Grammatical Inference and the Argument from the Poverty of the Stimulus

Alexander Clark (asc@aclark.demon.co.uk)
ISSCO / TIM, University of Geneva
UNI-MAIL, Boulevard du Pont-d'Arve,
CH-1211 Genève 4, Switzerland

Abstract

Formal results in grammatical inference clearly have some relevance to first language acquisition. Initial formalisations of the problem (Gold 1967) are however inapplicable to this particular situation. In this paper we construct an appropriate formalisation of the problem using a modern vocabulary drawn from statistical learning theory and grammatical inference and looking in detail at the relevant empirical facts. We claim that a variant of the Probably Approximately Correct (PAC) learning framework (Valiant 1984) with positive samples only, modified so it is not completely distribution free is the appropriate choice. Some negative results derived from cryptographic problems (Kearns *et al.* 1994) appear to apply in this situation but the existence of algorithms with provably good performance (Ron, Singer, & Tishby 1995) and subsequent work, shows how these negative results are not as strong as they initially appear, and that recent algorithms for learning regular languages partially satisfy our criteria. We conclude by speculating about the extension of these results beyond regular languages.

Introduction

For some years, the relevance of formal results in grammatical inference to the empirical question of first language acquisition by infant children has been recognised (Wexler & Culicover 1980). Unfortunately, for many researchers, with a few notable exceptions (Abe 1988), this begins and ends with Gold's famous negative results in the identification in the limit paradigm. This paradigm, though still widely used in the grammatical inference community, is clearly of limited relevance to the issue at hand, since it requires the model to be able to exactly identify the target language even when an adversary can pick arbitrarily misleading sequences of examples to provide. Moreover, the paradigm as stated has no bounds on the amount of data or computation required for the learner. In spite of the inapplicability of this particular paradigm, it is possible to construct, as we shall see, quite strong arguments that bear directly on this problem.

Grammatical inference is the study of machine learning of formal languages. It has a vast formal vocabulary and has been applied to a wide selection of different problems, where the "languages" under study can be (representations of) parts of natural languages, sequences of nucleotides, moves of a robot, or some other sequence data. For any conclusions that we draw from formal discussions to have any applicability to the real world, we must be sure to select, or construct, from the rich set of formal devices available an appropriate formalisation. Even then, we should be very cautious about making inferences about how the infant child must or cannot learn language: subsequent developments in GI might allow a more nuanced description in which these conclusions are not valid. The situation is complicated by the fact that the field of grammtical inference, much like the wider field of machine learning in general, is in a state of rapid change.

In this paper we hope to address this problem by justifying the selection of the appropriate learning framework starting by looking at the actual situation the child is in, rather than from an *a priori* decision about the right framework. We will not attempt a survey of grammatical inference techniques; nor shall we provide proofs of the theorems we use here. Though these arguments have often been used in support of nativist theories of language acquisition – i.e. theories that posit the existence of large amounts of detailed language specific innate knowledge that the child is provided with through some genetically specified language organ – we are not concerned with these debates here. In fact, as we shall see below, some of these arguments apply as strongly to principle and parameter based theories (Chomsky 1986) as to empiricist models. These arguments are more pertinent to questions about the autonomy and modularity of language learning: the question whether learning of some level of linguistic knowledge – morphology or syntax, for example – can take place in isolation from other forms of learning – such as word meaning.

Positive results can help us to understand how humans might learn languages by outlining the class of algorithms that *might* be used by humans, considered as computational systems at a suitable abstract level. Conversely, negative results might be helpful if they could demonstrate that no algorithms of a certain class could perform the task – in this case we could know that the human child learns his language in some other way.

On a terminological note we should mention the phrase "The argument from the poverty of the stimulus". This now classic argument holds that the stimulus available to the

child is too poor to allow learning to proceed. There are two versions of the argument as (Cowie 1999) argues convincingly. The first sort can be called *a posteriori* arguments. These rely on particular contingent facts about the linguistic experience of the child – in particular the quantity of data, and the frequency of particular constructions therein. We shall not be discussing these arguments here; the reader is referred to the vigorous debate in (Pullum & Scholz 2002). We focus here on the second class of argument, a more *a priori* and purely theoretical argument, that depends on mathematical results from learnability theory for its force.

