Computational Theories of Learning and Developmental Psycholinguistics

Jeffrey Heinz

September 20, 2010

1 Introduction

A computer is something that computes, and since humans make computations when processing information, humans are computers. What kinds of computations do humans make when they learn languages?

Answering this question requires the collaborative efforts of researchers in several different disciplines and sub-disciplines, including language science (e.g. theoretical linguistics, psycholinguistics, language acquisition), computer science, psychology, and cognitive science. The primary purpose of this chapter is explain to developmental psycholinguists and language scientists more generally, the main conclusions and issues in computational learning theories. This chapter is needed because

1. the mathematical nature of the subject makes it largely inaccessible to those without the appropriate training (though hopefully this chapter shows that the amount of training required is less than what is standardly assumed)

2. to correct falsehoods which exist in the literature about the relevance of computational learning theories to the problem of language learning.

The main points in this chapter are:

1. The central problem of learning is generalization.

2. Consensus exists that, for feasible learning to occur at all, restricted, structured hypothesis spaces are necessary.

3. Debates about statistical learning vs. symbolic learning are misplaced. To the extent meaningful debate exists at all, it is about how learning ought to be defined; in particular, it is about what kinds of experience computers are required to succeed on in order to say that they have "learned" something.

4. Computational learning theorists and developmental psycholinguists can profitably interact in the design of meaningful artificial language learning experiments.

In order to understand how a computer can be said to learn something, a definition of learning is required. Only then does it become possible to ask whether the behavior of the computer meets the necessary and sufficient conditions of learning required by the definition. Computational learning theories provide definitions of what it means to learn and then asks, under those definitions: What can be learned, how and why? Which definition is "correct" of course is where most of the issues lie.

At the most general level, a language learner is something that comes to know a language on the basis of its experience. All computational learning theories consider learners to be functions which map experience to languages (Figure 1). Therefore, in order to define learning both languages and experience need to be defined first.

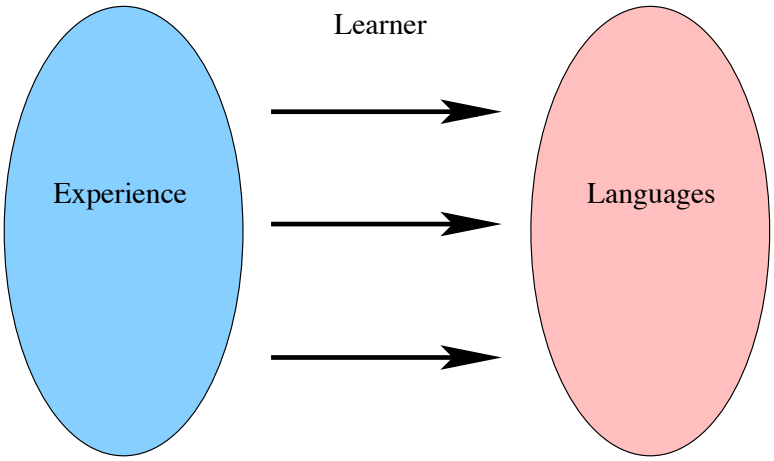

Figure 1: Learners are functions ϕ from experience to languages

2 Languages, grammars, and experience

2.1 Languages

Before we can speak of grammars, which are precise descriptions of languages, it will be useful to talk about languages themselves. In formal language theory, languages are mathematical objects which exist independently of any grammar. They are usually defined as subsets of all logically possible strings of finite length constructible from a given alphabet. This can be generalized to probability distributions over all those strings, in which case they are called stochastic languages.

The alphabet can be anything, so long as it is unchanging and finite. Elements of the alphabet can represent IPA symbols, phonological features, morphemes or words in the dictionary. If desired, the alphabet can also include structural information such as labeled phrasal boundaries. It follows that any description of sentences and words that language scientists employ can be described as a language or stochastic language with a finite alphabet.[1]

[1]Languages with infinite alphabets are also studied (Otto, 1985), but they will not be discussed in this chapter.

It useful to consider the functional characterizations of both languages and stochastic languages because they are the mathematical objects of interest to language scientists. As functions, a language L maps strings to one only if the string is in the language and all other logically possible strings are mapped to zero. Stochastic languages, as functions, map all logically possible strings to real values between zero and one such that they sum to one. Figure 2 illustrates functional characterizations of English as a language and as a stochastic language. The functional characterization of English as a language only makes

English as a language			English as a stochastic language		
John sang	$\rightarrow$	1	John sang	$\rightarrow$	1.2×10^{-12}
John and sang	$\rightarrow$	0	John and sang	$\rightarrow$	0
John sang and Mary danced	$\rightarrow$	1	John sang and Mary danced	$\rightarrow$	2.4×10^{-12}
...			...		

Figure 2: Functional characterizations of English as a language and a stochastic language.

binary distinctions between well-formed and ill-formed sentences. On the other hand, the functional characterization of English as a stochastic language makes multiple distinctions. In both cases, the characterizations are infinite in the sense that both assign nonzero values to infinitely many possible sentences. This is because there is no principled upper bound on the length of possible English sentences.[2]

How stochastic languages are to be interpreted ought to always be carefully articulated. For example, if the real numbers are intended to indicate probabilities then the functional characterization in Figure 2 says that "John sang" is twice as likely as "John and Mary sang" On the other hand, if the real numbers are supposed to indicate well-formedness, then the claim is that that "John sang" is twice as well-formed (or acceptable) as "John sang and Mary danced."[3]

As explained in the next section, from a computational perspective, the distinction between stochastic and non-stochastic languages is often unimportant. I use the word *pattern* to refer to both stochastic and non-stochastic languages in an intentionally ambiguous manner.

2.2 Grammars

Grammars are finite descriptions of patterns. It is natural to ask whether every conceivable pattern has a grammar. The answer is No. In fact most logically possible patterns cannot be described by *any* grammar at all of any kind. There is an analogue to real numbers. Real numbers are infinitely long and some are unpredictable in an important kind of way:

[2]If there were, then there would be a value n such that "John sang $\underbrace{\text{and John sang}}_{n-1 \text{ times}}$" would be well-formed but "John sang $\underbrace{\text{and John sang}}_{\text{n times}}$" would be ill-formed like "John and sang."

[3]There is a technical issue here. If there are infinitely many nonzero values, then it is not always the case that they can be normalized to yield a well-formed probability distribution. For example, if each sentence is equally acceptable, we would expect a uniform distribution. But the uniform distribution cannot be defined over infinitely many elements since the probability for each element goes to zero.

no algorithm exists (nor can ever exist) which can generate the real number correctly up to some arbitrary finite length; such reals are called uncomputable. In fact, uncomputable real numbers turn out to be the most common kind of real number (Turing, 1937)!

We assume that the functional characterizations of patterns (like English above) are computable (i.e. have grammars of some kind). Although there are many different kinds of grammatical formalisms, formal language theory provides a universal means for studying them on the basis of the kind of patterns they are able to describe. This is because the formal language theory studies properties of the languages and stochastic languages that are independent of the particular grammatical formalism used to describe them.[4]

The Chomksy Hierarchy classifies logically possible patterns into sets of nested regions. For further details regarding the Chomsky Hierachy, readers are referred to Harrison (1978); Hopcroft *et al.* (1979, 2001) and Thomas (1997). The Finite patterns are those whose func-

