CG learner. Specifically, I will show how any explicit theory of UG translates into an MDL evaluation metric that allows the child to compare different possible hypotheses within the hypothesis space defined by UG. If correct, the discussion below points to bare MDL as our starting point in studying learning and as the linguist's M for comparing contenders for the correct theory of UG.

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 competent adult speaker has a grammar, $G \in \mathbb{G}$
2. G is stored in memory
3. G is used to parse inputs

In order to make learning possible, we must allow a learner who currently represents G to also consider at least one other grammar G' and to switch from G to G' under certain conditions.[17] 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:

4. During language learning, a second grammar, $G' \in \mathbb{G}$ can be stored in memory and used to parse the input
5. 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 the grammar G for which the encoding of G (as defined by T) and of D (using G) is the shortest. By relying only on what the theory of UG under consideration is already committed to, this bare MDL learner offers a natural starting point for the study of learnability: alternatives in which the learner ignores the freely available MDL criterion and relies on some other mechanism instead should only be pursued given evidence

[17] Strictly speaking, maintaining more than one grammar is not always necessary. In particular, the learners proposed by Angluin (1982), Koshiba et al. (1997), Clark and Eyraud (2007), and Heinz (2010) all operate by considering just one grammar at a time and updating it as input comes along. All these learners, however, assume elaborate mechanisms for growing a grammar – usually tailor-made for the specific UGs they are designed to handle – that go well beyond the basic commitment to an explicit UG.

that the MDL null hypothesis is incorrect.[18,19] The argument for bare MDL as the null hypothesis can be taken to support approaches in the literature that use MDL for learning, such as the works mentioned in section 2.2.3, and in particular works such as de Marcken (1996) and Rasin and Katzir (2013) that use MDL not simply as a convenient heuristic but as the sole principle that maps an explicit UG to an evaluation metric.[20] Moreover, as mentioned in the introduction and discussed further in section 3.2, the generality of the mapping from UGs to learners provides a framework in which theories of UG can be compared with respect to their predictions about learning.

4 A simple example

4.1 Encoding

To see how the mapping from theories of UG to bare MDL learners works, let us consider a naive theory of UG. This theory, call it 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

[18]To date, the literature has provided very little that bears directly on the empirical question of whether children use MDL as a criterion for comparing hypotheses during learning. On the other hand, several works have provided arguments – often in conflicting directions – regarding a possible role for description length more broadly in the learning process. In particular, Feldman (2000), extending the results of Shepard et al. (1961), provides evidence for the cognitive relevance of MDL by showing that description length is correlated with learning difficulty in concept learning (see also Feldman 2006 and Goodman et al. 2008). In the same vein, Moreton and Pater (2012a,b) review the literature on artificial grammar learning in phonology and conclude that description length is a central factor determining learning difficulty in this domain. On the other hand, Kurtz et al. (2013) point to a more nuanced pattern of difficulty in concept learning, and Moreton et al. (2014) provide evidence for correlating difficulty with factors other than description length, both in phonological learning and in concept learning. I will not attempt to relate such results about learning difficulty with the question of what evaluation criterion is used by the learner.

[19]Heinz and Idsardi (2013) note a lack of correlation between the complexity of finite-state machines for capturing certain patterns and potentially relevant language classes to which these patterns correspond. Based on this, Heinz and Idsardi suggest that MDL is not an appropriate learning criterion in phonology. Note, however, that the complexity of a grammar is only one part of the MDL criterion: the size of the description of the data given the grammar is just as important as the size of the grammar itself, and without taking it into account it is generally not possible to draw conclusions about the adequacy of the criterion. In addition, Heinz and Idsardi discuss the length of very specific representations – namely, the finite-state machines they use to describe the relevant patterns – and these representations do not correspond to any of the main grammatical formalisms for phonology. Given different representations, grammar size can change. Finally, it is hard to see how the possible correlation of language families with the description length for the best grammar (with or without taking the data into account) is a relevant consideration. The question is whether, given an appropriate representation scheme, the grammar that yields the shortest description in any particular situation is also the one that humans arrive at.