We shall proceed as follows: after briefly describing FLA, we describe the various elements of a model of learning, or framework. We then make a series of decisions based on the empirical facts about FLA, to construct an appropriate model or models, avoiding unnecessary idealisation wherever possible. We proceed to some strong negative results, well-known in the GI community that bear on the questions at hand. The most powerful of these (Kearns *et al.* 1994) appears to apply quite directly to our chosen model. We then discuss an interesting algorithm (Ron, Singer, & Tishby 1995) which shows that this can be circumvented, at least for a subclass of regular languages. Finally, after discussing the possibilities for extending this result to all regular languages, and beyond, we conclude with a discussion of the implications of the results presented.

First Language Acquisition

Let us first examine the phenomenon we are concerned with: first language acquisition. In the space of a few years, children almost invariably acquire, in the absence of explicit instruction, one or more of the languages that they are exposed to. A multitude of subsidiary debates have sprung up around this central issue covering questions about critical periods – the ages at which this can take place, the exact nature of the evidence available to the child, and the various phases of linguistic use through which the infant child passes. In the opinion of many researchers, explaining this ability is one of the most important challenges facing linguists and cognitive scientists today.

A difficulty for us in this paper is that many of the idealisations made in the study of this field are in fact demonstrably false. Classical assumptions, such as the existence of uniform communities of language users, are well-motivated in the study of the "steady state" of a system, but less so when studying acquisition and change. There is a regrettable tendency to slip from viewing these idealisations correctly – as counter-factual idealizations – to viewing them as empirical facts that need to be explained. Thus, when looking for an appropriate formulation of the problem, we should recall for example the fact that different children do not converge to exactly the same knowledge of language as is sometimes claimed, nor do all of them acquire a language competently at all , since there is a small proportion of children who though apparently neurologically normal fail to acquire language. In the context of our discussion later on, these observations lead us to accept slightly less stringent criteria where we allow a small probability of failure and do not demand perfect equality of hypothesis and target.

Grammatical Inference

The general field of machine learning has a specialised subfield that deals with the learning of formal languages. This field, Grammatical Inference (GI), is characterised above all by an interest in formal results, both in terms of formal characterisations of the target languages, and in terms of formal proofs either that particular algorithms can learn according to particular definitions, or that sets of language cannot be learnt. In spite of its theoretical bent, GI algorithms have also been applied with some success. Natural language, however is not the only source of real-world applications for GI. Other domains include biological sequence data, artificial languages, such as discovering XML schemas, or sequences of moves of a robot. The field is also driven by technical motives and the intrinsic elegance and interest of the mathematical ideas employed. In summary it is not just about language, and accordingly it has developed a rich vocabulary to deal with the wide range of its subject matter.

In particular, researchers are often concerned with formal results – that is we want algorithms where we can *prove* that they will perform in a certain way. Often, we may be able to empirically establish that a particular algorithm performs well, in the sense of reliably producing an accurate model, while we may be unable to prove formally that the algorithm will always perform in this way. This can be for a number of reasons: the mathematics required in the derivation of the bounds on the errors may be difficult or obscure, or the algorithm may behave strangely when dealing with sets of data which are ill-behaved in some way.

The basic framework can be considered as a game played between two players. One player, the teacher, provides information to another, the learner, and from that information the learner must identify the underlying language. We can break down this situation further into a number of elements. We assume that the languages to be learned are drawn in some way from a possibly infinite class of languages, $\mathcal{L}$, which is a set of formal mathematical objects. The teacher selects one of these languages, which we call the *target*, and then gives the learner a certain amount of information of various types about the target. After a while, the learner then returns its guess, the hypothesis, which in general will be a language drawn from the same class $\mathcal{L}$. Ideally the learner has been able to deduce or induce or abduce something about the target from the information we have given it, and in this case the hypothesis it returns will be identical to, or close in some technical sense, to the target. If the learner can conistently do this, under whatever constraints we choose, then we say it can learn that class of languages. To turn this vague description into something more concrete requires us to specify a number of things.