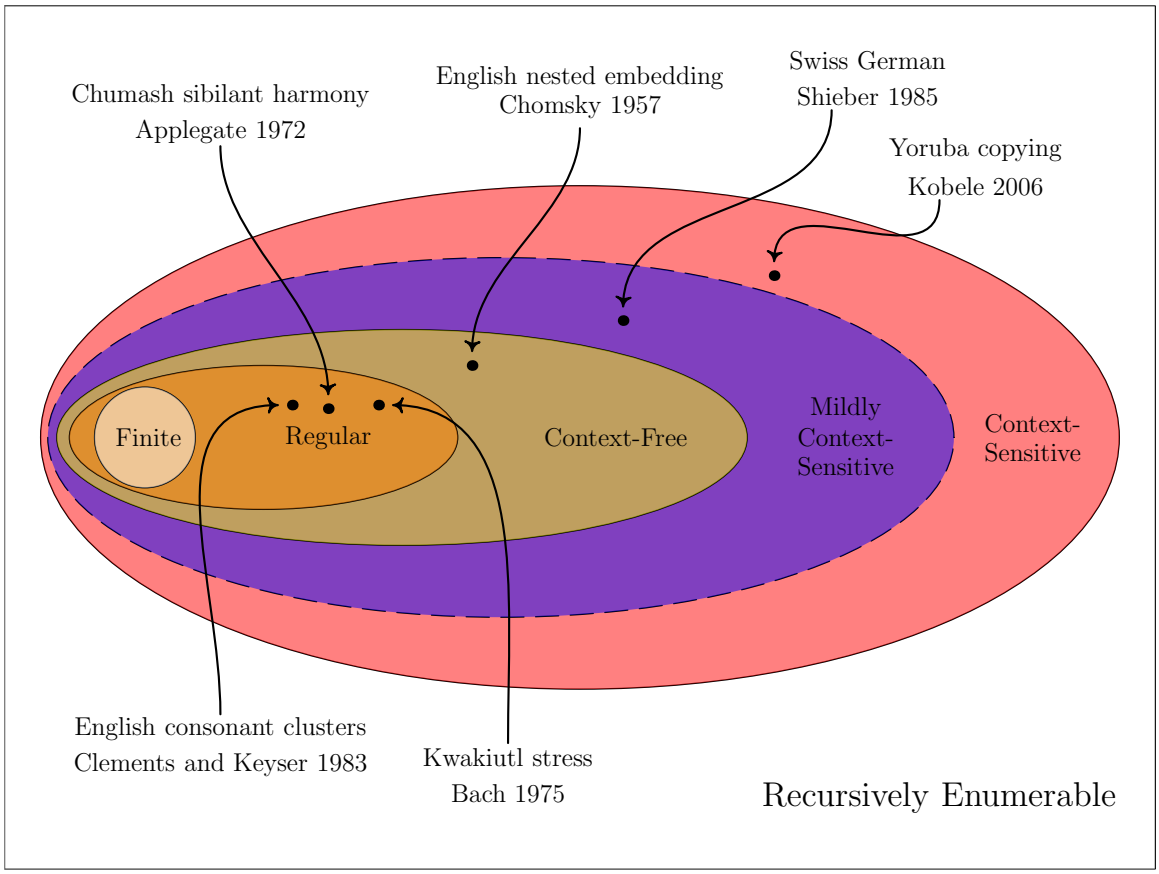

Figure 3: The Chomsky Hierarchy with natural language patterns indicated.

tional characterizations have only finitely many sentences with nonzero values. Regular languages are all those describable with finite state automata, and regular stochastic languages are those describable by probabilistic finite-state automata. In other words, regular

[4]In some cases, particular grammatical formalisms are especially good at illuminating important properties of the patterns themselves and so the grammatical formalism becomes strongly identified with the languages. This seems to be the case with (probabilistic) finite state automata and regular (stochastic) languages, for example.

patterns, whether stochastic or not, only require grammars that distinguish finitely many states. On the other hand, context-free patterns require grammars that distinguish infinitely many states. The class of recursively enumerable (r.e.) patterns are exactly those patterns describable with grammars; i.e. they are computable patterns.

It is of course of great interest to know what kinds of patterns natural languages are. Figure 3 shows where some natural language patterns fall in the Chomsky Hierarchy. For example, phonological patterns do not appear to require grammars that distinguish infinitely many states[5], unlike some syntactic patterns, which appear to require grammars which make infinitely many distinctions.

It is also important to understand the goals of computational research on natural language patterns. In particular establishing complexity bounds is different from the hypotheses which state both neccessary *and sufficient* properties of possible natural language patterns. For example the hypothesis that natural language patterns are mildy context-sensitive Joshi (1985), is a hypothesis that seeks to establish an upper bound on the complexity of natural language. Joshi is not claiming, as far as I know, that *any* mildly context-sensitive pattern is a possible natural language one. In my opinion, it is much more likely that possible natural language patterns belong to subclasses of the major regions of the Chomsky Hierarchy. For example, Heinz (2010a) hypothesizes that all phonological segmental phonotactic patterns belong to particular subregular classes and Clark and Eyraud (2007) present a learnable subclass of the context-free languages which describes syntactic patterns like nested embedding.

Although from the perspective of formal language theory, grammars are the mathematical objects of secondary interest, it does matter that learners return a grammar, instead of a language. This is for the simple reason that as mathematical objects, grammars are of finite length and the functional characterizations of patterns are infinitely long. Thus while Figure 1 describes learners as functions from experience to languages, they are more accurately described as functions from experience to grammars (Figure 4).

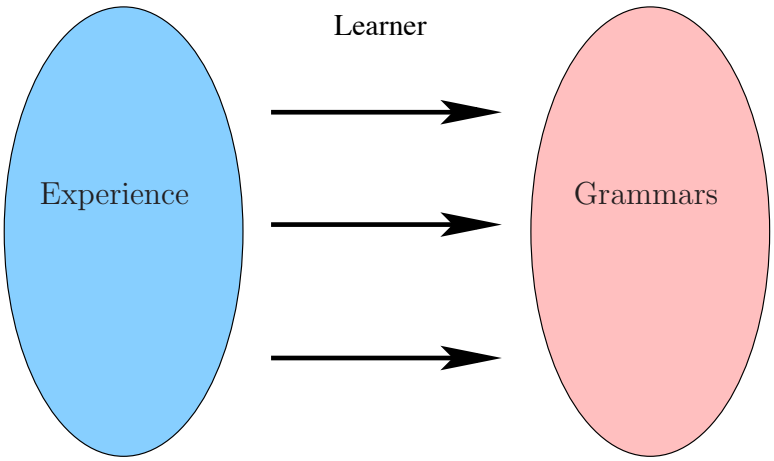

Figure 4: Learners are functions ϕ from experience to grammars.

[5]For more on the hypothesis that all phonological patterns are regular see Kaplan and Kay (1994); Eisner (1997) and Karttunen (1998).

2.3 Experience

There are many different kinds of experience learning theorists consider, but they agree that the experience is a finite sequence (Figure 5). It is necessary to decide what the elements

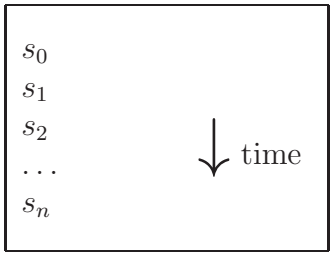

Figure 5: The learner's experience

s_i of the sequence are. In this chapter we distinguish four kinds of experience. *Positive evidence* refers to experience where each s_i is known to be a non-zero-valued sentence of the target pattern. *Positive and negative evidence* refers to experience where each s_i is given as belonging to the target pattern (has a nonzero value) or as not belonging (has a zero value). *Noisy evidence* refers to the fact that some of the experience is incorrect. For example, perhaps the learner has the experience that some s_i belongs to the target language, when in fact it does not (perhaps the learner heard a foreign sentence or someone misspoke). *Queried evidence* refers to experience learners may have because they specifically asked for it. In principle, there are many different kinds of queries learners could make. This chapter does not address these last two kinds; readers are referred to Angluin and Laird (1988) and Kearns and Li (1993) for noisy evidence, and to Angluin (1988a, 1990); Becerra-Bonache *et al.* (2006) and Tîrnauca (2008) for queries.

2.4 Learners as functions

Armed with the basic concepts and vocabulary all learning theorists use to describe target languages, grammars, and experience, it is now possible to define learners. They are simply functions that map experience to grammars. This is a very precise characterization of learners, but it is also very broad. Any learning procedure can be thought of as a function from experience to grammars, including connectionist ones (e.g. Rumelhart and McClelland (1986)), Bayesian ones (Griffiths *et al.*, 2008), learners based on maximum entropy (e.g. Goldwater and Johnson (2003)), as well as those embedded within generative models (Wexler and Culicover, 1980; Berwick1985, 1985; Niyogi and Berwick, 1996; Tesar and Smolensky, 1998; Niyogi, 2006). Each of these learning models, and I would suggest any learning model, takes as its input a finite sequence of experience and outputs some grammar, which defines a language or a stochastic language. Consequently, all of these particular proposals are subject to the results of formal learning theory.

