17 Learnability

ALEXANDER CLARK

1. Introduction

One of the paradoxes of modern linguistics is that though one of Chomsky's lasting contributions to linguistics is to locate the central empirical problem of linguistics as an acquisition problem, the attention of the field as a whole has been focused on the goals of finding descriptively adequate grammars for natural languages. To a certain extent this has reflected the relative levels of development of the underlying mathematical theory. While the theory of linguistic representation or description has been very well developed, starting with Chomsky's own seminal contributions in the 1950s, the theory of grammatical inference, the corresponding formal discipline that concerns itself with the problems of learning such representations, is much less mature. Indeed the history of grammatical inference has largely been a history of negative results, to such an extent that the only thing most linguists know about grammatical inference is the negative result of Gold (1967), which seems to indicate that the task is impossible. Thus reviews of learnability in linguistics (for example, Niyogi, 2006; Yang, 2008; Clark and Lappin, 2012; Heinz, in press) tend to focus on negative results, with, caricaturing somewhat, the nativists stressing the negative results and the researchers of a more empiricist persuasion downplaying them. Such negative results can in principle serve to rule out purported solutions to the language acquisition problem, and have in the past been taken to rule out a naive "blank slate" empiricism.

More interesting of course are positive results, since we are interested in explaining a phenomenon, language acquisition, which certainly does occur; and these positive results have been, to say the least, in very short supply. Again, Gold's early paper has proved enormously influential: the very limited range of positive results he considered in that paper have been widely, though incorrectly, taken to exhaust the possibilities for language acquisition. In particular his elementary proof that any finite class of languages can be learned from positive data alone has been taken as an important point of reference, and a justification for theories of grammar that take the class of possible human grammars to be finite.

The absence of other positive results has been in its own way more important than the negative results, motivating the strong linguistic nativism advocated by Chomsky (1981) and Pinker (1994), and causing problems for those who find these claims implausible.

The Handbook of Language Emergence, First Edition. Edited by Brian MacWhinney and William O'Grady.

While one can convincingly argue that the negative results have been overstated, that is unsatisfying: merely saying that something is not impossible is not an adequate scientific hypothesis. What is needed is a range of alternative positive answers: here is how language acquisition might take place. Once there are a variety of options on the table, one can explore them using the standard methodologies of science.

The seminal work of Dana Angluin on learning regular languages changed the situation enormously (Angluin, 1982, 1987). Her work showed for the first time that it is possible to learn languages just from a finite amount of information about the strings that are or are not in the language. The algorithms she defined are "inferential" – they deduce the structure of the language from examples, given certain assumptions. Unfortunately, her results were limited to the acquisition of regular grammars (or finite state automata), which, though infinite, lack many of the structural properties which are characteristic of natural language syntax. As a result, with a few exceptions (Pilato and Berwick, 1985), they have not been widely influential in linguistics, though they have been extensively studied in the technical literature on grammatical inference.

In recent years, this work has been greatly extended to encompass the acquisition of the sorts of grammars that are needed to account for natural language syntax: to context-free and mildly context-sensitive grammars. This newly developed theory, which we call "distributional learning," takes the old ideas of distributional learning of the American structuralists (Harris, 1951), and builds on them a rigorous mathematical theory of learnability. While this is still quite new, and thus incomplete, already we have efficient learnability results for classes of languages that plausibly include all natural languages. Of course, since the theory is incomplete, the set of positive results we currently have is still only a partial picture. More such results are on the way. The negative results therefore do come into play; they have an important role: to curb excessive optimism as these results tell us that certain types of positive result are impossible.

Here, we do not make the empirical claim that any of the algorithms is how language acquisition proceeds, but we do claim that from combinations of the results we already have, which we sketch below, there are learning algorithms that fit the gross facts of language acquisition.

In this chapter, we discuss the theory of learnability or grammatical inference, from a positive perspective. We start (in Section 2) by looking at the methodological issues involved in applying the tools of mathematical analysis to the empirical problem of language acquisition, and the various assumptions that we make, and by discussing the problems of grammatical inference. Then in Section 3 we review, non-technically, some recent developments in the field based on the classic ideas of distributional learning.

2. Methodology

We start by articulating some foundational assumptions that make it possible to discuss this from a theoretical perspective. First, we assume that the brain/mind can be fruitfully considered to be a computational system; without this assumption no mathematical analysis can start, and we will approach this formally using the machinery of mathematics and theoretical computer science. So rather than considering computer programs that we can run experimentally on natural language corpora, we study algorithms that are guaranteed to learn in some precise sense. The mathematical proofs guarantee the

correctness of the algorithms in a way that even the most thorough empirical experiments cannot. Moreover, they can give us direct insight into the classes of languages that are learnable under various paradigms. Running a computer program on a CHILDES corpus (MacWhinney, 1995), even if it works to some extent, tells us nothing about the properties of the language that make it work, nor about whether it will work on other corpora of child-directed speech from other languages not in the CHILDES collection. As Keller and Asudeh (2002) say: "A generative grammar is empirically inadequate (and some would say theoretically uninteresting) unless it is provably learnable." The real story of language acquisition – a decade-long interaction between a rapidly developing child and a community of adults – is far beyond what we can capture in a tractable mathematical model. Nevertheless, we can define simple models of learning that give insight into the basic possibilities of learning grammars from examples.

2.1 Outputs

We are interested in modeling at some level of abstraction the language acquisition process; a first step is to consider the inputs and outputs of the process, to define what the data is that the learner receives, and what sort of output the learner should produce. We start by considering the outputs of the process: what kind of object is the thing that is learned by the language acquisition device (LAD)? What type of object is the grammar or I-language that is produced?