- What sort of mathematical object should we use to represent a language?
- What is the target class of languages?
- What information is the learner given?
- What computational constraints does the learner operate under?

- How close must the target be to the hypothesis, and how do we measure it?

This paper addresses the extent to which negative results in GI, could be relevant to this real world situation. As always, when negative results from theory are being applied, a certain amount of caution is appropriate in examining the underlying assumptions of the theory and the extent to which these are applicable. As we shall see, in our opinion, none of the current negative results, though powerful, are applicable to the empirical situation. We shall accordingly, at various points, make strong pessimistic assumptions about the learning environment of the child, and show that even under these unrealistically stringent stipulations, the negative results are still inapplicable. This will make the conclusions we come to a little sharper. Conversely, if we wanted to show that the negative results did apply, to be convincing we would have to make rather optimistic assumptions about the learning environment.

Applying GI to FLA

We now have the delicate task of selecting, or rather constructing, a formal model by identifying the various components we have identified above. We want to choose the model that is the best representation of the learning task or tasks that the infant child must perform. We consider that some of the empirical questions do not yet have clear answers. In those cases, we shall make the choice that makes the learning task more difficult. In other cases, we may not have a clear idea of how to formalise some information source. We shall start by making a significant idealisation: we consider language acquisition as being a single task. Natural languages as traditionally describe have different levels. At the very least we have morphology and syntax; one might also consider inter-sentential or discourse as an additional level. We conflate all of these into a single task: learning a formal language; in the discussion below, for the sake of concreteness and clarity, we shall talk in terms of learning syntax.

The Language

The first question we must answer concerns the language itself. A formal language is normally defined as follows. Given a finite alphabet Σ, we define the set of all strings (the free monoid) over Σ as $\Sigma*$. We want to learn a language $L \subset \Sigma^*$. The alphabet Σ could be a set of phonemes, or characters, or a set of words, or a set of lexical categories (part of speech tags). The language could be the set of well-formed sentences, or the set of words that obey the phonotactics of the language, and so on. We reduce all of the different learning tasks in language to a single abstract task – identifying a possibly infinite set of strings. This is overly simplistic since transductions, i.e. mappings from one string to another, are probably also necessary. We are using here a standard definition of a language where every string is unambiguously either in or not in the language.. This may appear unrealistic – if the formal language is meant to represent the set of grammatical sentences, there are well-known methodological problems with deciding where exactly to draw the line between grammatical and ungrammatical sentences. An alternative might be to consider acceptability rather than grammaticality as the defining criterion for inclusion in the set. Moreover, there is a certain amount of noise in the input – There are other possibilities. We could for example use a fuzzy set – i.e. a function from $\Sigma^* \rightarrow [0, 1]$ where each string has a degree of membership between 0 and 1. This would seem to create more problems than it solves. A more appealing option is to learn distributions, again functions f from $\Sigma^* \rightarrow [0, 1]$ but where $\sum_{s \in L} f(s) = 1$. This is of course the classic problem of language modelling, and is compelling for two reasons. First, it is empirically well grounded – the probability of a string is related to its frequency of occurrence, and secondly, we can deduce from the speech recognition capability of humans that they must have some similar capability.

Both possibilities – crisp languages, and distributions – are reasonable. the choice depends on what one considers the key phenomena to be explained are – grammaticality judgments by native speakers, or natural use and comprehension of the language. We favour the latter, and accordingly think that learning distributions is a more accurate and more difficult choice.