3 What is learning?

It remains to be defined what it means for a function which maps experiences to grammars to be successful. After all, there are many logically possible such functions, but we are interested in evaluating particular learning proposals. For example, we may be interested in those learning functions that are humanlike, or which return humanlike grammars.

This chapter draws on a large set of learning literature. Readers are referred to Nowak *et al.* (2002) for an excellent, short introduction to computational learning theory. Niyogi (2006) and de la Higuera (2010) provide detailed, accessible treatments, and Jain *et al.* (1999), Zeugmann and Zilles (2008) Lange *et al.* (2008) Anthony and Biggs (1992), and Kearns and Vazirani (1994b) provide technical introductions. I have also relied on the following research: (Gold, 1967; Horning, 1969; Angluin, 1980; Osherson *et al.*, 1986; Angluin and Laird, 1988; Vapnik, 1995, 1998).

3.1 Learning Criteria

In addition to deciding what kinds of experience learners must necessarily succeed on, learning theorists must also define what counts as success. The general idea in the learning theory literature is that learning has been successful is the learner has *converged* to the right language. Is there some point n after which the learner's hypothesis doesn't change (much)? Convergence can be defined in different ways, to which we return below. Typically, learning theorists conceive of an, infinite stream of experience to which the learner is exposed so that it makes sense to talk about a convergence point. Is there a point n such that for all $m \geq n$, Grammar $G_m \simeq G_n$ (given some definition of $\simeq$)? Figure 6 illustrates.

datum	Learner's Hypothesis
s_0	$\phi(\langle s_0 \rangle) = G_0$
s_1	$\phi(\langle s_0, s_1 \rangle) = G_1$
s_2	$\phi(\langle s_0, s_1, s_2 \rangle) = G_2$
$\ldots$	
s_n	$\phi(\langle s_0, s_1, s_2, \ldots, s_n \rangle) = G_n$
$\ldots$	
s_m	$\phi(\langle s_0, s_1, s_2, \ldots, s_m \rangle) = G_m$

$\downarrow$ time

Figure 6: If, for all $m \geq n$, it is the case that $G_m \simeq G_n$ (given some definition of $\simeq$), then the learner is said to have converged.

The infinite streams of experience are also called *texts* (Gold 1967) and *presentations of data* (Angluin 1988). All three terms are used synonymously here.

Defining successful learning as convergence to the right language after some point n, raises another question with respect to experience: Which infinite streams require convergence? Generally two kinds of requirements have been studied. Some infinite streams are *full*; that is, every possible kind of information about the target language occurs at some point in the presentation of the data. For example, in the case of positive evidence, each sentence in the language would occur at some finite point in the stream of experience.

Table 1: Choices which make the learning problem easier and harder.

	Makes learning easier		Makes learning harder
a.	positive and negative evidence	A.	positive evidence only
b.	noiseless evidence	B.	noisy evidence
c.	queries permitted	C.	queries not permitted
d.	approximate convergence	D.	exact convergence
e.	full infinite streams	E.	any infinite sequence
f.	computable infinite streams	F.	any infinite sequence

Table 2: Choices providing a coarse classification of learning frameworks.

The second one is about computability, for which there are two concerns. First, there are as many infinite texts as there are real numbers and so most of these sequences are not computable. Should learners be required to succeed on these? Also, there are infinite sets of elements such that each element is computable, but the set itself is not. This is because there can be no program that is able to generate all (the programs for) the elements of the set itself.[6] So another strand of research limits the infinite sequences the learners are required to succeed on to *computable* data presentations. This means not only is the set of infinite sequences under consideration computable, but so is each element within the set as well.

3.2 Definitions of Learning

Table 2 summarizes the kinds of choices to be made when deciding what learning means.

The division of the choices into columns labeled "Makes learning easier" and "Makes learning harder" ought to be obvious. Learners only exposed to positive evidence have more work to do than those given both positive and negative evidence. Similarly, learners who have to work with noisy evidence will have a more difficult task than those given noise-free evidence. Learners allowed to make queries have access to more information than those not permitted to make queries. Exact convergence is a very strict demand, and approximate convergence is less so. Finally, requiring learners to succeed for every logically possible presentation of the data makes learning harder than requiring learners only to succeed for full or computable presentations simply because there are far fewer full and/or computable presentations of the data. Figure 7 shows the proper subset relationships among full and computable presentations of data.

Using the coarse classification provided by Table 2, I now classify several definitions of learning (these are summarized in Table 3 on page 14). For example, the learning paradigm known as *identification in the limit from positive data* (Gold, 1967) requires that the learner succeed with positive evidence only (A), noiseless evidence (b), and without queries (C).

[6] As an example, consider the halting problem. This problem takes as input a program p and an input i for p, and asks whether p will run forever on i, or eventually halt. It is known that there are infinitely many programs which do not halt on some inputs. For each such program p choose some input i_p. Since i_p is an input, it is finitely long and can be generated by some program. But the class of all such i_ps cannot be generated by any program. If it could, it follows there is a solution to the halting problem, but in fact, no such solution exists (or can exist) because the halting problem is known to be uncomputable; that is, no algorithm exists which solves it (Turing, 1937).

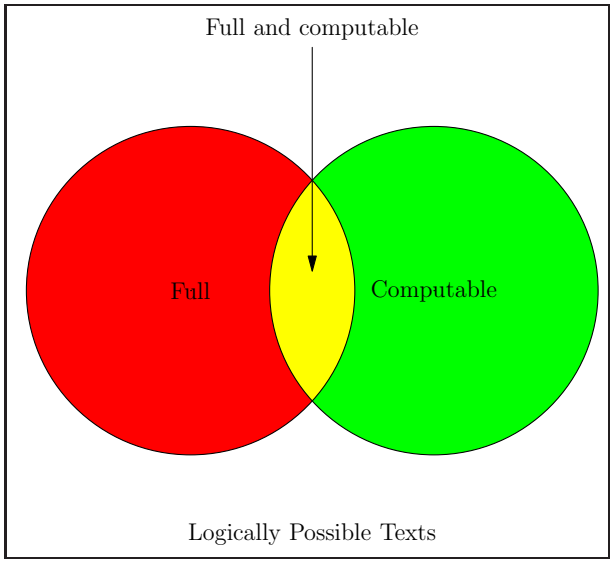

Figure 7: Proper subset relationships between all logically possible classes of texts, classes of full texts, computable classes of texts, and both full and computable classes of texts.

Exact convergence (D) is necessary: even if the grammar to which the learner converges generates a language which differs only in one sentence from the target language, this is counted as a failure. This framework is generous in that learners are only required to succeed on full infinite sequences (e) but must succeed for any such sequence, not just computable ones (F).

The learning paradigm known as *identification in the limit from positive and negative data* (Gold, 1967) is the same except the learner is exposed to both positive and negative evidence (a).

Pitt (1985) and Wiehagen *et al.* (1984) study a learning paradigm I'll call *Identification in the limit from positive data with probability p* where learners are probabilistic (have access to random information) and convergence is defined in terms of the probability that the learners identify the target language in the limit given any infinite sequence of experience. Angluin (1988b) considers a variant of Pitt's framework which I'll call *identification in the limit from distribution-free positive stochastic input with probability p* where the data presentations are generated probabilistically from fixed, but arbitrary probability distributions (including uncomputable ones). The term *distribution-free* refers to the fact that the distribution generating the data presentation is completely arbitrary. Thus these frameworks are similar to identification in the limit from positive data but make an easier choice with respect to convergence (d).

Gold (1967) also considered a paradigm I'm calling *identification in the limit from positive r.e. texts* which is the similar to identification in the limit from positive data except that the learner is only required to succeed on full, computable streams (f), and not any stream.

