J. Linguistics **43** (2007), 393–427. © 2007 Cambridge University Press
doi:10.1017/S0022226707004628 Printed in the United Kingdom

Machine learning theory and practice as a source of insight into universal grammar[1]

SHALOM LAPPIN

Department of Philosophy, King's College London

STUART M. SHIEBER

Division of Engineering and Applied Sciences, Harvard University

(Received 29 May 2006; revised 14 December 2006)

In this paper, we explore the possibility that machine learning approaches to natural-language processing being developed in engineering-oriented computational linguistics may be able to provide specific scientific insights into the nature of human language. We argue that, in principle, machine learning results could inform basic debates about language, in one area at least, and that in practice, existing results may offer initial tentative support for this prospect. Further, results from computational learning theory can inform arguments carried on within linguistic theory as well.

1. Introduction

It is widely believed that the scientific enterprise of theoretical linguistics on one hand and the engineering of language applications on the other are

[1] We gratefully acknowledge Michael Collins's aid in drafting section 5 of this paper. Antecedents of this paper were presented at the Workshop on Computational Linguistics, University College, London, 2004; Computational Linguistics in the Netherlands 2004, Leiden; the Linguistics Colloquium, University of Illinois at Urbana-Champaign, 2005; Construction of Meaning: Stanford Semantics and Pragmatics Workshop, Stanford University, 2005; the 2005 LSA Summer Institute, Cambridge, MA; ESSLI 2005, Edinburgh; the Computer Science Colloquium, University of Essex, 2006; the Cognitive Science and Linguistics Colloquium, Carleton University, Ottawa, 2006; and the Linguistics Department Colloquium, Stanford University, 2006. The first author presented some of the proposals contained in this paper in his Ph.D Research Seminar in the Philosophy Department, King's College, London, during the second semester, 2006. We are grateful to the participants of these forums and to three anonymous reviewers for the *Journal of Linguistics* for much useful feedback, some of which has led to significant modifications in our proposals. We would also like to thank Joan Bresnan, Alex Clark, Eve Clark, Jennifer Cole, Jonathan Ginzburg, Ray Jackendoff, Dan Jurafsky, Ruth Kempson, Chris Manning, Fernando Pereira, Steve Pinker, Geoff Pullum, Ivan Sag, Richard Samuels, Barbara Scholz, Gabriel Segal, Richard Sproat, and Charles Yang for helpful discussion of many of the ideas proposed in this paper. We are indebted to the Stanford Symbolic Systems Program, Ivan Sag, and Todd Davies for providing a welcoming venue for completion of the paper. The first author is grateful to the Leverhulme Foundation for supporting his work there. The first author's research was supported by grant number RES-000-23-0065 of the Economic and Social Research Council of the United Kingdom. Needless to say, we bear sole responsibility for these ideas.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

separate endeavors with little to contribute to each other at the moment in the way of techniques and results. In this paper, we explore the possibility that machine learning approaches to natural-language processing (NLP) being developed in engineering-oriented computational linguistics (CL) may be able to provide specific scientific insights into the nature of human language. We argue that, in principle, machine learning (ML) results could inform basic debates about language, in one area at least, and that in practice, existing results may offer initial tentative support for this prospect. Further, results from computational learning theory can inform arguments carried on within linguistic theory as well.

A basic conundrum of modern linguistic research is the question of how natural languages can be acquired. It is uncontroversial that the learning of a natural language (or of anything else) requires some assumptions concerning the structure of the phenomena being acquired. So, for example, consider a computer program that is designed to learn to classify e-mails as spam or non-spam. One could assume that spam is identified by the magnitude of a weighted average over a set of indicator features f_i, each of which specifies whether a given word w_i appears in the e-mail. We could then use a corpus of e-mail messages, each marked as to their spam status, to train the weights appropriately. Such a learning approach (essentially perceptron learning over the features) assumes (erroneously, of course, but often effectively) that the phenomenon being investigated, spam status, is characterizable independently of any properties of the e-mail but those manifested in the word-existence features, and that their impact on spam status is strictly linear. Alternatively, some other set of features, method for computing their combination, and so forth might be used, reflecting other assumptions about the nature of the phenomenon being investigated.