In Chomsky's phrase, language is a system of "discrete infinity": though acoustically it varies continuously, linguistically it consists of discrete words arranged sequentially, and the sentences can be of unbounded length. We take a fairly standard view: we assume we have some internal grammar that generates some unobserved (or "latent") hierarchical structures that are then mapped to phonological and semantic representations. We wish to draw our net as widely as possible. If we commit to one particular representation, such as the version of the Standard Theory studied in Wexler and Culicover (1980), and that representation is abandoned for empirical or other reasons (as the Standard Theory was), then the analysis becomes outmoded. Recently a broad consensus has developed in the mathematical linguistics community (Stabler, 2011) on the appropriate class of grammars. It transpires that formalisms that appear very different superficially are in fact mathematically equivalent in a strong sense. This equivalence even spans one of the most fundamental divides in syntactic theories: between theories that use movement and those that do not. Most current proposals are equivalent to some subclass of the class of multiple context-free grammars (Seki, Matsumura, Fujii, and Kasami, 1991): this includes tree adjoining grammars of various types, minimalist grammars, and so on (Joshi, Vijay-Shanker, and Weir, 1991; Borsley, 1996; Stabler, 1997). Thus we can consider learning approaches that output grammars of this type, and be confident that they are adequate.[1] As Stabler (2013) puts it, "This consensus is stable and rather well understood."

2.2 Inputs

We now consider the inputs to the learning algorithm. Classically, the input to the learner has been considered to be only strings: sequences of sounds that the child passively

observes (Gold, 1967; Chomsky, 1962). We take these to be sequences of categorized speech sounds, as phonemes or phones, glossing over the problems of low-level acoustic processing, phonology and such like, all of which raise interesting issues.

The normal situation of the child is much richer in a number of respects; the child can interact with the parent/caregiver in a number of ways, has information about the situational context of the utterance, and so on. In particular the child can observe what objects are present, and what events are happening as utterances are being made, together with other indicators of salience such as which objects are held in the parent's hand, the gaze of the parent, and other factors. Moreover, the child is not entirely passive: the child can act by moving, pointing or looking in particular directions, picking up objects in response to requests and questions, and so on; additionally, the child can generate utterances of its own, well-formed or ill-formed, and these actions will have direct or indirect effects on what happens (Tomasello, 2003; E. Clark, 2009).

Of course, language acquisition is not just a case of learning which sentences are grammatical and which not: The child also learns the meanings and communicative functions of sentences. This can clearly not be learned just from the strings, but requires information about the situational context.

In terms of the architecture of the grammar, then, we have three types of object: we have the surface strings of acoustic symbols, the meanings of the utterances, and the putative hierarchical structures that underlie these pairings. We assume that in all plausible models the child has access to the sequence of speech sounds. The next more controversial question is about the degree to which learners have access to the meanings of the utterances they hear. Here opinions are divided.

The most optimistic proposal is that the learner has access to the complete semantics of the utterance: the learner is thus presented with sound/meaning pairs, where the meaning is taken to be some hierarchically structured semantic representation, a well-formed formula in some innate language of thought (Wexler and Culicover, 1980; Pinker, 1995; Siskind, 2000; Kwiatkowski, Zettlemoyer, Goldwater, and Steedman, 2010). The assumption is then that the child is able to infer the meaning of the utterance by combining some ability to reason about the intentions of the speaker with information about the various salient events happening, and with some partial information about the syntactic structure of the utterance. Certainly, in the later stages of language acquisition this may be possible; but in the very early phases of language acquisition, when the child does not know what the words are or what the syntactic structure is, it seems implausible.[2]

Finally, we consider the question of whether children have direct information about the hierarchical structures. Clearly they do not: some models assume the trees as input, but this is only in the context of the existence of some other learning component that can infer the trees (Wexler and Culicover, 1980); from a learnability point of view this begs the question. We therefore must assume that the child does not have access to these structures. Any structures that the learner uses must be constructed by the learner during the course of acquisition. Thus our learning model must account for these inside the theory, and not posit them as inputs.

Clearly, having one model that is intended to cover the entire process of language acquisition from the earliest stages to adult grammar is a gross simplification; so perhaps the models considered here are best thought of as models of the earliest period of language acquisition, when the child has learned the acoustic categories of the language

and nothing else. Once the learner has some partial knowledge of the syntactic structure of the language, of the lexical categories and of the meanings of frequent words, the learner can leverage its existing knowledge in a number of ways. The problem is at its most acute when the child is youngest and so that is where we should focus our analysis.

2.3 The problems

Grammatical inference is hard but not impossible; it is important to understand the various difficulties in order to see how they can be overcome. Classically, the difficulties arise from two distinct yet interacting factors. First there are problems to do with whether there is enough information in the input data for the learner to succeed; we will call these, rather loosely, information-theoretic problems. The second class of problems concern the computational issue of using this information to construct a hypothesis; it may be that, though there is enough information available in some mathematical sense, there are computational problems that cannot be solved efficiently. We can call these computational complexity problems. The information-theoretic problem has been studied extensively for 50 years, and the computational complexity problems for nearly as long (Gold, 1978), and they are now well understood; recently a third problem has come into sharper focus, which we call the strong learning problem, which interacts with our assumptions about the semantic information available in the input, which we discuss later.

Attention in linguistics has focused largely on the first of these, the information-theoretic problems, which has been considered to be a "logical problem" (Baker and McCarthy, 1981). There are a number of reasons for this focus: the study of grammatical inference predates the development of the theory of computational complexity, and, furthermore, these problems are to some extent more fundamental than the complexity problems, since if the information is inadequate then the question of the computational complexity of inference cannot even be formulated coherently. The negative results here, such as those of Gold (1967), show that, for any learner, the class of languages learned must be limited in some way. These results are often taken to show that the learner must have constraints on its hypothesis space, on the set of grammars that it might output. However, the arguments here are unsound: they conflate the hypothesis space of the learner with the learnable class of grammars, which can and sometimes must be very different. The hypothesis space of the learner must contain, but in general is not equal to, the class of grammars that can be learned. Though the arguments from Gold's theorem show that the latter must be limited, they say nothing about the former, and it is the former – the hypothesis space of the learner – that is the object we are primarily interested in, in the case of language acquisition. See Clark and Lappin (2013) for detailed discussion of this point.