Horning (1969); Osherson *et al.* (1986) and Angluin (1988b) study learning stochastic languages from positive data, and similarly restrict the presentations of data for which the learner is required to succeed on. In this chapter I discuss Angluin's framework since she generalizes the results of earlier researchers to obtain the strongest result. In this framework,

which I'll call *Identification in the limit from approximately computable positive stochastic input"* the presentations of data must be generated according to fixed "approximately computable" probability distributions. In this way, this definition of learning is like identification in the limit from positive data from r.e. texts because learners need only converge on full and computable presentations of the data in order to have successfully learned.[7]

Finally, Probably Approximately Correct (PAC) learning (Valiant, 1984; Anthony and Biggs, 1992; Kearns and Vazirani, 1994a) makes a number of different assumptions. Both positive and negative evidence are permitted (a). Noise and queries are not permitted (b,c). Convergence need only be approximate (d), but the learner must succeed for any kind of data presentation, both non-full and uncomputable (E,F). What counts as convergence is tied to the degree of "non-fullness" of the data presentation.

3.3 Classes of languages

Finally, computational learning theories are concerned with learners of *classes of languages* and not just single languages. This is primarily because every language can be learned by a constant function (Figure 8). For example, with any of definitions above, it is easy to

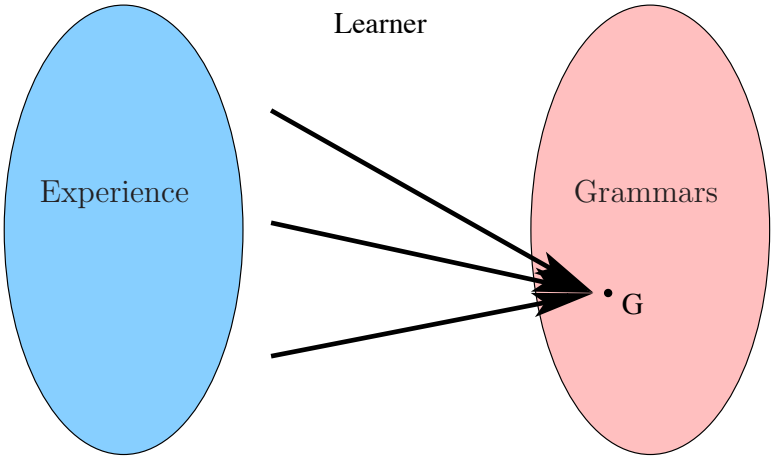

Figure 8: Learners which are constant functions map all possible experience to a single grammar.

state a learner for English (and just English). Just map all experience (no matter what it is) to a grammar for English. Even if we don't know what this grammar is yet, the learning problem is "solved" once we know it. Obviously, such "solutions" to the learning problem are meaningless, even if mathematically correct.

For this reason, computational learning theories ask whether a collection of more than one language can be learned by the same learner. This more meaningfully captures the kinds of question language scientists are interested in: Is there a single procedure that not only learns English, but also Spanish, Arabic, Inuktitut, and so on?

[7]If a data presentation is being generated from a computable stochastic language, then it is also full. This is because for any sentence with nonzero probability, the probability of this sentence occurring increases monotonically to one as the size of the experience grows. For example, we expect that the unlikely side of a biased coin will appear if it is to be flipped enough times.

4 Results of computational learning theories

Computational learning theorists have proven, given the above definitions, classes of languages that can and cannot be learned. Generally, formal learning theorists are interested in large classes of learnable languages because they want to see what is possible in principle. If classes of languages are learnable, the next important question is whether they are *feasibly* learnable. This means whether learners can succeed with reasonable amounts of time and effort where reasonable is defined in standard ways according to computational complexity theory (Garey and Johnson, 1979; Papadimitriou, 1994).

This section provides the largest classes known to be provably learnable under the different definitions of learning above. Where possible, I also indicate whether such classes can be feasibly learned.

4.1 No feasible learners for major regions of the Chomsky Hierarchy

Gold (1967) proved three important results. First, a learner exists which identifies the the class of r.e. languages in the limit from positive and negative data. Second, a learner exists which identifies the finite languages in the limit from positive data, but no learner exists which can identify any superfinite class in the limit from positive data. Superfinite classes of languages are those that include all finite languages and at least one infinite language. It follows from this result that none of the major regions of the Chomsky Hierarchy are identifiable in the limit from positive data by *any* learner which can be defined as mapping experience to grammars. Again, this includes Bayesian learners, connectionist models, etc. It is this result with which Gold's paper has become identified. Gold's third (and usually overlooked) result is that if learning is defined so that learners need only succeed given full presentations which (1) contain only positive evidence and (2) are computable[8], then a learner exists which can learn any recursively enumerable language (identification in the limit from positive r.e. texts).

Angluin (1988b), developing work begun in Horning (1969) and extended by Osherson *et al.* (1986) presents a similar result to Gold's third result for stochastic languages. She shows that if the learning criteria is that learners are only required to succeed for presentations of the positive data generable by those stochastic languages, then the class of r.e. stochastic languages is learnable (because such presentations of data are computable since r.e. stochastic languages are computable).[9]

This result contrasts sharply with other frameworks that investigate the power of probabilistic learners. Pitt (1985) and Wiehagen *et al.* (1984) show that the class of languages identifiable in the limit from positive data with probability p is the same as the class of languages identifiable in the limit from positive data whenever $p > 2/3$. Angluin (1988) concludes "These results show that if the probability of indentification is required to be above some threshold, randomization is no advanatge..." Angluin also shows that for all

[8]Gold referred to these texts as those generable by a partially recursive functions.

[9]Technically, she shows that the class of *approximately computable* distributions is learnable. The crucial feature of this class is that its elements are enumerable and computable, which is why I take some liberty in calling them r.e. stochastic languages.

p, the class of languages identifiable in the limit from positive data with probability p from distribution-free stochastic input is exactly the same as the the class of languages identifiable in the limit from positive data with probability p. Angluin observes that the "assumption of positive stochastic rather than positive [data presentations] is no help, if we require convergence with any probability greater than 2/3." She concludes "the results show that if no assumption is made about the probability distribution [generating the data presentations], stochastic input gives no greater power than the ability to flip coins."

Finally, in the PAC learning framework (Valiant, 1984), not even the finite class of languages is learnable (Blumer *et al.*, 1989).

In the cases where learners are known to exist in principle, we may also examine their feasibility. In the case of the identification in the limit from positive and negative data, (Gold, 1978) shows that there are no feasible learners for even the regular class of languages. In other words, while learners exist in principle for the r.e. class, they consume too much time and resources in the worst-cases. In the case of identification in the limit from r.e. texts and identification in the limit from approximately computable positive stochastic input, the learners known to exist in principle are also not feasible.[10]

Table 3 summarizes the results discussed in this section.

4.2 Other results

The previous section appears to paint a dismal picture—either large regions of the Chomsky Hierarchy are not learnable even in principle, or if they are, they are not feasibly learnable.

However there are many feasible learners for classes of language even in the frameworks with the most demanding criteria, such as identification in the limit from positive data and PAC-learning. This rich literature includes Angluin (1980, 1982); Muggleton (1990); Garcia *et al.* (1990); García and Ruiz (1996); Fernau (2003); García and Ruiz (2004); Clark and Thollard (2004); Oates *et al.* (2006); Clark and Eyraud (2007); Becerra-Bonache *et al.* (2008); Heinz (2008, 2009); Yoshinaka (2008, 2009); Heinz (2010b) and Heinz (2010a). The language classes discussed in those papers are not major regions of the Chomsky Hierarchy, but are *subclasses* of such regions.

Some of these languages classes are of infinite size and include infinite languages—but they crucially exclude some finite languages so they are not superfinite language classes. I return to this point below when discussing why the fundamental problem of learning is generalization.

Also, the proofs that these classes are learnable are constructive so concrete learning algorithms whose behavior is understood exist. The algorithms are successful because they are "aware" of the structure in the class, or equivalently, of its defining properties. They utilize such structure, or the properties of the language class, to generalize correctly. Often the proofs of the algorithm's success involve characterizing the kind of finite experience learners need to in order to generalize correctly.

To sum up, even though identification in the limit from positive data and PAC-learning make the learning problem harder by requiring learners to succeed for any data presentation,

[10]These learners essentially compute an ordered list of all r.e. languages (or r.e. stochastic languages). They find the first grammar in this list compatible with the experience so far.