In the field of machine learning, the prior structure imputed to the phenomenon is referred to as the MODEL, which allows for variation in a set of parameters that may be of different sorts – continuous or discrete, fixed or combinatorial. Learning algorithms are procedures for setting these parameters on the basis of samples (OBSERVATIONS) of the phenomenon being acquired. The success of the algorithm relative to the observations can be verified by testing to see if it generalizes correctly to unseen instances of the phenomenon. The role of the model and algorithm is to provide a learning BIAS.[2] In linguistics, this prior structure is sometimes referred to as UNIVERSAL GRAMMAR (UG), the innate (that is experientially prior) aspect of the human language endowment that allows (biases) language acquisition. In the sequel,

[2] This use of the term BIAS has no pejorative import. Rather, it expresses the fact that the model biases learning according to a priori expectations.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

we will uniformly use the term LEARNING BIAS for this model or universal grammar aspect.[3]

It is uncontroversial, because it is definitional and therefore trivial, that human language acquisition must have some bias or innate component in this technical sense.[4] There is, however, a non-trivial open question – some would call it the key question in linguistics – as to the detailed structure of this bias, the NATURE OF UNIVERSAL GRAMMAR. Which aspects are derivable from general elements of cognition and which are task-specific properties of natural-language acquisition? In the former case, we might expect simpler models, more uniform in structure, and in the latter case, structures that are more complex, idiosyncratic, and highly articulated. A primary point of this paper, however, is that THERE IS NO SUCH THING AS A NO-BIAS MODEL.

All things being equal, methodological simplicity leads us to prefer an uncomplicated, uniform, task-general learning model, which we will term a WEAK-BIAS model in contrast to STRONG-BIAS models that are highly articulated, non-uniform, and task-specific.[5] But all things might not be equal. Weak-bias models might not be, and it is often argued are not, sufficient for the purpose of acquiring linguistic knowledge. Variations of such arguments have been advanced under the name POVERTY OF THE STIMULUS. They assert that without a certain task-specific complexity to the learning bias, the training data are insufficient to permit acquisition of a model that generalizes appropriately to the full range of unseen instances of some phenomenon under investigation. It is a predominant view in linguistics that the complexity of the natural languages that adults speak requires much of the

[3] We take bias in this sense to be a feature of a learning algorithm, and we do not make any assumptions that it is an innate, biologically determined property of human learners. We also do not claim that learning algorithms that provide computationally viable procedures for grammar induction correspond to psychological processing mechanisms. We treat the bias of a successful learning algorithm as specifying a universal grammar only in the sense that it characterizes the hypothesis space of learnable languages for that procedure.

[4] Nonetheless, one sees expressions of the view that statistical learning methods are somehow faulty for lacking any bias. For instance:

> 'But a more serious problem with DDL [data-driven learning], both present and future, has to do with the wild statistical disparities between what is presented to children and how children actually learn. As pointed out by Fodor & Pylyshyn (1988) and others, a DDL model without innate knowledge, or learning priors, can do nothing but recapitulate the statistical distributions of adult input. But children often learn their languages in ways that clearly defy such distributions.' (Legate & Yang 2002)

In fact, without 'learning priors', not even the statistical distribution can be recapitulated, any distribution being based on the structure of its parameters. Without a learning prior, a learning method can merely recapitulate the training data, without generalizing to a distribution, and even such recapitulation constitutes a bias (against generalization). The issue, of course, is whether the innate knowledge or learning prior is strong or weak.

[5] The range from weak to strong bias is, of course, a continuum, not a dichotomy. However, we will continue to use the terms loosely to distinguish models that fall at the respective ends of the spectrum.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

intricate structure of language to be innate, given the limited evidence to which children are exposed. Chomsky (2000: 6f.) summarizes this view:

> A careful look at the interpretation of expressions reveals very quickly that from the very earliest stages, the child knows vastly more than experience has provided. That is true even of simple words. At peak periods of language growth, a child is acquiring words at a rate of about one an hour, with extremely limited exposure under highly ambiguous conditions. The words are understood in delicate and intricate ways that are far beyond the reach of any dictionary, and are only beginning to be investigated. When we move beyond single words, the conclusion becomes even more dramatic. Language acquisition seems much like the growth of organs generally; it is something that happens to a child, not that the child does. And while the environment plainly matters, the general course of development and the basic features of what emerges are predetermined by the initial state. But the initial state is a common human possession. It must be, then, that in their essential properties and even down to fine detail, languages are cast in the same mold.

An influential approach to investigating this key question has been to construct linguistically motivated infrastructures – representations of linguistically-oriented parameters and constraints – that purport to capture universal aspects of language as proposals for the learning bias. Chomsky (2000: 8) describes this Principles and Parameters (P&P) model in the following terms:

> We can think of the initial state of the faculty of language as a fixed network connected to a switch box; the network is constituted of the principles of language, while the switches are options to be determined by experience. When switches are set one way, we have Swahili; when they are set another way, we have Japanese. Each possible human language is identified as a particular setting of the switches – a setting of parameters, in technical terminology. If the research program succeeds, we should be able literally to deduce Swahili from one choice of settings, Japanese from another, and so on through the languages that humans acquire. The empirical conditions of language acquisition require that the switches can be set on the basis of the very limited properties of information that is available to the child.