The class of languages

A common confusion in some discussions of this point is between languages and classes of languages. Learnability is a property of *classes* of languages. If there is only one language in the class of languages to be learned then the learner can just guess that language and learn it. A class with two languages is again trivially learnable if you have an efficient algorithm for testing membership. It is only when the set of languages is exponentially large or infinite, that the problem becomes non-trivial, from a theoretical point of view. The class of languages we need is a class of languages that includes all attested human languages and additionally all "possible" human languages. Natural languages are thought to fall into the class of mildly context-sensitive languages, (Vijay-Shanker & Weir 1994), so clearly this class is large enough. It is, however, not necessary that our class be this large. Indeed it is essential for learnability that it is not. As we shall see below, even the class of regular languages contains some subclasses that are computationally hard to learn. Indeed, we claim it is reasonable to define our class so it does *not* contain languages that are clearly not possible human languages. Figure 1 shows diagramatically the relationships between the various classes we discuss in this paper.

Information sources

Next we must specify the information that our learning algorithm has access to. Clearly the primary source of data is the *primary linguistic data* (PLD), namely the utterances that occur in the child's environment. These will consist of both child-directed speech and adult-to-adult speech. These are generally acceptable sentences that is to say sentences that are in the language to be learned. These are called *positive* samples. One of the most long-running debates in this field

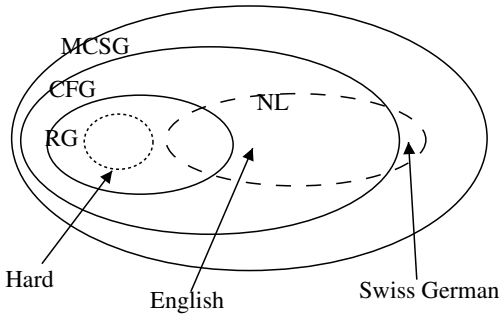

Figure 1: Diagram showing possible classes of languages. RG is the class of regular languages, CFG is context free, MSCG is the class of mildly context sensitive languages that include linear indexed grammars. The dashed ellipse shows a putative class of possible natural languages, that includes all attested natural languages, but excludes some computationally difficult languages shown in the dotted ellipse.

is over whether the child has access to negative data – unacceptable sentences that are marked in some way as such. The consensus (Marcus 1993) appears to be that they do not. In middle-class Western families, children are provided with some sort of feedback about the well-formedness of their utterances, but this is unreliable and erratic, not a universal of global child-raising. Furthermore this appears to have no effect on the child. Children do also get indirect pragmatic feedback if their utterances are incomprehensible. In our opinion, both of these would be better modelled by what is called a membership query: the algorithm may generate a string and be informed whether that string is in the language or not. However, we feel that this is too erratic to be considered an essential part of the process. Another question is whether the input data is presented as a flat string or annotated with some sort of structural evidence, which might be derived from prosodic or semantic information. Unfortunately there is little agreement on what the constituent structure should be – indeed many linguistic theories do not have a level of constituent structure at all, but just dependency structure.

Semantic information is also claimed as an important source. The hypothesis is that children can use lexical semantics, coupled with rich sources of real-world knowlege to infer the meaning of utterances from the situational context. That would be an extremely powerful piece of information, but it is clearly absurd to claim that the meaning of an utterance is uniquely specified by the situational context. If true, there would be no need for communication or information transfer at all. Of course the context puts some constraints on the sentences that will be uttered, but it is not clear how to incorporate this fact without being far too generous. In summary it appears that only positive evidence can be unequivocally relied upon though this may seem a harsh and unrealistic environment.