Definition of Learning	Feasible learnability of the major regions of the Chomsky Hierarchy
Identification in the limit from positive data (Gold 1967)	Finite languages are learnable but no superfinite class of languages is learnable (e.g. regular, context-free, context-sensitive, r.e.).
Identification in the limit from positive data with probability p (Pitt, 1985; Wiehagen *et al.*, 1984)	For all $p > 2/3$, classes identifiable in the limit from positive data with probability p are the same as those identifiable in the limit from positive data.
Identification in the limit from distribution-free positive stochastic input with probability p (Angluin, 1988b)	For all p, classes identifiable in the limit from distribution-free positive stochastic input with probability p are the same as classes identifiable in the limit from positive data with probability p.
Identification in the limit from positive and negative data (Gold 1967)	R.e. languages are learnable but the regular languages are not feasibly learnable (Gold 1978).
Identification in the limit from positive r.e. texts (Gold 1967)	R.e. languages are learnable but they are not not feasibly learnable.
Identification in the limit from approximately computable positive stochastic input (Angluin, 1988b) (supersedes Horning 1969, Osherson et al. 1986)	R.e stochastic languages are learnable but they are not feasibly learnable.
Probably Approximately Correct (PAC) (Valiant 1984)	Finite languages not learnable (Blumer *et al.*, 1989), and hence neither are the regular, context-free, context-sensitive, and r.e. languages.

Table 3: Foundational results in computational learning theory

so that no learners exist, even in principle, for superfinite classes of languages, there are feasibly learnable language classes in these frameworks. Furthermore, many of the above researchers have been keen to point out the natural language patterns which belong to these learnable subclasses.

5 Interpreting results of computational learning theories

5.1 Wrong reactions

How have the above results been interpreted? Are those interpretations justified? Perhaps the most widespread myth about formal learning theory is the oft-repeated claim that Gold (1967) is irrelevant because

1. Horning (1969) showed that statistical learning is more powerful than symbolic learning.

2. unrealistic assumptions are made

As shown below, these claims have been made by influential researchers in cognitive science, computer science, computational linguistics, and psychology.

In this section I rebut these charges. The authors cited below repeatedly fail to distinguish different definitions of learnability, fail to identify Gold (1967) with anything other than identification in the limit from positive data, and/or false statements about the kinds of learning procedures Gold (1967) considers. With respect to the claim that identification in the limit makes unrealistic assumptions, I beleive it is fair to debate the assumptions underlying any learning framework. However, the arguments put forward by the authors below are not convincing, usually because they say they very little about what the problematic assumptions are and how their proposed framework overcomes them without introducing unrealistic assumptions of their own.

Consider how Horning is used to downplay Gold's work. For example, Abney writes

> ... though Gold showed that the class of context free grammars is not learnable, Horning showed that the class of stochastic context free grammars is learnable.

The first clause only makes sense if, by "Gold", Abney is referring to identification in the limit from positive data. After all, Gold did show that the context-free languages are learnable not only from positive and negative data, but also from positive data only if the learners are only required to succeed on computable data presentations (identification in the limit from positive r.e. texts).

As for the second clause, Abney leaves it to the reader to infer that Gold and Horning are studying different definitions of learnability. Abney emphasizes the stochastic nature of Horning's target grammars as if that is the key difference in their results, but it should be clear from Sections 4 and Table 3 that the gain in learnability is not coming from the stochastic nature of the target patterns, but rather from the fact that, in Horning's framework, learners are only required to succeed for full and computable data presentations.

Again, Gold (1967) showed this result too, in a non-stochastic setting (identification in the limit from positive r.e. texts).

Similarly, in the introductory 1999 text to computational linguistics Manning and Schütze (1999, 386-387) write

> . Gold (1967) showed that CFGs [context-free grammars] cannot be learned (in the sense of identification in the limit that is whether one can identify a grammar if one is allowed to see as much data produced by the grammar as one wants) without the use of negative evidence (the provision of ungrammatical examples). But PCFGs [probabilistic context-free grammars] can be learned from positive data alone (Horning 1969). (However, doing grammar induction from scratch is still a difficult, largely unsolved problem, and hence much emphasis has been placed on learning from bracketed corpora. . .)

Like Abney, Manning and Schütze do not mention Gold's third result that CFGs can be learned if the data presentations are limited to computable ones. To their credit, they acknowledge the hard problem of learning PCFGs despite Horning's (and later Angluin's) results. Horning's and Angluin's learners are completely impractical and are unlikely to be the basis for any feasible learning strategy for PCFGs. So these positive learning results offer no little insight on how PCFGs which describe natural language patterns may actually be induced from the kinds of corpus data that Manning and Schütze have in mind.

Klein (2005, 4-5) also writes:

> . . . Golds formalization is open to a wide array of objections. First, as mentioned above, who knows whether all children in a linguistic community actually do learn the same language? All we really know is that their languages are similar enough to enable normal communication. Second, for families of probabilistic languages, why not assume that the examples are sampled according to the target languages distribution? Then, while a very large corpus wont contain every sentence in the language, it can be expected to contain the common ones. Indeed, while the family of context-free grammars is unlearnable in the Gold sense, Horning (1969) shows that a slightly softer form of identification is possible for the family of probabilistic context-free grammars if these two constraints are relaxed (and a strong assumption about priors over grammars is made).

Again, by "Gold's formalization", Klein must be referring to identification in the limit from positive data. Klein's first point is that it is unrealistic to use exact convergence as a requirement because we do not know if children in communities all learn exactly the same language, and it is much more plausible that they learn languages that are highly similar, but different in some details. Hopefully by now it is clear that Klein is misplacing the reason why it is impossible to identify in the limit from positive data superfinite classes of languages. It is not because of exact convergence, it is because learners are required to succeed for any full presentation of the data, not just the computable ones. In frameworks that allow looser definitions of convergence (PAC-learning, identification in the limit from positive data with probability p), the main results are more or less the same as in the identifcation in the limit from positive data. The real reason for Horning's success is made

clear in Angluin (1988): identification in the limit from approximately computable positive stochastic data only requires learners to succeed for data presentations that are computable. As for the choice of exactness as the definition of convergence being unrealistic, it is a useful abstraction that lets one ignore the variation that exists in reality to concentrate on the core properties of natural language that make learning possible.

Klein then claims that it is much more reasonable to assume that the data presentations are generated by a fixed unchanging probability distribution defined by the target PCFG. This idealization may lead to fruitful research, but it is hard to accept it as realistic. That would mean that for each of us, in our lives, every sentence we have heard up until this point, and will hear until we die, is being generated by a fixed unchanging probability distribution. It is hard to see how this could be true given that what is actually said is determined by so many non-linguistic factors.[11] So if realism is one basis for the "wide array of objections" that Klein mentions, the alternative proposed does not look much better.

Johnson and Reizeler (2002) are much more cautious in their rhetoric:

> Turning to theoretical results on learning, it seems that statistical learners may be more powerful than non-statistical learners. For example, while Golds famous results showed that neither finite state nor context-free languages can be learnt from positive examples alone (Gold, 1967), it turns out that probabilistic context-free languages can be learnt from positive examples alone (Horning, 1969).[1]

Note the judicious use of "may" in the first sentence. Although the second sentence fails to ackowledge Gold's result of learning the class of r.e. languages from positive r.e. texts, their footnote addresses the different learning definitions in identification in the limit from positive data and in what is essentially identification in the limit from approximately computable positive stochastic input (Angluin 1988).

> That is, a class of probabilistic languages may be statistically learnable even though its categorical counterpart is not. Informally this is because the statistical learning framework makes stronger assumptions about the training data (i.e., that it is distributed according to some probabilistic grammar from the class) and accepts a weaker criterion for successful learning (convergence in probability).

However, there is no discussion whether the stronger assumptions about the trainign data is warranted.

Like Klein above, Bates and Elman (1996) also argue that Gold (1967) is irrelevant because of unrealistic assumptions. They write:

> A formal proof by Gold [1967] appeared to support this assumption, although Gold's theorem is relevant only if we make assumptions about the nature of the learning device that are wildly unlike the conditions that hold in any known nervous system [Elman et al. 1996].

[11] Even if we abstract away from actual words and ask whether if strings of lingusitic categories are generated by a fixed underlying PCFGS, the claim is probably false. Imperative structures often have different distributions of categories than declaratives than questions and the extent to which these are used in discourse depends entirely on non-linguistic factors in the real world.