Nonetheless, the Gold analysis put the focus sharply on one particular aspect of the problem: the absence of negative evidence. Gold's results suggested that without explicit negative evidence only very small classes of languages could be learned. In particular the problem of recovering from overly general hypotheses without correction seemed to be impossibly hard. If the learner's hypothesis failed to generate a particular grammatical sentence, then the learner could notice this since he/she would observe one of these sentences and realize the error, but the converse problem of overgeneralization seemed much harder. If the learner has a hypothesis that generates some ungrammatical sentences, then in the Gold model there seemed no way that the child could recover. It

was rapidly realized that the probabilistic nature of the input could serve an important role. If the learner uses a probabilistic model, then it could detect overgeneralization. The learner would expect to see sentences of a certain type appear with a particular frequency. If the actual frequency that the learner observes is less than this expectation, then this is a cue to the learner that the model overgeneralizes in some way. This "indirect negative evidence," as it came to be called (Chomsky, 1981: 9), allows the learner to recover from these errors. Fairly soon after Gold's results, some formal results appeared (Horning, 1969) which indicated that this was mathematically well founded, and these results have been extended enormously since then (Chater and Vitányi, 2007; Cohen and Smith, 2012). These modern results indicate that random positive examples do contain enough information to pick out a correct grammar. From a modern perspective the standard Gold model is misleading, because the examples from which the learner learns are not generated randomly, or helpfully, but rather may be generated adversarially; the learner is required to succeed even when the examples are being generated with an intent to mislead. And under this model, indeed, recovering from overgeneralization can be hard. Results like that of Cohen and Smith (2012) show that, under reasonable assumptions about the distribution of examples, this part of the learning problem is tractable.

The second problem, which has received much less attention, concerns computational complexity: how can the child efficiently construct a grammar given the information it receives? That is to say, suppose the child does in fact receive enough information in principle to pick out the right grammar; given the fact that the human brain is a finite computational system with limited resources, how can the child work out which is the correct grammar? This constraint, which applies in language acquisition as much as in any other area of cognition (van Rooij, 2008), more directly indicates the properties of the learner. Here we restrict ourselves to learners that are computationally efficient in a standard technical sense – polynomial time. Many results over the years have shown that hidden inside various learning tasks are intractable computational problems: Kearns and Valiant, 1994; Angluin and Kharitonov, 1995; Abe and Warmuth, 1992; Cohen and Smith, 2012. In the next section we will discuss one family of computationally efficient learning algorithms.

Lurking behind these problems is the real target of our inquiry: the nature of the learner. Clearly the child brings innate biases to bear on the learning process; on the basis of a finite amount of data, the child converges to one hypothesis amongst many that are compatible with the data. The existence of these biases is not a point about which there is any substantive disagreement.[3] The debate is about the nature of these biases, and in particular about whether they are domain-specific or not, and if they are how rich and complex they might be.

3. Analogy and Distributional Learning

One naive view of language acquisition is that children learn language by a process of generalization and abstraction based on some notions of similarity, typically derived from distributional properties of the language. For example, a child might hear the following sequence of examples: "look at the dog," "look at the cat," "look at the car," and on the basis of these examples assume that the words "cat," "dog," and "car" are members of the same lexical category. Therefore, given some additional sentence, say

"I want a dog," the learner might conclude, without ever having observed them, that the sentences "I want a car" and "I want a cat" are also grammatical. This has a certain superficial plausibility, but this naive view was recognized early on as being acutely problematic; indeed Chomsky (1955) devotes some attention to pointing out some of the many problems with this approach. For example, the famous pair of examples "John is easy to please," "John is eager to please" serves, if nothing else, to illustrate one problem. The words "eager" and "easy" occur in some of the same contexts, but are clearly different syntactically, as the contrast between "John is eager to die" and *"John is easy to die" illustrates.

Nonetheless, it is worth examining this naive approach, in the full awareness of its inadequacy, to get some insight into the possibilities of this type of learning. Rather surprisingly, though such approaches have often been appealed to as heuristics (e.g. Adriaans, 1999) it was not until quite recently that an adequate formal account of this approach was developed, in Clark and Eyraud (2007).

We start by considering this most basic model. Clark and Eyraud (2007) formalize the simplest intuition of string substitutability as the basis for a learning algorithm. In order to explain this approach, we need to start by defining a few technical terms. First, we use the term *string* to mean a sequence of one or more words. So an individual word like "cat" or "the" is a string, and so is a sequence like "the cat" or "dog on the." Some of these strings may correspond to syntactic constituents and some may not. Secondly, we want to define the notion of *substitutability*. The most basic notion here is when two strings are completely substitutable, which relation we call *congruence*. Call the two strings u and v. These two strings are *congruent* if, whenever we have a grammatical sentence that contains u, we can replace that occurrence of u with v and the result will also be grammatical and, conversely, if we have an ungrammatical sentence and we swap u and v, then the result will remain ungrammatical. We will denote this by $u \equiv v$. This is a very strong criterion, unreasonably strong for the case of natural languages, where it is quite hard to find two strings that are completely substitutable. For example "Monday" and "Tuesday" are plausibly congruent in this sense, but beyond such simple examples phenomena such as lexical ambiguity complicate the picture. Now this precise mathematical property depends in principle on checking an infinite number of sentences for grammaticality; we therefore need to consider a slightly weaker relationship that we call *weak* substitutability. Two strings u and v are weakly substitutable in this sense if and only if there is *some* context within which they both occur. So, for example, "eager" and "easy" are weakly substitutable in this sense, since they can both occur in the context "John is _ to please," but are not congruent. We can write this weaker property as $u \cong v$.