[20]For de Marcken (1996) MDL is a substitute for Structural Risk Minimization, but it is still the sole contributor to the actual evaluation metric used by the learner. While de Marcken's focus is different from that of the present work – in particular, his emphasis of a specific representational framework that he develops can obscure the general applicability of MDL as an immediate CG learning criterion for any explicit UG – his work provides a particularly clear example of how pure MDL can fit in with a linguistically motivated UG.

on the right-hand side.[21] T_1 marks the end of the grammar with an additional separator. For example, the grammar below will be listed as `ABA#ABC#A#BCD#...#EFG##`:

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

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 above, 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.

[21] 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 an MDL learner.

4.2 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:

$$G \quad := \quad \begin{cases} \gamma \rightarrow \ \sigma_1 \ \gamma \\ \vdots \\ \gamma \rightarrow \ \sigma_n \ \gamma \end{cases}$$

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, the search in the simulations below relies on Simulated Annealing (SA, 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. Indeed, it is quite possible that, even if they use the MDL criterion, humans will turn out to be incapable of exploring the search space effectively. If that is the case, the search component could make the learner – and with it the entire innate component – considerably more restrictive than suggested by the representational abilities of UG and by the MDL criterion.[22]

SA proceeds by comparing a current hypothesis to one of its neighbors, chosen at random, in terms of goodness, which in the present case is the total description length. That is, when a current hypothesis G is compared to one of its neighbors, G', $|G| + |D|G|$ is compared to $|G'| + |D|G'|$. If G' is better than G (that is, $|G'| + |D|G'| < |G| + |D|G|$), the search switches to G'. Otherwise, the choice of whether to switch to G' is made probabilistically and depends both on how much worse G' is and on a *temperature* parameter. The higher the temperature, the more likely the search is to switch from G to its bad neighbor G'. Similarly, the closer G and G' are in terms of overall description length, the more likely the search is to switch to G'. The

[22]The idea that a significant part of the restrictiveness of the innate component may be the result of constraints on learning has been pursued in the literature in various contexts. See Saffran (2003), Heinz (2007), and Heinz and Idsardi (2013), for example.

temperature is initially set to a relatively high value, and it is gradually lowered as the search progresses, making the search increasingly greedy. The search ends when the temperature descends below a certain threshold.

For any grammar G, the neighbor grammar G' is generated as a variant of G in which one of the changes in the following list occurs:

1. An element, possibly a new nonterminal, is added to one of the rules.
2. An element is deleted from one of the rules.
3. A new rule of the form $X \rightarrow \epsilon$ is created for some category X.
4. A nonterminal in the right-hand side of a rule is replaced with its expansion according to some rule in the grammar.
5. A nonterminal X in the right-hand side of a rule is replaced with a new nonterminal Y, and a unit rule $Y \rightarrow X$ is added to the grammar.

The modification is chosen according to a uniform distribution over possible changes. All decisions in a given modification are made randomly as well (category for insertion, positions for insertion or deletion, etc.).

4.3 Results

In section 4.1 above we saw the specification of T_1, a simple-minded CFG UG, and in section 4.2 we saw the details of a search procedure that turns the MDL evaluation metric induced by T_1 into a learner. In this section we will see the results of running this learner on two extremely simple data sets: one that is the concatenation of words from an artificial lexicon and another that involves palindromes. Both tasks are loosely based on patterns that arise in natural language. The concatenation data set requires that the learner address the challenge of segmenting the input, a challenge solved by human learners, who are exposed to inputs that are for the most part unsegmented. The palindrome data set requires that the learner address the challenge of acquiring center embedding, a common pattern in natural languages. Despite this loose correspondence with natural language, the goal of the present section is not the realistic modeling of learning in humans – both T_1 and the data sets are far too simplistic to be informative in this respect – but rather to show how a bare MDL learner induced by an explicit UG operates, and how significant patters arise from the representational abilities of the UG in question guide the search for the best hypothesis given the data.