Presentation

We have now decided that the only evidence available to the learner will be unadorned positive samples drawn from the target language. There are various possibilities for how the samples are selected. The choice that is most favourable for the learner is where they are slected by a helpful teacher to make the learning process as easy as possible (Goldman & Mathias 1996). While it is certainly true that carers speak to small children in sentences of simple structure (Motherese), this is not true for all of the data that the child has access to, nor is it universally valid. Moreover, there are serious technical problems with formalising this, namely what is called 'collusion' where the teacher provides examples that encode the grammar itself, thus trivialising the learning process. Though attempts have been made to limit this problem, they are not yet completely satisfactory. The next alternative is that the examples are selected randomly from some fixed distribution. This appears to us to be the appropriate choice, subject to some limitations on the distributions that we discuss below. The final option, the most difficult for the learner, is where the sequence of samples can be selected by an intelligent adversary, in an attempt to make the learner fail, subject only to the weak requirement that each string in the language appears at least once. This is the approach taken in the identification in the limit paradigm (Gold 1967), and is clearly too stringent. The remaining question then regards the distribution from which the samples are drawn: whether the learner has to be able to learn for every possible distribution, or only for distributions from a particular class, or only for one particular distribution.

Resources

Beyond the requirement of computability we will wish to place additional limitations on the computational resources that the learner can use. Since children learn the language in a limited period of time, which limits both the amount of data they have access to and the amount of computation they can use, it seems appropriate to disallow algorithms that use unbounded or very large amounts of data or time. As normal, we shall formalise this by putting polynomial bounds on the *sample complexity* and *computational complexity*. Since the individual samples are of varying length, we need to allow the computational complexity to depend on the total length of the sample. A key question is what the parameters of the sample complexity polynomial should be. We shall discuss this further below.

Convergence Criteria

Next we address the issue of reliability: the extent to which all children acquire language. First, variability in achievement of particular linguistic milestones is high. There are numerous causes including deafness, mental retardation, cerebral palsy, specific language impairment and autism. Generally, autistic children appear neurologically and physically normal; but about half may never speak. Autism, on some accounts, has an incidence of about 0.2% Therefore we can require learning to happen with arbitrarily high probability, but requiring it to happen with probability one is unreasonable. A related question concerns convergence: the

extent to which children exposed to a linguistic environment end up with the same language as others. Clearly they are very close since otherwise communication could not happen, but there is ample evidence from studies of variation (Labov 1975), that there are non-trivial differences between adults, who have grown up with near-identical linguistic experiences, about the interpretation and syntactic acceptability of simple sentences, quite apart from the wide purely lexical variation that is easily detected. An example in English is "Each of the boys didn't come".

Moreover, language change *requires* some children to end up with slightly different grammars from the older generation. At the very most, we should require that the hypothesis should be close to the target. The function we use to measure the 'distance' between hypothesis and target depends on whether we are learnng crisp languages or distributions. If we are learning distributions then the obvious choice is the Kullback-Leibler divergence – a very strict measure. For crisp languages, the probability of the symmetric difference with respect to some distribution is natural.

PAC-learning

These considerations lead us to some variant of the Probably Approximately Correct (PAC) model of learning (Valiant 1984). We require the algorithm to produce with arbitrarily high probability a good hypothesis. We formalise this by saying that for any $\delta > 0$ it must produce a good hypothesis with probability more than $1 - \delta$. Next we require a good hypothesis to be arbitrarily close to the target, so we have a precision ϵ and we say that for any $\epsilon > 0$, the hypothesis must be less than ϵ away from the target. We allow the amount of data it can use to increase as the confidence and precision get smaller. We define PAC-learning in the following way: given a finite alphabet Σ, and a class of languages $\mathcal{L}$ over Σ, an algorithm PAC-learns the class $\mathcal{L}$, if there is a polynomial q, such that for every confidence $\delta > 0$ and precision $\epsilon > 0$, for every distribution D over Σ^* , for every language L in $\mathcal{L}$, whenever the number of samples exceeds $q(1/\epsilon, 1/\delta, |\Sigma|, |L|)$, the algorithm must produce a hypothesis H such that with probability greater than $1 - \delta$, $Pr_D(H \Delta L > \epsilon)$. Here we use $A \Delta B$ to mean the symmetric difference between two sets. The polynomial q is called the sample complexity polynomial. We also limit the amount of computation to some polynomial in the total length of the data it has seen. Note first of all that this is a worst case bound – we are not requiring merely that on average it comes close. Additionally this model is what is called 'distribution-free'. This means that the algorithm must work for every combination of distribution and language. This is a very stringent requirement, only mitigated by the fact that the error is calculated with respect to the same distribution that the samples are drawn from. Thus, if there is a subset of Σ^* with low aggregate probability under D, the algorithm will not get many samples from this region but will not be penalised very much for errors in that region. From our point of view, there are two problems with this framework: first, we only want to draw positive samples, but the distributions are over all strings in Σ^*, and include some that give a zero probability to all strings in the language concerned. Secondly, this is too pessimistic because the distribution has no relation to the language: intuitively it's reasonable to expect the distribution to be derived in some way from the language, or the structure of a grammar generating the language.