Let us assume now, counterfactually, that natural languages are such that whenever two strings are weakly substitutable they are congruent, that if $u \cong v$ then $u \equiv v$. This is not true for English, nor, we assume, for any other language, and though we shall remove this idealizing assumption later on, for the moment we will proceed on the basis that it does hold.

Under this assumption, it is easy for a learner to determine whether or not two strings are congruent: the learner waits until it sees two strings that differ only in having u and v swapped, at which point it knows that they are congruent. If it observes two sentences lur and lvr, where l and r are arbitrary, possibly empty, strings, then $u \equiv v$. This assumption then licenses a certain amount of generalization beyond the sentences that the learner has actually observed. But how can we efficiently construct a grammar based on this insight?

First, note that the congruence relation is an equivalence relation, and as a result we can divide all strings into non-overlapping equivalence classes, where each string is in the class of strings that it is congruent with. These classes, called the congruence classes of the languages, represent objective clusters of distributionally identical words or strings, and can be used for the basis of a grammar. In short, the learner constructs a grammar where the nonterminals of the grammar correspond to these congruence classes, in the sense that each nonterminal will only generate the strings that are in that congruence class. So for example, if we have a congruence class of the days of the week, Monday through Sunday, then we will have a nonterminal for this class, and that nonterminal will just generate those seven words, and no others.

It turns out that languages that are not regular[4] have an infinite number of congruence classes, and thus there is a need for some process to select which of these classes should be used as the basis of the grammar. We shall return to this problem later, but for the moment we consider a learner that naively picks the congruence classes of all the substrings of the sentences that it observes.

Given these nonterminals, the learner must then define a set of productions or rules; this is trivial because of the following fact. Suppose we have four strings, u, u', v, v', such that $u \equiv u'$ and $v \equiv v'$; in other words we have two pairs of congruent strings. Then it is easy to see that $uv \equiv u'v'$. In other words the concatenation of u and v will be congruent with the concatenation of u' and v'. If we write X for the nonterminal corresponding to the congruence class that contains both uv and $u'v'$ and Y for the one corresponding to the congruence class containing u and u', and similarly Z for the one for v and v', then we should have a rule of the form $X \to YZ$. Once we have decided what the nonterminals represent, finding the rules is mechanical and in this case quite trivial. This approach gives us an algorithm that can learn every context-free language that satisfies the substitutability property.

3.1 Discussion

This learner is so simple that it is easy to understand its properties, properties that are often shared in a more obscure manner by more complex learners. So, before we move to consider the limitations of this algorithm as a model for language acquisition, and sketch more powerful learners that overcome those limitations, it is worth pausing to consider some of the issues it raises in miniature.

First, the model is in a certain sense inferential or deductive. It constructs a grammar on the basis of identifying mathematically well-defined patterns from a finite collection of examples; as a result it is computationally efficient in the standard technical sense;[5] from these patterns it creates a grammar in a principled way. It also learns rapidly: it is guaranteed to converge once it has seen a small number of examples.[6]

The learner is, in addition, correct for a certain class of languages. For languages in this class, it is mathematically guaranteed to converge exactly to a grammar that precisely generates the language it is trying to learn. That is, it is guaranteed to succeed at learning the right set of strings. The class of languages it can learn is infinite and contains infinite languages, including infinitely many languages that are properly context-free. The grammars it produces are of a very classic type: context-free grammars. These are rule-based symbolic grammars that generate hierarchically structured parse-trees. Thus, the approach here is very different from the connectionist learning literature, which in

general rejects explicit rule-based grammars. This learner uses positive data only; it does not interact at all with the environment or teacher. It is purely passive, and learns under a worst-case environment, even when the examples are being generated by an adversary who is trying to make the learner fail, that is, under an unrealistically strict learning model.

Finally, and more abstractly, the learner is objective or, for lack of a better word, empiricist. The representational primitives of the grammar, the nonterminals that correspond to the notion of syntactic category, are objectively based on properties of the language considered as a set of strings. In this case, they correspond to distributional equivalence classes, but other approaches may make a different representational decision, as we shall see later.

This is the most basic model of distributional learning, and like all elementary models it is inadequate in several respects. As Clark and Eyraud (2007) admit, this learning algorithm relies on a property – substitutability – that natural languages simply do not have. This is partly because the learning model is overly restrictive and neglects the importance of probabilistic data. It is also limited in a number of other ways: the class of representations, context-free grammars, is too small, and the result is only a "weak" one, in that it generates only the correct set of strings and not an appropriate set of structures. Nonetheless, it establishes an important point: there are mathematically precise, computationally efficient models that can learn classes of context-free grammars. While not quite the first result in context-free learning (see Lee, 1996 for a useful survey of early work), it has proven to be fruitful, and many of the limitations of this original paper have been eliminated in later work.

3.2 Extensions

We now consider how the limitations of this work have been overcome in subsequent extensions. We start by considering one important and controversial modification we make to the learning model. The substitutable learner was entirely passive; it did not interact at all, and moreover it learned under an adversarial paradigm. As a result, from the negative results we know above, it is impossible for the learner to succeed for classes of grammars that include all finite languages, and in this case it relies on a closure property that does not hold. There are a number of reasonable ways to proceed: one way is to consider a probabilistic learning model where we continue to assume an entirely passive learner, but where we consider the examples the learner observes to be generated according to some random process. Obviously, one needs to place some constraints on what the random process might be. Given a suitable set of constraints on the possible distributions of examples, it is possible to define learners that are correct and efficient, but can learn much larger classes of languages (Clark and Thollard, 2004a; Clark, 2006; Luque and Infante-Lopez, 2010; Hsu, Kakade, and Zhang, 2012; Shibata and Yoshinaka, 2013). This is possible, but extremely difficult in two respects. First, we need to make a collection of assumptions in the learning model that are hard to justify: assumptions about the process that generates the data, and about the parameters of the distribution, that are motivated entirely by the requirements of the learner rather than by observations about what the actual distribution of examples is. Secondly, the highly technical nature of the proofs involved make it hard to see how these can be extended to much richer models.