4.3.1 Segmentation

The first data set is based on the one described by Saffran et al. (1996). In Saffran et al.'s experiment, in which 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`).[23] 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:[24]

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

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

G_{1000} :

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

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} :

$$\begin{array}{ll} 5144 \rightarrow t\,i\,b\,u\,d\,o\,5144 & 5144 \rightarrow p\,a\,b\,i\,k\,u\,5144 \\ 5144 \rightarrow g\,o\,l\,a\,t\,u\,5144\,r & 5144 \rightarrow d\,a\,r\,o\,p\,i\,5144 \end{array}$$

[23] 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, repetitions are not prohibited. As far as I can tell, this does not affect the point made here.

[24] 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.

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 4.1 is not aware of the tasks of segmentation and lexicon induction, and it does not represent probabilities in its rules. Consequently, the bare MDL learner for 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.

4.3.2 Palindromes

For our second simulation, along the lines of Horning's paradigm, we will use an input that exhibits nested dependencies. Such dependencies are common in natural language: they are present in the nesting of object-extracted relative clauses in English, for example, as well as in the basic structure of verb-argument dependencies in German clauses. It has been suggested by Fitch and Hauser (2004) that humans acquire such patterns in experiments of artificial-language learning, though the experiment and the claim remain controversial (see Perruchet and Rey 2005, among others).[25]

In the nesting data set I will use a segmented input. We can specify the learner's 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.[26] At least in simple cases, the 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, the learner cannot see past the first 25 characters of each element in the sequence):

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} :

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

[25] The palindrome language is a member of certain interesting infinite classes that can also be learned under the demanding criterion of *iitl*, as shown by Koshiba et al. (1997).

[26] Note, however, that the 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.

Grammar length: 35 Encoding length: 2154 Energy: 2189.0

G_{4200} :

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

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 TL theories of UG and CG theories of learning into closer contact. I reviewed some of the central arguments within each discipline for and against rich UGs and for and against learning, concluding that linguists' notions of rich UGs are well-founded, but that cognition-general learning approaches are viable as well. Differently from what is often suggested in the literature, I argued that the two can and should co-exist and support each other. Specifically, I used 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 an MDL evaluation metric that can serve as the central component of a CG learner. This mapping from theories of UG to learners 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 suggested that such learners should be our null hypothesis regarding the child's learning mechanism, and that furthermore, the mapping from theories of UG to learners provides a framework for comparing theories of UG.

References

Abels, Klaus, and Ad Neeleman. 2010. Nihilism masquerading as progress. *Lingua* 120:2657–2660.

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

Angluin, Dana. 1982. Inference of reversible languages. *Journal of the Association for Computing Machinery* 29:741–765.

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

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.

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

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

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87:84–125.

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral Dissertation, MIT, Cambridge, MA.

Berwick, Robert C., Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35:1207–1242.

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

Bonatti, Luca, Marcela Peña, Marina Nespor, and Jacques Mehler. 2005. Linguistic constraints on statistical computations. *Psychological Science* 16:451–459.

Braine, M. D. S. 1971. On two types of models of the internalization of grammars. In *The ontogenesis of grammar*, ed. D. J. Slobin, 153–186. Academic Press.

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.

Bresnan, Joan, and Ronald M. Kaplan. 1982. Grammars as mental representations of language. In *The mental representation of grammatical relations*. MIT Press.

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, 11–53. New York: Wiley.

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

Chang, Nancy Chih-Lin. 2008. Constructing grammar: A computational model of the emergence of early constructions. Doctoral Dissertation, EECS Department, University of California, Berkeley, Berkeley, CA.