We argue that the general machine learning techniques now being used with some success in engineering-oriented computational linguistics provide an alternative approach to this question, and one that could provide quite a different answer. In particular, to the extent that a given machine learning experiment is successful in acquiring a particular phenomenon, it shows that the learning bias that the model embodies is sufficient for acquisition of that phenomenon. If, further, the bias is relatively weak, containing few

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

assumptions and little task-specificity, the experiment elucidates the key question by showing that arguments against the model based on its inability to yield the relevant linguistic knowledge are groundless. (Of course, other arguments against the model might still apply.)

In addition, we argue that theoretical results in machine learning – more specifically, results from computational learning theory – give further support for alternatives to the P&P model. Computational learning theory addresses issues of learnability under various mathematical assumptions: in particular, it focuses on whether learning is possible in different scenarios, and on the resources required for successful learning in terms of computation and number of training examples. We focus on theoretical results from PAC/VC learning (Probably Approximately Correct learning, with Vapnik and Chervonenkis dimensionality, discussed in section 5), although similar results are found in models of online learning, and Bayesian methods. Results from computational learning theory have underpinned much of the work in more applied machine learning, including research in engineering-oriented computational linguistics. These results also suggest that the learning framework implicit in the P&P approach – n fixed, binary parameters – is only one of a range of plausible definitions for 'UG' or 'innate bias', at least if learnability arguments alone are used to motivate arguments about the nature of UG.

To summarize, we support the view, which we take to be uncontroversial, that some form of universal grammar underlies the ability of humans to acquire language. Where it lies on the scale from weak to strong bias is contentious, however, and machine learning experiments and theory can clarify that position.

In order to provide pertinent background for our argument, we review the poverty-of-stimulus argument in section 2. We sketch the basic design of machine learning systems in section 3. We then show, using parsing as an exemplar problem,[6] how machine learning methods (hence, any experiments based on them) inherently involve learning biases, that is, claims about the nature of UG, but we demonstrate that these assumptions can be quite weak, in contrast to the richly articulated set of structures and conditions that many theoretical linguists seek to attribute to the initial state of the language acquisition device (section 4). We first look at supervised machine learning in section 4.1, focusing on the work of Collins (1999) on probabilistic parsing,

[6] We are limiting ourselves to parsing in order to go into some detail in clarifying the role of learning bias in machine learning and language acquisition. There has been a great deal of work on the application of machine learning methods to a wide variety of tasks in NLP in addition to parsing. This work has achieved impressive results in the full spectrum of NL tasks, including areas like speech recognition, morphological analysis, part of speech tagging, anaphora resolution, semantic interpretation, domain theory identification, discourse coherence, and dialogue interpretation. For an overview of some of this work in the context of the issues we are discussing here see the discussion by Lappin (2005).

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

arguing that much of the success of these methods is based on distributional bias, as opposed to the categorical bias assumed in the linguistics literature. Then in section 4.2 we consider recent work by Klein & Manning (2002, 2004) on unsupervised grammar induction, which makes even fewer learning bias assumptions than supervised learning.

In addition to empirical work on machine learning for natural-language processing, theoretical research can elucidate language acquisition arguments as well. We summarize pertinent results from computational learning theory in section 5, highlighting their import for linguistic theorizing. In section 6 we focus on a number of important distinctions between the concept of parameter current in theoretical linguistics on one hand and in probabilistic language modeling on the other. We suggest that while the former may well be problematic on empirical grounds, the latter provides a well motivated basis for a weak-bias model of UG that can, in principle, support a computationally viable theory of language acquisition.

Finally, in section 7 we discuss the implications of the surprisingly good performance of machine learning in grammar induction for theories of grammar and language acquisition. We suggest that arguments from the insufficiency of weak-bias models may not provide motivation for elaborate notions of UG. Most importantly, however, regardless of the particular conclusions one draws from the current contingent state of machine-learned linguistic performance, the more general point remains that results of such experiments could in principle falsify stimulus-poverty arguments for strong learning bias. In this way at least, the methods of engineering-oriented computational linguistics can inform theoretical linguistics.

2. Poverty-of-stimulus arguments for a strong learning bias

Linguists have argued for a highly articulated task-specific language acquisition device to account for the speed and efficiency with which humans acquire natural language on the basis of the relevant evidence, regarded intuitively to be sparse. In our terms this involves assuming a strong learning bias that encodes a rich set of constraints on the properties of the object to be acquired from the data. The strong-bias view relies primarily on the poverty-of-stimulus argument to motivate the claim that a powerful task-specific bias is required for language acquisition. According to this argument, the complex linguistic knowledge that a child achieves within a short time, with very limited data, cannot be explained through general learning procedures of the kind involved in other cognitive tasks.