The alternative is to consider a more idealized model. Any model that is mathematically tractable will make some unrealistic assumptions; rather than making many questionable assumptions that are obscure and hard to understand, we make one very simple assumption that is easy to understand, and work with that. The assumption we make is that the learner can ask whether a given string is grammatical or not. In terms of language acquisition this means that we assume that the child receives some feedback on whether its utterances are well formed or not. In learning theory we formulate this as an oracle: the learner constructs a string and submits it to the oracle, which returns a yes or no answer depending on whether it is in the language or not. The oracle knows what the language is and is assumed to answer perfectly. We call these questions membership queries. In the real world there are of course no oracles; but there are parents and caregivers who also know whether sentences are grammatical or not, and the child does interact with them in a number of ways. The statistical properties of the input are such that, as has been noted before, it is plausible that a certain type of indirect negative evidence is available: from the absence of examples that one would expect to see, one can conclude that they are ungrammatical (Chomsky, 1981: 9). This is undoubtedly a controversial assumption, but there are a number of reasons why it is nonetheless reasonable. First, for the case of regular languages, where the theory of learning is well developed, there are algorithms that use only probabilistic positive data to learn from, and which can learn the whole class of regular languages (Oncina and Garcia, 1992; Clark and Thollard, 2004b; Hsu et al., 2012); these results show that membership queries can be replaced with "indirect negative evidence." This work is now being applied to recently developed algorithms for context-free grammars (Shibata and Yoshinaka, 2013). Secondly, there are strong theoretical results extending Horning (1969), such as Chater and Vitányi (2007), which suggest that under very mild conditions one can learn arbitrary languages from positive data alone; similar, more specific results for context-free grammars confirm this (Cohen and Smith, 2012). Finally, in computational experiments in grammatical inference, the absence of negative data is rarely if ever a problem (Starkie, Coste, and van Zaanen, 2004). People learn from what there is, not from what there is not. Observations will only ever tell you what happens, and learning proceeds happily in the absence of evidence about what does not happen. It is only because of a misinterpretation of the Gold negative result that the problem of no negative evidence has loomed so large in the linguistics literature.

Learning algorithms that use membership queries can learn larger classes of languages. Angluin (1987) showed that there is an efficient learner that can learn the whole class of regular languages, a class that is impossible to learn in the Gold paradigm. In a similar way, the substitutable learner has been extended in a number of ways using membership queries. Clark and Yoshinaka (2012) presents a learner which uses a representation called parallel multiple context-free grammars (Seki et al., 1991), which augments context-free grammars with two additional operations: a copying operation, and an ability to manipulate tuples of strings, which allows for the treatment of movement/displacement phenomena. This allows the algorithm to learn even quite exotic phenomena, for example case-stacking in Australian languages such as Kayardild, and various other forms of copying that occur in natural language (Kobele, 2006). This learner can learn a class of languages that appears to be weakly adequate for the description of natural language syntax. Moreover, the clusters of strings on which it bases the nonterminals no longer correspond as before to congruence classes

of distributionally identical strings, but rather to a hierarchy of clusters of strings that are distributionally similar but not necessarily identical: a much better fit to the linguistic facts.

In this model, we consider the idea of an environment and of the distribution in a much more general way, as specifying the relation between the yield of a subderivation and the entire surface sequence of words. More precisely, we consider it as a function that takes the subyield and constructs the whole sentence. In the case of a context-free grammar these functions have a very simple form: they simply concatenate a string before the yield and a string after. Given a noun phrase whose yield is "a biscuit," the result of integrating it into an entire sentence will be something like "I want a biscuit now." This reduces to the standard idea of a context: inserting something into the context "I want _ now." Richer models will have richer notions of yield and context; the yield might be a tuple of strings, where one part of the string might be a constituent that is in the process of moving from one place to another, and the context might include an operation that duplicates some part of the yield, thus allowing for copying. For almost any derivational model, or any type of phrase structure grammar, we can define an appropriate model of context and yield, and use these techniques to learn a grammar subject to various constraints.

It is not the case though that these learners can learn everything. Though they can learn all finite and all regular languages, it seems that there are some simple context-free languages that cannot be learned using these approaches. So, for example, consider the standard example of a context-free language that is not regular, the language that consists of any number of *a*s followed by an equal number of *b*s:

$$\{a^n b^n \ : \ n > 0\}$$

This language is easily learnable using these distributional techniques. The slightly modified language that consists of any number of *a*s followed by a number of *b*s which is *not* equal to the number of *a*s is not learnable using these techniques:

$$\{a^m b^n \ : \ m, n > 0, \ m \neq n \}$$

This language, though a simple context-free language, is highly unnatural from a linguistic point of view. Distributional learning then makes the correct prediction that these languages are not possible human languages.

Additionally, the learnable classes are not closed under the standard language theoretic operations of union, concatenation, and so on. Thus, consider a language that forms polar questions by reversing the sequence of words in the declarative sentence. So, in this language, the polar question corresponding to "John is happy" would be "happy is John?" If we denote the set of declarative sentences by D, then the language would contain $D \cup D^R$, where D^R is the set of reversed declarative sentences that are the questions. There are languages where D is learnable but $D \cup D^R$ is not. Beyond languages that are not learnable at all, some languages will be learned before others. Given a finite amount of information about the strings in the language, the learner comes to a conclusion about the rest of the language, even though there are infinitely many possible options compatible with the data it has seen so far. In some respects this is a version of a simplicity measure: some languages will be considered before others, and some will not be considered at all. These algorithms therefore have a clear bias, as all

learning algorithms must – they are based on specific notions of similarity – but it seems that the principles are not specific to the domain of language in any meaningful way.