Charikar, Moses, Eric Lehman, Ding Liu, Rina Panigrahy, Manoj Prabhakaran, Amit Sahai, and Abhi Shelat. 2005. The smallest grammar problem. *Information Theory, IEEE Transactions on* 51:2554–2576.

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. 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.

Clark, Alexander, and Rémi Eyraud. 2007. Polynomial identification in the limit of context-free substitutable languages. *Journal of Machine Learning Research*

8:1725–1745.
Clark, Alexander, and Shalom Lappin. 2011. *Linguistic nativism and the poverty of the stimulus*. Wiley-Blackwell.
Clark, Robin, and Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24:299–346.
Crain, Stephen, Drew Khlentzos, and Rosalind Thornton. 2010. Universal Grammar versus language diversity. *Lingua* 120:2668–2672.
Crain, Stephen, and Paul Pietroski. 2002. Why language acquisition is a snap. *The Linguistic Review* 19:163–183.
Debowski, L. 2011. On the vocabulary of grammar-based codes and the logical consistency of texts. *Information Theory, IEEE Transactions on* 57:4589–4599.
Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.
Dowman, Mike. 2007. Minimum description length as a solution to the problem of generalization in syntactic theory. Ms., University of Tokyo, Under review.
Dyer, Fred C, and Jeffrey A Dickinson. 1994. Development of sun compensation by honeybees: How partially experienced bees estimate the sun's course. *Proceedings of the National Academy of Sciences* 91:4471–4474.
Endress, Ansgar, Ghislaine Dehaene-Lambertz, and Jacques Mehler. 2007. Perceptual constraints and the learnability of simple grammars. *Cognition* 105:577–614.
Endress, Ansgar, and Jacques Mehler. 2009a. Perceptual constraints in phonotactic learning. Ms. To appear in Journal of Experimental Psychology: Human Perception and Performance.
Endress, Ansgar, Marina Nespor, and Jacques Mehler. 2009. Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences* 13:348–353.
Endress, Ansgar D., and Luca L. Bonatti. 2013. Single vs. multiple mechanism models of artificial grammar learning. Under review.
Endress, Ansgar D., and Jacques Mehler. 2009b. Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology* 62:2187–2209.
Evans, Nicholas, and Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32:429–492.
Feher, Olga, Haibin Wang, Sigal Saar, Partha P. Mitra, and Ofer Tchernichovski. 2009. De novo establishment of wild-type song culture in the zebra finch. *Nature* 459:564–568.
Feldman, Jacob. 2000. Minimization of boolean complexity in human concept learning. *Nature* 407:630–633.
Feldman, Jacob. 2006. An algebra of human concept learning. *Journal of Mathematical Psychology* 50:339–368.
Fitch, W.T., and M.D. Hauser. 2004. Computational constraints on syntactic processing in a nonhuman primate. *Science* 303:377–380.
Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors, and Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science* 33:287–300.
Garcia, John, Walter Hankins, and Kenneth Rusiniak. 1974. Behavioral regulation of the milieu interne in man and rat. *Science* 185:824–831.