By now we are familiar with authors identifying Gold (1967) solely with identification in the limit from positive data. What assumptions does Gold make that are "wildly unlike the conditions that hold in any known nervous system?" Gold only assumes that learners are functions from finite sequences of experience to grammars. It is not clear to me why this assumption not applicable to nervous systems, or any other computer. Perhaps Bates and Elman are taking issues with exact convergence, but as mentioned above, learning frameworks that allow looser definitions of convergence do not change the main results and even Elman *et al.* (1996) uses abstract models.

Perfors *et al.* (2010), while describing an approach to language learning that balances preferences for simple grammars with good fits to the data, writes

> Traditional approaches to formal language theory and learnability are unhelpful because they presume that a learner does not take either simplicity or degree of fit into account (Gold 1967). A Bayesian approach, by contrast, provides an intuitive and principled way to calculate the tradeoff... Indeed it has been formally proven that an ideal learner incorporating a simplicity metric will be able to predict the sentences of the language with an error of zero as the size of the corpus goes to infinity (Solomonoff 1978, Chater and Vitànyi 2007); in many more traditional approaches, the correct grammar cannot be learned even when the number of sentences is infinite (Gold 1967). However learning a grammar (in a probabilistic sense) is possible, given reasonable sampling assumptions, if the learner is sensitive to the statistics of language (Horning 1969).

The first sentence is simply false. While it is true that Gold does not specifically refer to learners which take either simplicity or degree of fit into account, that in no way implies his results do not apply to such learners. Gold's results apply to any algorithms that can be said to map finite sequences of experience to grammars, and the Bayesian models Perfors et al. propose are such algorithms. The fact that Gold doesn't specifically mention these particular traits emphasizes how general and powerful Gold's results are. If Perfors et al. really believe Bayesian learners can identify a superfinite class of languages in the limit from positive data, they should go ahead and try to prove it. (Unfortunately for them, Gold's proof is correct so we already know it is useless trying.)

The last sentence is what we come to expect from authors who are peddling statistical learning models. The statment that learners that are "sensitive to the statistics of language" can learn probabilistic grammars is attributed to Horning with no substantial discussion of the real issues. Readers are left believing in the power of statistical learning, even though the real issue is whether learning is defined in a way as to require learners to succeed on full and computable data presentations versus all full data presentations. Again Gold showed that any r.e. language can be learned from positive r.e. texts. Angluin (1988) showed that learners that are "sensitive to the statistics of language" are not suddenly more powerful (Identification in the limit from distribution-free positive stochastic input with probability p). Finally Perfors et al. hide the real issue behind the phrase "under reasonable sampling conditions." In the above discussion of Klein's comments, I think there is every reason to question how reasonable those assumptions are. But I would be happy if the debate could at least get away from "the sensitivity of the learner to statistics" rhetoric to whether the assumption that actual data presentations are generated according to fixed unchanging

computable probability distributions is reasonable. That would be progress and would reflect one actual lesson from computational learning theory.

As for Chater and Vitányi (2007), which extends work begun in Solomonoff (1978), they provide a more accurate, substantial and overall fairer portrayal of Gold's (1967) paper than these others, and corroborate some of the points made in this chapter. However, a couple of inaccuracies remain. Consider the following passage:

> Gold (1967) notes that the demand that language can be learned from every text may be too strong. That is, he allows the possibility that language learning from positive evidence may be possible precisely because there are restrictions on which texts are possible. As we have noted, when texts are restricted severely, e.g., they are independent, identical samples from a probability distribution over sentences, positive results become provable (e.g., Pitt, 1989); but the present framework does not require such restrictive assumptions. (p. 153)

This quote is misleading in a couple of ways. First, Gold (1967) goes much farther than just suggesting learning from positive evidence alone may be possible if the texts are restricted; in fact, he shows this (identification in the limit from positive r.e. texts). Secondly, the claim that their framework does not assume that the stream of experience which is the input to the "ideal language learner" is not generated by inedpendently drawn samples from a computable distribution is false.[12] Section 2.1 of their paper explains exactly how the input to the learner is generated. They explain very clearly that they add probabilities to a Turing machine, much in the same way probabilities can be added to any automaton. In this case, the consequence is they are able to describe r.e. stochastic languages. In fact they conclude this section with the following sentence:

> The fundamental assumption concerning the nature of the linguistic input outlined in this subsection can be summarized as the assumption that the linguistic input is generated by some monotone computable probability distribution $\mu_C(x)$. (p. 138)

It is unclear how this sentence and the claim on p. 153 that they make no restrictive assumptions on the text can both be true. In fact, one important lesson from computational learning theory that this chapter is trying to get across is that assuming that the data presentations (the linguistic input in Chater and Vitányi's terms) are drawn from some computable class of presentations is *the primary factor* in determining whether all patterns can be learned in principle or whether only very restricted classes can be. It certainly looks from their discussion of their learning framework, that successful learning is defined as success only on data presentations that are computable.

In fact, the "ideal language learner" does not appear to differ substantially from the one discussed by Angluin (1988) in the framework *identification in the limit from approximately computable positive stochastic input*. I conjecture that Chater and Vitányi's learner is a particular instantiation of the Angluin's learner in Theorem 17. Their ideal langauge learner

[12]The reference to Pitt 1989 is also odd given that this paper does not actually provide the positive results the authors suggest as it discusses identification in the limit from positive data with probability p. Horning 1969, Osherson, Stob, and Weinstein, or Angluin 1988 are much more appropriate references here.

essentially enumerates all possible r.e. stochastic languages by the length of their description. It then finds the first r.e. stochastic language in this list (i.e. a shortest grammar) consistent with the data. The only difference between this learner and Angluin's is that they specify a natural way to do the enumeration, whereas Angluin is agnostic regarding the details of the enumeration (because her proof only requires the existence of some enumeration).

Also, there is no indication whatsoever that the "ideal language learner" is feasible. In fact, the enumeration of all stochastic r.e. languages points to the contrary. Chater and Vitányi discuss the feasibility of the learner towards the end of their paper, where they point to a "crucial set of open questions" regarding "how rapidly learners can converge well enough" with the kinds of data in a child's linguistic environment. Of course, if my conjecture is correct, their idealized language learner cannot feasibly learn the r.e. stochastic languages, even if it can do so in principle (like Angluin's 1988 learner). Of course it may be that there is some subclass of the r.e. stochastic languages that the algorithm is able to learn feasibly, and which may include natural language patterns. In my view, research in this direction would be a positive development.

Finally, it is worth emphasizing that frameworks which require learners to succeed only on full and computable data presentations are weaker than frameworks which require learners to succeed on all full data presentations, computable and uncomputable, for the simple reason that there are more data presentations of the latter type (Figure 7). Learners successful in these more difficult frameworks (mentioned in Section 4.2) are more robust in the sense that they are guaranteed to succeed even when the data presentations to which they are exposed are not computable. The fact that there are feasible learners which can learn interesting classes of languages under strong definitions of learning (e.g. PAC-learnable classes) underscores how powerful those learning results are.

5.2 Correct reactions

Gold (1967:453-454) provides three ways to interpret his three main results:

> 1. The class of natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally... In particular the results on [identification in the limit from positive data] imply the following: The class of possible natural languages, if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality.
> 2. The child receives negative instances by being corrected in a way that we do not recognize...
> 3. There is an a priori restriction on the class of texts [presentations of data; i.e. infinite sequences of experience] which can occur...

The first possibility follows directly from the fact that no superfinite class of languages is identifiable in the limit from positive data. The second and third possibilities follow from Gold's other results on identification in the limit from positive and negative data and on identification in the limit from positive r.e. texts.

Each of these research directions can be fruitful, if honestly pursued. For the case of language acquisition, Gold's three suggestions can be investigated empirically. We ought to ask

1. What evidence exists that possible natural language patterns form subclasses of major regions of the Chomsky Hierarchy?

2. What evidence exists that children receive positive and negative evidence in some, perhaps implicit, form?

3. What evidence exists that each stream of experience each child is exposed to is guaranteed to be generated by a fixed, computable process (i.e. computable probability distribution or primitive recursion function)? I.e. what evidence exists that the data presentations are a priori limited?