3.3 Strong learners

A more fundamental problem with these approaches is pointed out by Berwick, Pietroski, Yankama, and Chomsky (2011), which we can cast in terms of the classical distinction between strong and weak generative capacity. The results that we have discussed so far have all been what we can call weak learning results. They receive as input flat sequences of words, and generalize to a grammar that generates an infinite set of sentences. While this is mathematically well founded, and is a standard model in the formal analysis of learnability from Gold onwards (see e.g. Niyogi, 2006; Yang, 2008), Berwick et al. argue that it is irrelevant to the real problems of language acquisition, which concern the learning of the hierarchically structured expressions that underlie the sound/meaning relationship that lies at the heart of linguistics. From their perspective then, these results are simply addressing the wrong problem, and as a result shed no light on the problem of language acquisition. Of course, Berwick et al. are completely correct that a theory of language acquisition needs to account for the acquisition of the syntax/semantics interface as well. The methodological question is how to approach this.

There are, broadly speaking, two options. One is to consider the model where the learner receives sentence/meaning pairs, and must converge to a grammar that generates the correct infinite set of sentence/meaning pairs. We call this weak semantic learning. This has some history: indeed it is sometimes considered to be the default model for language acquisition. In this model it is assumed that somehow the learner can infer the meaning of an utterance from some partial knowledge of the meaning of the words, the situational context and the intentions of the speaker (Wexler and Culicover, 1980; Pinker, 1995). The learner then receives as input not just the utterance considered as a sequence of words, but also a representation of its meaning.

The other is to consider a strong learning model where the learner only receives the surface strings, but must learn to generate a correct set of structural descriptions. Strong learning is often thought to be too hard to be a productive research strategy. In this model, the learner receives only the utterances as input, but is required to acquire a grammar that not only generates exactly the right set of grammatical strings, but also generates some appropriate syntactic structures or structural descriptions (SDs).

Formally, we can frame this as saying that the learner must learn a grammar that is structurally equivalent to the target grammar, in some appropriate technical sense: in the case of context-free grammar, this might be isomorphism of the two grammars. This model again has some problems: first, the degree of convergence seems too strong. We do not in fact observe a convergence of structural descriptions as required by this model, since we do not observe the structural descriptions at all. All we observe is a convergence of the sentence/meaning pairs, which does not require in general a convergence of the structural descriptions themselves, though such a convergence is often assumed without argument. It is entirely possible that two different native speakers might have grammars which assign slightly different SDs to the same strings, and yet agree on the available

readings for those sentences; moreover, it is not clear that constituent structure trees, rather than some weaker notion of dependency structure, is the right model for SDs. Secondly, it is generally taken to be too hard. Until very recently, it seemed for technical reasons implausible to suppose that there were learning algorithms capable of strong learning.[7] We will remain agnostic here as to which is the most appropriate model.

For weak semantic learning, the ease of the learning task depends on what precisely the form of the semantic representation is taken to be. If the semantic expression has some hierarchical structure that corresponds to the syntactic structure of the utterance, then, unsurprisingly, this is enormously helpful as the learner can exploit this structure, mapping it back onto the surface string. In the most extreme case, one can take the semantic structure to be isomorphic to the syntactic structure, which massively simplifies the learning problem but at the cost of a highly unrealistic assumption about the syntax–semantics interface. Phenomena like expletive pronouns, the variety of different lexicalization patterns across languages, coordination, and ellipsis, to name but a few, conspire to make the mapping from semantics to syntax quite complex, even if the converse mapping from syntax to semantics is quite well behaved.

The most extreme assumption then is that the "semantic" information is in fact a labeled parse tree, in which case the learning problem is entirely trivial. Less extreme assumptions lead to harder learning problems (Sakakibara, 1992; Dudau-Sofronie, Tellier, and Tommasi, 2003). If we merely take the semantic representation to be some formula that is computed from the structural description, but may not have a structure that is directly related to it, then the learning problem is much harder. Yoshinaka and Kanazawa (2011) show some preliminary results along these lines, again using distributional learning.

There has also been recent progress with strong learning. Any strong learner is *a fortiori* a weak learner; therefore it is natural to try to build a strong learner on top of a weak learner, in other words to start by solving the easier problem – the weak learning problem – and then later to extend the learner to solve the harder problem. To explain how one might do this, it is important to understand how a weak learner can fail to be a strong learner. A weak learner is guaranteed to converge to a grammar that generates the right set of strings. However, if we run it several times on different inputs it is possible that it may converge to a structurally different grammar each time: to different grammars that each define the same set of strings. A deceptively simple solution is to add a component that converts a grammar into a standard form for each language in a certain class. That is to say, for each language L in a class of languages, we define a single grammar G_L that generates L. Then we have a component that, given any grammar G that generates L, will return G_L. If we have a component like this, which we can call a canonicalizer, then we can turn a weak learner into a strong learner: we use the weak learner to learn a grammar G, then we give this grammar to the canonicalizer and return the output. If the weak learner is correct, and the canonicalizer functions properly, then the overall system will be a strong learner.

Of course, things are not quite as straightforward as all this. There are two problems: first, canonicalizing an arbitrary context-free grammar is impossible, or at least not computable, given the undecidability of the equivalence of context-free grammars. Secondly, we will only have one grammar for each language: while this is essential given

the strictures of the learning model, it greatly limits the classes of grammars that can be learned. From one perspective this second factor is not a problem but an advantage: as a result this model makes some very strong predictions about the set of string/meaning pairs that is possible. To take one simple example, suppose we have two words that are distributionally identical, or congruent in our terms, u and v. In the strong learning model, this implies that they must also be syntactically identical, in the sense that the degree of ambiguity of a sentence with u in it must be exactly the same as the degree of ambiguity of that sentence with that u swapped with a v. The congruence of u and v implies immediately that if lur is grammatical then so is lvr. The stronger implication is that if lur has n distinct readings then so does lvr. This from one perspective is completely obvious, and indeed it seems to hold, but this is not predicted by any other theory of grammar that we are aware of. Clark (2013) presents a preliminary result along these lines, developing ideas from Clark (2011). The technical trick used there is to select as nonterminals only those congruence classes that cannot be represented as a combination of other congruence classes.