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.
Goodman, N.D., J.B. Tenenbaum, J. Feldman, and T.L. Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive Science* 32:108–154.
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.
Hackl, Martin, Jorie Koster-Hale, and Jason Varvoutis. 2012. Quantification and acd: Evidence from real-time sentence processing. *Journal of Semantics* 29:145–206.
Harbour, Daniel. 2011. Mythomania? methods and morals from 'the myth of language universals'. *Lingua* 121:1820 – 1830.
Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31:190–222.
Heinz, Jeffrey. 2007. The inductive learning of phonotactic patterns. Doctoral Dissertation, University of California, Los Angeles.
Heinz, Jeffrey. 2010. String extension learning. In *ACL*, 897–906.
Heinz, Jeffrey, and William Idsardi. 2013. What complexity differences reveal about domains in language*. *Topics in cognitive science* 5:111–131.
Horn, Laurence. 1972. On the semantic properties of the logical operators in English. Doctoral Dissertation, UCLA.
Horn, Laurence. 2011. *Histoire d'*O: Lexical pragmatics and the geometry of opposition*, 383–416. Bern: Peter Lang.
Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.
Hsu, Anne S., and Nick Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34:972–1016.
Hsu, Anne S., Nick Chater, and Paul M.B. Vitányi. 2011. The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition* 120:380 – 390.
Hunter, Tim, and Jeffrey Lidz. 2012. Conservativity and learnability of determiners. *Journal of Semantics* .
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.
Kapur, Shyam. 1991. Computational learning of languages. Doctoral Dissertation, Cornell University, Ithaca, NY.
Katzir, Roni, and Raj Singh. 2013. Constraints on the lexicalization of logical operators. *Linguistics and Philosophy* 36:1–29.
Kieffer, J.C., and En-hui Yang. 2000. Grammar-based codes: a new class of universal lossless source codes. *Information Theory, IEEE Transactions on* 46:737–754.
Kirby, S. 2002. Learning, bottlenecks and the evolution of recursive syntax. *Linguistic*

evolution through language acquisition: Formal and computational models 173–203.

Kirby, Simon. 2000. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. *The evolutionary emergence of language: Social function and the origins of linguistic form* 303–323.

Kirby, Simon, Kenny Smith, and Henry Brighton. 2004. From UG to universals. *Studies in Language* 28:587–607.

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.

Koshiba, Takeshi, Erkki Mäkinen, and Yuji Takada. 1997. Learning deterministic even linear languages from positive examples. *Theoretical Computer Science* 185:63 – 79.

Kurtz, Kenneth J, Kimery R Levering, Roger D Stanton, Joshua Romero, and Steven N Morris. 2013. Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39:552–572.

Legate, Julie Anne, and Charles Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review* 19.

Lempel, A., and J. Ziv. 1976. On the complexity of finite sequences. *Information Theory, IEEE Transactions on* 22:75–81.

Levinson, Stephen C., and Nicholas Evans. 2010. Time for a sea-change in linguistics: Response to comments on 'the myth of language universals'. *Lingua* 120:2733–2758.

Li, Ming, and Paul Vitányi. 1997. *An introduction to kolmogorov complexity and its applications*. Berlin: Springer Verlag, 2nd edition.

Lidz, Jeffrey, Sandra Waxman, and Jennifer Freedman. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition* 89:B65–B73.

Magri, Giorgio. 2013. The complexity of learning in Optimality Theory and its implications for the acquisition of phonotactics. *Linguistic Inquiry* 44:433–468.

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT, Cambridge, MA.

Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46:53–85.

Marcus, Gary F. 2000. Pabiku and ga ti ga: Two mechanisms infants use to learn about the world. *Current Directions in Psychological Science* 9:145–147.

Matthewson, Lisa. 2012. On how (not) to uncover cross-linguistic variation. In *Proceedings of NELS 42*.

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.

Miller, George, and Noam Chomsky. 1963. Finitary models of language users. In *Handbook of Mathematical Psychology*, ed. R. Duncan Luce, Robert R. Bush, and