One alternative that has been suggested is the PAC learning with simple distributions model introduced by (Denis 2001). This is based on ideas from complexity theory where the samples are drawn according to a universal distribution defined by the conditional Kolmogorov complexity. While mathematically correct this is inappropriate as a model of FLA for a number of reasons. First, learnability is proven only on a single very unusual distribution, and relies on particular properties of this distribution, and secondly there are some very large constants in the sample complexity polynomial.

The solution we favour is to define some natural class of distributions based on a grammar or automaton generating the language. Given a class of languages defined by some generative device, there is normally a natural stochastic variant of the device which defines a distribution over that language. Thus regular languages can be defined by a finite-state automaton, and these can be naturally extended to Probabilistic finite state automaton. Similarly context free languages are normally defined by context-free grammmars which can be extended again to to Probabilistic or stochastic CFG. We therefore propose a slight modification of the PAC-framework. For every class of languages $\mathcal{L}$, defined by some formal device define a class of distributions defined by a stochastic variant of that device. $\mathcal{D}$. Then for each language L, we select the set of distributions whose support is equal to the language: $D_L^+ = \{D \in \mathcal{D} : \forall s \in \Sigma^* s \in L \Leftrightarrow P_D(s) > 0\}$. Samples are drawn from one of these distributions.

There are two technical problems here: first, this doesn't penalise over-generalisation. Since the distribution is over positive examples, negative examples have zero weight, so we need some penalty function over negative examples or alternatively require the hypothesis to be a subset of the target. Secondly, this definition is too vague. The exact way in which you extend the "crisp" language to a stochastic one can have serious consequences. When dealing with regular languages, for example, though the class of languages defined by deterministic automata is the same as that defined by non-deterministic languages, the same is not true for their stochastic variants (Esposito *et al.* 2002). Additionally, one can have exponential blow-ups in the number of states when determinising automata. Similarly, with CFGs, (Abney, McAllester, & Pereira 1999) showed that converting between two parametrisations of stochastic Context Free languages are equivalent but that there are blow-ups in both directions. We do not have a completely satisfactory solution to this problem at the moment; an alternative is to consider learning the distributions rather than the languages.

In the case of learning distributions, we have the same framework, but the samples are drawn according to the distribution being learned T, and we require that the hypothesis H has small divergence from the target: $D(T||H) < \epsilon$. Since the divergence is infinite if the hypothesis gives prob-

ability zero to a string in the target, this will have the consequence that the target must assign a non-zero probability to every string.

Negative Results

Now that we have a fairly clear idea of various ways of formalising the situation we can consider the extent to which formal result apply. First we consider negative results, which in Machine Learning come in two types. First, information-theoretic bounds on sample complexity, derived from the Vapnik-Chervonenkis (VC) dimension of the space of languages, a measure of the complexity of the set of hypotheses. If we add a parameter to the sample complexity polynomial that represents the complexity of the concept to be learned then this will remove these problems. This can be the size of a representation of the target which will be a polynomial in the number of states, or simply the number of non-terminals or states. This is very standard in most fields of machine learning.

The second problem relates not to the amount of information but to the computation involved. Results derived from cryptographic limitations on computational complexity, can be proved based on widely held and well supported assumptions that certain hard cryptographic problems are insoluble. In what follows we assume that there are no efficient algorithms for common cryptographic problems such as factoring Blum integers, inverting RSA function, recognizing quadratic residues or learning noisy parity functions.