4. Conclusion

Looking back critically at these algorithms, it is clear that none of the models we have looked at is completely adequate as a model of language acquisition. Though they are all computationally efficient, the models that can learn rich enough classes of grammars are only weak learners, that use membership queries, whereas the strong learners that do not use membership queries can only learn very small classes of grammars. Nonetheless, it is plausible to consider combinations of these learners: a strong learner that can learn parallel multiple context-free grammars using only a suitable source of probabilistic data. Such a learner is not to hand at the time of writing, but it seems plausible that one does exist: it is not ruled out by the negative results, and all of its components have already been developed independently; such a learner would be an adequate model of language acquisition. A blanket rejection of empiricist learning techniques as impossible in principle or as vague and imprecise is thus no longer tenable: there are completely specified, provably correct distributional learning algorithms that under a variety of assumptions can learn hierarchically structured grammars of the right types for natural language syntax. Of course these learners start with a rich set of innate biases; these biases though are of a very different type to those normally considered by linguists, since they are more abstract and have little or no language-specific knowledge in them. We posit a UG of course, as all theories must, but a version of UG which is small and as far as we can see domain-general; learnability theory does not, as we understand it now, motivate anything more.

Acknowledgments

I am grateful to Shalom Lappin, Ryo Yoshinaka, and others for many helpful discussions over the years and to Jeff Heinz and the editors for helpful comments on an earlier draft.

NOTES

1 In fact we use in the end a slightly more powerful but still efficient formalism, Parallel Multiple Context-Free Grammars, which have an additional copying operation, though this extra power may not be necessary.
2 See Clark and Lappin (2011) for further discussion.
3 It is worth noting that there is some debate about what exactly the term innate means; see Mameli and Bateson (2006) for discussion.
4 A regular language is one that can be generated by a regular grammar or finite-state automaton, the lowest level of the Chomsky hierarchy. It is well known that these grammars are inadequate to represent natural language syntax.
5 It is also a perfectly practical algorithm which has been implemented and tested on artificial examples.
6 For technical details see the original paper and Yoshinaka (2008).
7 It is worth remembering that Angluin's results are in fact strong learning results in this sense, albeit for a class of grammars without interesting structural descriptions.

REFERENCES

Abe, N. and M. K. Warmuth. 1992. On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning* 9: 205–260.

Adriaans, Pieter. 1999. Learning shallow context-free languages under simple distributions. Technical Report ILLC Report PP-1999-13, Institute for Logic, Language and Computation, Amsterdam.

Angluin, D. 1982. Inference of reversible languages. *Journal of the Association for Computing Machinery* 29(3): 741–765.

Angluin, D. 1987. Learning regular sets from queries and counterexamples. *Information and Computation* 75(2): 87–106.

Angluin, D. and M. Kharitonov. 1995. When won't membership queries help? *Journal of Computer and System Sciences* 50(2): 336–355.

Baker, C. L. and J. J. McCarthy. 1981. *The Logical Problem of Language Acquisition*. Cambridge, MA: MIT Press.

Berwick, R. C., P. Pietroski, B. Yankama, and N. Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35: 1207–1242.

Borsley, Robert D. 1996. *Modern Phrase Structure Grammar*. Oxford: Blackwell.

Chater, N. and P. Vitányi. 2007. "Ideal learning" of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51(3): 135–163.

Chomsky, Noam. 1955. The logical structure of linguistic theory. PhD thesis, MIT.

Chomsky, Noam. 1962. Explanatory models in linguistics. In Ernest Nagel, Patrick Suppes, and Alfred Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, pp. 528–550. Studies in Logic and the Foundations of Mathematics 44. Stanford, CA: Stanford University Press.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Clark, Alexander. 2006. Pac-learning unambiguous NTS languages. In Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino, and Etsuji Tomita (eds.), *Grammatical Inference: Algorithms and Applications*, pp. 59–71. Berlin/Heidelberg: Springer.

Clark, Alexander. 2011. A language theoretic approach to syntactic structure. In Makoto Kanazawa, András Kornai, Marcus Kracht, and Hiroyuki Seki (eds.), *The Mathematics of Language*, pp. 39–56. Berlin/Heidelberg: Springer.

Clark, Alexander. 2013. Learning trees from strings: A strong learning algorithm for some context-free grammars. *Journal of Machine Learning Research* 14(December): 3537–3559.

Clark, Alexander and Rémi Eyraud. 2007. Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research* 8 (August): 1725–1745.

Clark, Alexander and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Malden, MA: Wiley-Blackwell.

Clark, Alexander and Shalom Lappin. 2012. Computational learning theory and language acquisition. In R. Kempson, T. Fernando, and N. Asher (eds.), *Philosophy of Linguistics*, pp. 445–475. Amsterdam: North Holland.

Clark, Alexander and Shalom Lappin. 2013. Complexity in language acquisition. *Topics in Cognitive Science* 5(1): 89–110.

Clark, Alexander and Franck Thollard. 2004a. PAC-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research* 5: 473–497.

Clark, Alexander and Franck Thollard. 2004b. Partially distribution-free learning of regular languages from positive samples. In *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 85–91. Stroudsburg, PA: Association for Computational Linguistics.

Clark, Alexander and Ryo Yoshinaka. 2012. Beyond semilinearity: Distributional learning of parallel multiple context-free grammars. In Jeffrey Heinz, Colin de la Higuera, and Tim Oates (eds.), *The Eleventh International Conference on Grammatical Inference. JMLR: Workshop and Conference Proceedings* 21, pp. 84–96.