My contention is that we have plenty of evidence with respect to question (1), some evidence with respect to (2), and virtually no evidence with respect to (3).

Consider question (1). Although theoretical linguists and language typologists repeatedly observe an amazing amount of variation in the world's languages, there is consensus that there are limits to the variation, though stating exact universals is difficult (Greenberg, 1963, 1978; Mairal and Gil, 2006; Stabler, 2009). Even language typologists who are suspicious of hypothesized language universals, once aware of the kinds of patterns that are logically possible, agree that not any logically possible pattern could be a natural language pattern.

Here is a simple example: many linguists have observed that languages do not appear to count past two. For example, no language requires sentences with at least n words to have the n word be a verb where $n \geq 3$ (unlike verb-second languages like German). This is a logically possible language pattern. Here is another one: if an even number of adjectives modify a noun, then they follow the noun in noun phrases, but if an odd number of adjectives follow a noun they precede the noun in noun phrases. These are both r.e. patterns.

According to Chater and Vitányi (2007), if the linguistic input a child received contained sufficiently many examples of noun phrases which obeyed the even-odd adjective rule above, they would learn it. It's an empirical hypothesis, but I think children would fail to learn this rule no matter how many examples they were given. Chater and Vitányi can claim that there is a simpler pattern consistent with data (e.g. adjectives can optionally precede or follow nouns) that children settle on because their lives and childhoods are too short for there to be enough data to move from the simpler generalization to the correct one. This also leads to an interesting, unfortunately untestable, prediction, that if humans only had longer lives and childhoods, we could learn such bizarre patterns like the even-odd adjective rule. In other words, they would explain the absence of even-odd adjective rule in natural languages as just a byproduct of short lives and childhoods, whereas I would attribute it to linguistic principles which exclude it from the collection of hypotheses children entertain. But there is a way to Chater and Vitányi can address the issue: How much data does "the ideal language learner" require to converge to the correct unattested pattern?

The harder learning frameworks—identification in the limit from positive data and PAC—bring more insight into the problem of learning and the nature of learnable classes of patterns.

First, these definitions of learning makes clear that the central problem in learning is generalizing beyond one's experience because under these definitions, generalizing to infinite

Table 4: Birds (a,b) are "warbler-barblers". Which birds (c-g) do you think are "warbler-barblers"?

patterns requires the impossibility of being able to learn certain finite patterns (Gold's first point above). I think humans behave like this. Consider the birds in Table 4. If I tell you birds (a,b) are "warbler-barblers" and ask which other birds (c,d,e,f,g) are warbler-barblers you're likely to decide that birds (c,f,g) could be warbler-barblers but birds (d,e) definitely not. You'd be very surprised to learn that in fact birds (a,b) are the only warbler-barblers of all time ever. Humans never even consider the possibility that there could just be exactly two "warbler-barblers". This insight is expressed well by Gleitman (1990:12):

> The trouble is that an observer who notices *everything* can learn *nothing* for there is no end of categories known and constructible to describe a situation [emphasis in original].

Chater and Vitányi (2007) can say that grammars to describe finite languages are more complex than regular or context-free grammars, and they are right, provided the finite language is big enough. Again, the question is what kind experience does "the ideal language learner" need in order to learn a finite language with exactly n sentences, and is this humanlike? This question should be asked of all proposed language learning models, and it is interesting to contrast "the ideal language learner" with Yoshinaka's (2008,2009) learners which generalize to $a^n b^n$ and $a^n b^n c^n$ with at most a few examples (and so those learners cannot learn, under any circumstances, the finite language that contains only those few examples).

Second, classes which are learnable within these difficult frameworks have the potential to yield new kinds of insights about which properties of natural languages make them learnable. As discussed in section 4.2, there are many positive results of interesting subclasses of major regions of the Chomsky Hierarchy which are identifiable in the limit from positive data and/or PAC-learnable, and which describe natural language patterns. The learners for those classes succeed because of the structure inherent to the hypothesis space—structure which can reflect deep, universal properties of natural language. Under weaker definitions of learning, where the r.e. class of patterns is learnable, such insights are less likely to be forthcoming.

As for Gold's second point, there has been some empirical study as to whether children use negative evidence in language acquisition (Brown and Hanlon, 1970; Marcus, 1993). Also learning frameworks which permit queries (Angluin, 1988a, 1990), especially correction queries (Becerra-Bonache *et al.*, 2006; Tîrnauca, 2008), can be thought as allowing learners to access implicit negative evidence.

As for third question above, I don't know of any research that has addressed it. Nonetheless, it should be clear that the commonly-cited statistical learning frameworks that have shown probabilistic context-free langauges are learnable (Horning, 1969), and in fact the r.e. stochastic languages are learnable (Angluin, 1988b; Chater and Vitányi, 2007) are pursuing

Gold's third suggestion. It also ought to be clear that the postive results that show r.e. patterns can be learned from positive, full and computable data presentations are "in principle" learners only. As far as is known, they can not learn these classes feasibly. Of course it may be possible that such techniques can feasibly learn interesting subclasses of major regions of the Chomsky Hierarchy which are relevant to natural language. If shown, this would be an interesting complement to the research efforts pursuing Gold's first suggestion, and could also reveal universal properties of natural language that contribute to their learnabilty.

6 Artificial language learning experiments

Many questions raised in the last section can in principle be addressed by artificial language learning experiments. These experiments probe the generalizations people make on the basis of brief finite exposure to artificial languages (Marcus *et al.*, 1999; Fitch and Hauser, 2004; Wilson, 2006; Berent *et al.*, 2008). The performance of computational models of learning can be compared with the performance of humans (Wilson, 2006; Michael C. Frank *et al.*, 2007) on these experiments.

But the relationship can go beyond comparison and evaluation to design. Well-defined learnable classes which contain natural language patterns are the bases for experiments. As mentioned, there are non-trivial interesting classes of languages which are PAC-learnable, which are identifiable in the limit from positive data, and which contain natural language patterns. The proofs are constructive and a common technique is identifying exactly the finite experience the proposed learners need to generalize correctly to each language in a given class. This critical finite experience is called the characteristic sample. The characteristic sample essentially becomes the training stimuli for the experiments. Other sentences in the language that are not part of the characteristic sample become test items. Finally, more than one learner can be compared by finding test items in the symmetric difference of the different patterns multiple learners return from the experimental stimuli. These points are also articulated by Rogers and Hauser (2010) in the context of formal language theory and I encourage readers to study their paper.

7 Conclusion

In this chapter I have tried to explain what computational learning theories are, and the lessons language scientists can draw from them. I believe there is a bright future for research which honestly integrates the insights of computational learning theories with the insights and methodologies of developmental psycholinguistics.

References

Abney, Steven. 1996. Statistical methods and linguistics. In *The balancing act: Combining symbolic and statistical approaches to language*, edited by J.L. Klavans ad P. Resnik, 1–26. Cambridge, MA: MIT Press.

Angluin, D. 1988a. Queries and concept learning. *Machine Learning* 2:319–342.

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

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

Angluin, Dana. 1988b. Identifying languages from stochastic examples. Tech. Rep. 614, Yale University, New Haven, CT.

Angluin, Dana. 1990. Negative results for equivalence queries. *Machine Learning* 5:121–150.

Angluin, Dana, and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2:343–370.

Anthony, M., and N. Biggs. 1992. *Computational Learning Theory*. Cambrisge University Press.

Bates, Elizabeth, and Jeffrey Elman. 1996. Learning rediscovered. *Science* 274:1849–1850.

Becerra-Bonache, L., J. Case, S. Jain, and F. Stephan. 2008. Iterative learning of simple external contextual languages. In *19th International Conference on Algorithmic Learning Theory (ALT'08)*, vol. 5254, 359–373. Springer. Expanded journal version accepted for the associated Special Issue of *TCS*, 2009.

Becerra-Bonache, Leonor, Adrian Horia Dediu, and Cristina Tîrnauca. 2006. Learning dfa from correction and equivalence queries. In *ICGI*, vol. 4201 of *Lecture Notes in Computer Science*, 281–292. Springer.