There may be algorithms that will learn with reasonable amounts of data but that require unfeasibly large amounts of computation to find. There are a number of powerful negative results on learning in the purely distribution-free situation we considered and rejected above. (Kearns & Valiant 1989) showed that acyclic deterministic automata are not learnable even with positive and negative examples. Similarly, (Abe & Warmuth 1992) showed a slightly weaker representation dependent result on learning with a large alphabet for non-deterministic automata, by showing that there are strings such that maximising the likelihood of the string is NP-hard. Again this does not strictly apply to the partially distribution free situation we have chosen.

However there is one very strong result that appears to apply. A straightforward consequence of (Kearns *et al.* 1994) shows that Acyclic Deterministic Probabilistic FSA over a two letter alphabet cannot be learned under another cryptographic assumption (the noisy parity assumption). Therefore any class of languages that includes this comparatively weak family will not be learnable in out framework.

But this rests upon the assumption that the class of possible human languages must include some cryptographically hard functions. It appears that our formal apparatus does not distinguish between these cryptographic functions which hav been consciously designed to be hard to learn, and natural languages which presumably have evolved to be easy to learn since there is no evolutionary pressure to make them hard to decrypt – no intelligent predators eavesdropping for example. Clearly this is a flaw in our analysis: we need to find some more nuanced description for the class of possible human languages that excludes these hard languages.

Positive results

There is a positive result that shows a way forward. A PDFA is μ-distinguishable the distributions generated from any two states differ by at least μ in the L_∞-norm, i.e. there is a string with a difference in probability of at least μ. (Ron, Singer, & Tishby 1995) showed that μ-distinguishable acyclic PDFAs can be PAC-learned using the KLD as error function in time polynomial in $n, 1/\epsilon, 1/\delta, 1/\mu, |\Sigma|$. They use a variant of a standard state-merging algorithm. Since these are acyclic the languages they define are always finite. This additional criterion of distinguishability suffices to guarantee learnability. This work can be extended to cyclic automata (Clark & Thollard 2003), and thus the class of all regular languages, with the addition of a further parameter which bounds the expected length of a string generated from any state. The use of distinguishability seems innocuous; in syntactic terms it is a consequence of the plausible condition that for any pair of distinct non-terminals there is some fairly likely string generated by one and not the other. Similarly strings of symbols in natural language tend to have limited length. An alternate way of formalising this is to define a class of distinguishable automata, where the distinguishability of the automata is lower bounded by an inverse polynomial in the number of states. This is formally equivalent, but avoids adding terms to the sample complexity polynomial. In summary this would be a valid solution if all human languages actually lay within the class of regular languages. We have implemented this algorithm and it is a practical algorithm, Note also the general properties of this kind of algorithm: provably learning an infinite class of languages with infinite support using only polynomial amounts of data and computation.

Discussion

This topic has been discussed before – a recent survey is (Nowak, Komarova, & Niyogi 2002). However, the suitability of a Gold paradigm seems to have been accepted uncritically in most previous discussions when in our view it is clearly inappropriate. As we have seen, no current negative results apply with the exception of (Kearns *et al.* 1994), and there are some promising positive results for regular languages. Extension to context free grammars or beyond requires or larger requires much further work. There are at least three areas of difficulty. First, determinism is an important part of the algorithms we have discussed here. Completely non-deterministic grammars are very hard to learn. There is some hope that mildly non-deterministic grammars might be learnable (Esposito *et al.* 2002). Secondly, some of the decision problems for CFGs that one might want to use in an algorithm, are undecidable. Finally, simple CFGs for natural languages have exponentially large numbers of non-terminals (Gazdar *et al.* 1985). On a more positive note, empirical work is promising in a number of specific fields, (Klein & Manning 2002; Clark 2002; 2003), though these algorithms have no formal guarantees of convergence to a correct grammar.