Eugene Galanter, volume 2, 419–491. New York, NY: Wiley.
Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25:83–127.
Moreton, Elliott, and Joe Pater. 2012a. Structure and substance in artificial-phonology learning, part i: Structure. *Language and Linguistics Compass* 6:686–701.
Moreton, Elliott, and Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part ii: Substance. *Language and Linguistics Compass* 6:702–718.
Moreton, Elliott, Joe Pater, and Katya Pertsova. 2014. Phonological concept learning. Ms., Submitted, June 2014.
Nevill-Manning, Craig, and Ian Witten. 1997. Compression and explanation using hierarchical grammars. *The Computer Journal* 40:103–116.
Niyogi, Partha. 2006. *The computational nature of language and learning*. The MIT Press.
Niyogi, Partha, and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61:161–193.
Niyogi, Partha, and Robert C. Berwick. 1997. Evolutionary consequences of language learning. *Linguistics and Philosophy* 20:697–719.
Niyogi, Partha, and Robert C. Berwick. 2009. The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences* 106:10124–10129.
Onnis, Luca, Matthew Roberts, and Nick Chater. 2002. Simplicity: A cure for overgeneralization in language acquisition? In *Proceedings of the 24th Annual Conference of the Cognitive Society*, ed. W. D. Gray and C. D. Shunn. London.
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.
Perfors, Amy, Joshua Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118:306–338.
Perruchet, Pierre, and Arnaud Rey. 2005. Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review* 12:307–313.
Piantadosi, Steven T., and Edward Gibson. 2013. Quantitative standards for absolute linguistic universals. *Cognitive Science* n/a–n/a.
Pitt, L. 1989. Probabilistic inductive inference. *Journal of the ACM* 36:383–433.
Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.
Rasin, Ezer, and Roni Katzir. 2013. On evaluation metrics in Optimality Theory. Ms., MIT and TAU (submitted), September 2013.
Reuland, Eric, and Martin Everaert. 2010. Reaction to: The myth of language universals and cognitive science"—evans and levinson's cabinet of curiosities: Should we

pay the fee? *Lingua* 120:2713–2716.
Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.
Rissanen, Jorma, and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, 149. Amer Mathematical Society.
Ross, J. R. 1967. Constraints on variables in syntax. Doctoral Dissertation, MIT, Cambridge, MA.
Saffran, Jenny R. 2003. Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science* 12:110–114.
Saffran, Jenny R., Elissa L. Newport, and Richard N. Aslin. 1996. Statistical learning by 8-month old infants. *Science* 274:1926–1928.
Sandler, Wendy, Irit Meir, Carol Padden, and Mark Aronoff. 2005. The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America* 102:2661–2665.
Sauerland, Uli, and Jonathan Bobaljik. 2013. Syncretism distribution modeling: Accidental homophony as a random event. In *Proceedings of GLOW in Asia IX 2012*.
Senghas, Ann, Sotaro Kita, and Asli Özyürek. 2004. Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science* 305:1779–1782.
Shepard, Roger N, Carl I Hovland, and Herbert M Jenkins. 1961. Learning and memorization of classifications. *Psychological Monographs: General and Applied* 75:1–42.
Smith, K., S. Kirby, and H. Brighton. 2003. Iterated learning: A framework for the emergence of language. *Artificial Life* 9:371–386.
Smith, Kirk H. 1966. Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology* 72:580–588.
Sobel, D.M., J.B. Tenenbaum, and A. Gopnik. 2004. Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive science* 28:303–333.
Solomonoff, Ray J. 1964a. A formal theory of inductive inference, part II. *Information and Control* 7:224–254.
Solomonoff, Ray J. 1964b. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.
Solomonoff, Ray J. 1978. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory* 24:422–432.
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.
Steedman, Mark. 1989. Grammar, interpretation, and processing from the lexicon. In *Lexical representation and process*, ed. William Marslen-Wilson, 463–504. MIT Press.
Steedman, Mark, and Jason Baldridge. 2011. Combinatory categorial grammar. In *Non-transformational syntax*, ed. Robert Borsley and Kersti Börjars, chapter 5, 181–224. Blackwell.

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.

Tily, H., and T.F. Jaeger. 2011. Complementing quantitative typology with behavioral approaches: Evidence for typological universals. *Linguistic Typology* 15:497–508.

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.

Wexler, Kenneth, and Peter W. Culicover. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.

Yang, Charles. 2010. Three factors in language variation. *Lingua* 120:1160–1177.

Yang, Charles D. 2002. *Knowledge and learning in natural language*. Oxford University Press.

Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in Cognitive Sciences* 8:451–456.

Yoshinaka, Ryo. 2011. Efficient learning of multiple context-free languages with multidimensional substitutability from positive data. *Theoretical Computer Science* 412:1821–1831.

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.