Berent, I., T. Lennertz, J. Jun, M. A. Moreno, and P. Smolensky. 2008. Language universals in human brains. *Proceedings of the National Academy of Sciences* 105:5321–5325.

Berwick1985. 1985. *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Blumer, Anselm, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* 36:929–965.

Brown, R., and C. Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In *Cognition and the developmental of language*, edited by J. Hayes, 11–53. New York: Wiley.

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.

Clark, Alexander, François Coste, and Lauren Miclet, eds. 2008. *Grammatical Inference: Algorithms and Applications*, vol. 5278 of *Lecture Notes in Computer Science*. Springer.

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

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

Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual ACL and 8th EACL*, 313–320. Madrid.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press and Bradford Book.

Fernau, Henning. 2003. Identification of function distinguishable languages. *Theoretical Computer Science* 290:1679–1711.

Fitch, W.T., and M.D. Hauser. 2004. Computational constraints on syntactic processing in nonhuman primates. *Science* 303:377–380.

García, Pedro, and José Ruiz. 1996. Learning k-piecewise testable languages from positive data. In *Grammatical Interference: Learning Syntax from Sentences*, edited by Laurent Miclet and Colin de la Higuera, vol. 1147 of *Lecture Notes in Computer Science*, 203–210. Springer.

García, Pedro, and José Ruiz. 2004. Learning k-testable and k-piecewise testable languages from positive data. *Grammars* 7:125–140.

Garcia, Pedro, Enrique Vidal, and José Oncina. 1990. Learning locally testable languages in the strict sense. In *Proceedings of the Workshop on Algorithmic Learning Theory*, 325–338.

Garey, M. R., and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1:3–55.

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

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

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, edited by Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Language*, 73–113. Cambridge: MIT Press.

Greenberg, Joseph. 1978. Initial and final consonant sequences. In *Universals of Human Language: Volume 2, Phonology*, edited by Joseph Greenberg, 243–279. Stanford University Press.

Griffiths, T.L., C. Kemp, and J. B. Tenenbaum. 2008. Bayesian models of cognition. In *The Cambridge handbook of computational cognitive modeling*, edited by Ron Sun. Cambridge University Press.

Harrison, Michael A. 1978. *Introduction to Formal Language Theory*. Addison-Wesley Publishing Company.

Heinz, Jeffrey. 2008. Left-to-right and right-to-left iterative languages. In Clark *et al.* (2008), 84–97.

Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. *Phonology* 26:303–351.

Heinz, Jeffrey. 2010a. Learning long-distance phonotactics. *Linguistic Inquiry* 41.

Heinz, Jeffrey. 2010b. String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

de la Higuera, Colin. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.

Hopcroft, John, Rajeev Motwani, and Jeffrey Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Hopcroft, John, Rajeev Motwani, and Jeffrey Ullman. 2001. *Introduction to Automata Theory, Languages, and Computation*. Boston, MA: Addison-Wesley.

Horning, J. J. 1969. A study of grammatical inference. Doctoral dissertation, Stanford University.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change)*. 2nd ed. The MIT Press.

Johnson, Mark, and Stephan Reizeler. 2002. Statistical models of language learning and use. *Cognitive Science* 26:239–253.

Joshi, A. K. 1985. Tree-adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions? In *Natural Language Parsing*, edited by D. Dowty, L. Karttunen, and A. Zwicky, 206–250. Cambridge University Press.

Kaplan, Ronald, and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20:331–378.

Karttunen, Lauri. 1998. The proper treatment of optimality in computational phonology. In *FSMNLP'98*, 1–12. International Workshop on Finite-State Methods in Natural Language Processing, Bilkent University, Ankara, Turkey.

Kearns, Michael, and Ming Li. 1993. Learning in the presence of malicious errors. *SIAM Journal of Computing* 22.

Kearns, Michael, and Umesh Vazirani. 1994a. *An Introduction to Computational Learning Theory*. MIT Press.

Kearns, M.J., and U.V. Vazirani. 1994b. *An Introduction to Computational Learning Theory*. Cambridge MA: MIT Press.

Klein, D. 2005. The unsupervised learning of natural language. Doctoral dissertation, Stanford University.

Lange, Steffen, Thomas Zeugmann, and Sandra Zilles. 2008. Learning indexed families of recursive languages from positive data: A survey. *Theoretical Computer Science* 397:194–232.

Mairal, Ricardo, and Juana Gil, eds. 2006. *Linguistic Universals*. Cambridge University Press.

Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, G. F., S. Vijayan, S. B. Rao, and P. Vishton. 1999. Rule learning by seven-month-old infants. *Science* 283:77–80.

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

Michael C. Frank, Sharon Goldwater, Vikash Mansinghka, Tom Griffiths, and Joshua Tenenbaum. 2007. Modeling human performance on statistical word segmentation tasks. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*.

Muggleton, Stephen. 1990. *Inductive Acquisition of Expert Knowledge*. Addison-Wesley.

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

Niyogi, Partha, and Robert Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61:161–193.

Nowak, Martin A., Natalia L. Komarova, and Partha Niyogi. 2002. Computational and evolutionary aspects of language. *Nature* 417:611–617.

Oates, Tim, Tom Armstrong, and Leonor Becerra Bonache. 2006. Inferring grammars for mildly context-sensitive languages in polynomial-time. In *Proceedings of the 8th International Colloquium on Grammatical Inference (ICGI)*, 137–147.

Osherson, Daniel, Scott Weinstein, and Michael Stob. 1986. *Systems that Learn*. Cambridge, MA: MIT Press.

Otto, F. 1985. Classes of regular and context-free languages over countably infinite alphabets. *Discrete Applied Mathematics* 12:41–56.

Papadimitriou, Christon. 1994. *Computational Complexity*. Addison Wesley.

Perfors, Amy, Joshua B. Tenenbaum, Edward Gibson, and Terry Regier. 2010. How recursive is language. In *Recursion and Human Language*, edited by Harry van der Hulst, chap. 9, 159–175. Berlin, Germany: De Gruyter Mouton.

Pitt, Leonard. 1985. Probabilistic inductive inference. Doctoral dissertation, Yale University. Computer Science Department, TR-400.

Rogers, James, and Marc Hauser. 2010. The use of formal languages in artificial language learning: a proposal for distinguishing the differences between human and nonhuman animal learners. In *Recursion and Human Language*, edited by Harry van der Hulst, chap. 12, 213–232. Berlin, Germany: De Gruyter Mouton.

Rumelhart, D. E., and J. L. McClelland. 1986. On learning the past tenses of english verbs. In *Parallel Distributed Processing, volume 2*, edited by J.L. McClelland and D. E. Rumelhart, 216–271. Cambridge MA: MIT Press.

Solomonoff, Ray J. 1978. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory* 24:422–432.

Stabler, Edward P. 2009. Computational models of language universals. In *Language Universals*, edited by Christiansen, Collins, and Edelman, 200–223. Oxford: Oxford University Press.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in optimality theory. *Linguistic Inquiry* 229–268.

Thomas, Wolfgang. 1997. Languages, automata, and logic. vol. 3, chap. 7. Springer.

Tîrnauca, Cristina. 2008. A note on the relationship between different types of correction queries. In Clark *et al.* (2008), 213–223.

Turing, Alan. 1937. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society* s2:230–265.

Valiant, L.G. 1984. A theory of the learnable. *Communications of the ACM* 27:1134–1142.

Vapnik, Vladimir. 1995. *The nature of statistical learning theory*. New York: Springer.

Vapnik, Vladimir. 1998. *Statistical Learning Theory*. New York: Wiley.

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

Wiehagen, R., R. Frievalds, and E. Kinber. 1984. On the power of probabilistic strategies in inductive inference. *Theoretical Computer Science* 28:111–133.

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

Yoshinaka, R. 2009. Learning mildly context-sensitive languages with multidimensional substitutability from positive data. In *20th International Conference on Algorithmic Learning Theory (ALT'09)*, vol. 5809 of *Lecture Notes in Artificial Intelligence*.

Yoshinaka, Ryo. 2008. Identification in the limit of k , l -substitutable context-free languages. In Clark *et al.* (2008), 266–279.

Zeugmann, Thomas, and Sandra Zilles. 2008. Learning recursive functions: A survey. *Theoretical Computer Science* 397:4–56.