It is worth pointing out that the negative result of (Kearns *et al.* 1994) applies with equal force to Principles and Pa-

rameters based models of language acquisition. The specific class of noisy parity functions they prove are unlearnable, are parametrised by a number of binary parameters in a way very reminiscent of Chomskyan theories. The mere fact that there are a finite number of parameters does not suffice to guarantee learnability, if the resulting class of languages is exponentially large.

In summary, we have proposed a formal analysis of first language acquisition. We have discussed how even the strongest negative result currently known, does not rule out a purely autonomous learning system, and that on the contrary there are encouraging positive results for regular languages. Learnability can be guaranteed by rather banal statistical properties of the input distributions; it is not necessary to hypothesise highly structured classes of possible languages.

Acknowledgements

The work presented here has been done partially within the framework of the PASCAL project of the European Unions IST programme and has benefited from a subsidy by the Swiss OFES.

References

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

Abe, N. 1988. Feasible learnability of formal grammars and the theory of natural language acquisition. In *Proceedings of COLING 1988*, 1–6.

Abney, S.; McAllester, D.; and Pereira, F. 1999. Relating probabilistic grammars and automata. In *Proceedings of ACL '99*.

Chomsky, N. 1986. *Knowledge of Language : Its Nature, Origin, and Use*. Praeger.

Clark, A., and Thollard, F. 2003. Pac-learnability of probabilistic deterministic finite-state automata. Manuscript.

Clark, A. 2002. Memory-based learning of morphology with stochastic transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 513–520.

Clark, A. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth Annual Meeting of the European Association for Computational Linguistics EACL 2003*, 59–66.

Cowie, F. 1999. *What's Within? Nativism Reconsidered*. Oxford University Press.

Denis, F. 2001. Learning regular languages from simple positive examples. *Machine Learning* 44(1/2):37–66.

Esposito, Y.; Lemay, A.; Denis, F.; and Dupont, P. 2002. Learning probabilistic residual finite state automata. In Adriaans, P.; Fernau, H.; and van Zaannen, M., eds., *Grammatical Inference: Algorithms and Applications, ICGI '02*, volume 2484 of *LNAI*, 77–91. Berlin, Heidelberg: Springer-Verlag.

Gazdar, G.; Klein, E.; Pullum, G.; and Sag, I. 1985. *Generalised Phrase Structure Grammar*. Basil Blackwell.

Gold, E. M. 1967. Language indentification in the limit. *Information and control* 10(5):447 – 474.

Goldman, S. A., and Mathias, H. D. 1996. Teaching a smarter learner. *Journal of Computer and System Sciences* 52(2):255–267.

Kearns, M., and Valiant, G. 1989. Cryptographic limitations on learning boolean formulae and finite automata. In *21st annual ACM symposium on Theory of computating*, 433–444. New York: ACM.

Kearns, M.; Mansour, Y.; Ron, D.; Rubinfeld, R.; Schapire, R.; and Sellie, L. 1994. On the learnability of discrete distributions. In *Proc. of the 25th Annual ACM Symposium on Theory of Computing*, 273–282.

Klein, D., and Manning, C. D. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the ACL*.

Labov, W. 1975. Empirical foundations of linguistic theory. In Austerlitz, R., ed., *The Scope of American Linguistics*. Peter de Ridder Press.

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

Nowak, M. A.; Komarova, N. L.; and Niyogi, P. 2002. Computational and evolutionary aspects of language. *Nature* 417:611–617.

Pullum, G. K., and Scholz, B. C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1-2):9–50.

Ron, D.; Singer, Y.; and Tishby, N. 1995. On the learnability and usage of acyclic probabilistic finite automata. In *COLT 1995*, 31–40. Santa Cruz CA USA: ACM.

Valiant, L. 1984. A theory of the learnable. *Comunication of the ACM* 27(11):1134 – 1142.

Vijay-Shanker, K., and Weir, D. J. 1994. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory* 27(6):511–546.

Wexler, K., and Culicover, P. W. 1980. *Formal Principles of Language Acquisition*. MIT Press.