Clark, Eve. 2009. *First Language Acquisition*. Cambridge: Cambridge University Press.

Cohen, S. B. and N. A. Smith. 2012. Empirical risk minimization for probabilistic grammars: Sample complexity and hardness of learning. *Computational Linguistics* 38(3): 479–526.

Dudau-Sofronie, Daniela, Isabelle Tellier, and Marc Tommasi. 2003. A learnable class of classical categorial grammars from typed examples. In *Proceedings of the 8th Conference on Formal Grammar*, pp. 77–88.

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

Gold, E. M. 1978. Complexity of automaton identification from given data. *Information and Control* 37(3): 302–320.

Harris, Z. S. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Heinz, Jeffrey. In press. Computational theories of learning and developmental psycholinguistics. In Jeffrey Lidz, William Snyder, and Joe Pater (eds.), *The Cambridge Handbook of Developmental Linguistics*. Cambridge: Cambridge University Press.

Horning, James Jay. 1969. A study of grammatical inference. PhD thesis, Stanford University.

Hsu, D., S. M. Kakade, and T. Zhang. 2012. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences* 78(5): 1460–1480.

Joshi, A. K., K. Vijay-Shanker, and D. J. Weir. 1991. The convergence of mildly context-sensitive grammar formalisms. In Peter Sells, Stuart Shieber, and Thomas Wasow (eds.), *Foundational Issues in Natural Language Processing*, pp. 31–81. Cambridge, MA: MIT Press.

Kearns, M. and L. Valiant. 1994. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM* 41(1): 67–95.

Keller, F. and A. Asudeh. 2002. Probabilistic learning algorithms and optimality theory. *Linguistic Inquiry* 33(2): 225–244.

Kobele, G. M. 2006. Generating copies: An investigation into structural identity in language and grammar. PhD thesis, University of California, Los Angeles.

Kwiatkowski, Tom, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1223–1233. Stroudsburg, PA: Association for Computational Linguistics.

Lee, Lillian. 1996. Learning of context-free languages: A survey of the literature. Technical Report TR-12-96, Harvard University.

Luque, F. and G. Infante-Lopez. 2010. PAC-learning unambiguous k, l-NTS languages. In *ICGI 2010 Grammatical Inference: Theoretical Results and Applications*, pp. 122–134. Berlin/Heidelberg: Springer.

MacWhinney, Brian. 1995. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum.

Mameli, M. and P. Bateson. 2006. Innateness and the sciences. *Biology and Philosophy* 21(2): 155–188.

Niyogi, P. 2006. *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: MIT Press.

Oncina, J. and P. Garcia. 1992. Inferring regular languages in polynomial update time. *Pattern Recognition and Image Analysis* 1: 49–61.

Pilato, Samuel F. and Robert C. Berwick. 1985. Reversible automata and induction of the English auxiliary system. In *ACL '85: Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, pp. 70–75. Stroudsburg, PA: Association for Computational Linguistics.

Pinker, Steven. 1994. *The Language Instinct*. London: Allen Lane.

Pinker, Steven. 1995. Language acquisition. In Daniel Osherson, Lila R. Gleitman, and Mark Liberman (eds.), *An Invitation to Cognitive Science. Volume 1: Language*, pp. 135–182. 2nd ed. Cambridge, MA: MIT Press.

Sakakibara, Y. 1992. Efficient learning of context-free grammars from positive structural samples. *Information and Computation* 97(1): 23–60.

Seki, H., T. Matsumura, M. Fujii, and T. Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science* 88(2): 229.

Shibata, Chihiro and Ryo Yoshinaka. 2013. PAC learning of some subclasses of context-free grammars with basic distributional properties from positive data. In *Proceedings of Algorithmic Learning Theory Conference*, pp. 143–157. Berlin/Heidelberg: Springer.

Siskind, Jeffrey Mark. 2000. Learning word-to-meaning mappings. In Peter Broeder and Jaap Murre (eds.), *Models of Language Acquisition*, pp. 121–153. Oxford: Oxford University Press.

Stabler, E. 1997. Derivational minimalism. In C. Retoré (ed.), *Logical Aspects of Computational Linguistics (LACL '96)*, pp. 68–95. Berlin: Springer.

Stabler, E. P. 2011. Computational perspectives on minimalism. In Cedric Boeckx (ed.), *The Oxford Handbook of Linguistic Minimalism*, pp. 617–641. Oxford: Oxford University Press.

Stabler, Edward P. 2013. The epicenter of linguistic behavior. In Montserrat Sanz, Itziar Laka, and Michael K. Tanenhaus (eds.), *Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures*, pp. 316–323. Oxford: Oxford University Press.

Starkie, B., F. Coste, and M. van Zaanen. 2004. The Omphalos context-free grammar learning competition. In G. Paliouros and Y. Sakakibara (eds.), *Proceedings of the International Colloquium on Grammatical Inference*, pp. 16–27. Berlin/Heidelberg: Springer.

Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

van Rooij, I. 2008. The tractable cognition thesis. *Cognitive Science: A Multidisciplinary Journal* 32(6): 939–984.

Wexler, Kenneth and Peter W. Culicover. 1980. *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

Yang, C. 2008. The great number crunch. *Journal of Linguistics* 44(1): 205–228.

Yoshinaka, R. 2008. Identification in the limit of *k*-*l*-substitutable context-free languages. In Alexander Clark, François Coste, and Laurent Miclet (eds.), *Grammatical Inference: Algorithms and Applications*, pp. 266–279. Berlin/Heidelberg: Springer.

Yoshinaka, R. and M. Kanazawa. 2011. Distributional learning of abstract categorial grammars. In Sylvain Pogodalla and Jean-Philippe Prost (eds.), *Logical Aspects of Computational Linguistics*, pp. 251–266. Berlin/Heidelberg: Springer.