The argument can be summarized as follows: There is insufficient evidence in pertinent language observations to accurately induce a particular general phenomenon regarding language with a weak-bias learning method alone. The particular phenomenon appears in language. Therefore, language

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

observations are not used with a weak-bias learning method alone in inducing the generalization; strong bias is required. This is, at root, a 'What else could it be?' argument. It asserts that, given the complexity of grammatical knowledge and the lack of evidence for discerning its properties, we are forced to the conclusion that much of this knowledge is not learned at all but must already be present in the initial design of the language learner.

In part, the basis for this claim is the assumption that first language learners have access to very limited amounts of data (the 'pertinent language observations'), largely free of negative information (corrections), and thus inadequate to support inductive projection of grammars under the attested conditions of acquisition without the assumed bias.

The claims we make here are independent of the issue of what exactly the data available to the language learner are. Whatever they are determined to be, machine learning can potentially be used, as argued below, to support their sufficiency for learning of the pertinent phenomena.

Nonetheless, it is worth digressing to mention that particular aspects of this view of data poverty have been increasingly subject to challenge. With respect to the absence of negative data, Chouinard & Clark (2003) provide a detailed cross-linguistic study of several language acquisition corpora in the CHILDES collection (MacWhinney 1995) suggesting that, in fact, there is a substantial amount of negative evidence available to children, and this evidence plays a significant role in the acquisition process. They found that in the initial phase of language learning (2–2.5 years) 50–67% of the child's errors are explicitly reformulated by parents into congruent contrast utterances in which the error is replaced by a correct form. They also show that children attend to these corrections in a high proportion of cases. Saxton et al. (2005) offer evidence that adult clarification requests significantly facilitate the transition from ill-formed to correct syntactic structures in child grammar induction. Negative evidence greatly facilitates data-driven grammar induction by significantly reducing the search space of possible rules and constructions. It effectively reduces language acquisition to an instance of supervised learning. Similarly, contextual information available to the language learner can, in principle, provide further constraints on the learning process.

The lack of pertinent data for particular phenomena has also been questioned. For example, English auxiliary inversion for question formation has previously been cited (among others, by Crain 1991) as a phenomenon that could not be learned from the linguistic data available to the child by virtue of the lack of exposure to the crucial examples, and so had to be attributed to the bias of the language acquisition model. Contra this claim, Pullum & Scholz (2002) suggest that careful examination of the linguistic data to which children are exposed reveals that it is far richer than poverty-of-stimulus theorists suggest, finding attested examples of auxiliary fronting of the sort previously assumed to be absent.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Further, even if the putatively crucial examples are absent, learning might still be possible. Scholz & Pullum (2006), citing work by Lewis & Elman (2002) observe that when important facts concerning the relative distributional frequency of certain interrogative forms are factored into the input data, appropriate generalizations concerning English question formation can be predicted by straightforward ML inference techniques.

Clark & Eyraud (2006) propose an algorithm for learning a context-free grammar that handles auxiliary fronting, and sequences of auxiliaries and modals correctly, where this algorithm can produce the rules of the grammar on the basis of a very small set of positive data. Here, the only bias is the allowance for the possibility of hierarchical structure implicit in the choice of context-free grammars. Whether this bias is language-specific or not is controversial; in any case, even if thought to be language-specific, it is quite weak.

Indeed, the mere possibility of expressing hierarchical structure in a grammar does not guarantee its utility. Nonetheless, Perfors et al. (2006) demonstrate that the posterior probability of a simple hierarchical grammar exceeds that of a regular (that is, non-hierarchical) grammar of similar coverage. This again 'suggests that there may be sufficient evidence in the input for an ideal rational learner to conclude that language is structure-dependent without having an innate language-specific bias to do so'.

To summarize the digression, claims that the data available to a language learner are missing crucial instances (negative data, critical pairs) may be premature. Nonetheless, as noted above, our claim about the potential for ML methods to clarify the status of poverty-of-stimulus arguments is independent of the exact nature of the evidence.

It is important to distinguish poverty-of-stimulus from other sorts of arguments that might be advanced in favor of a particular theory of grammar. One might, for example, argue that a proposed system of grammar provides the best theory for capturing the syntactic (and other grammatical) relations exhibited by phrases and sentences in a language, or across languages. Such a theory is motivated by its coverage of the observed properties of a language or a set of languages.[7] Alternatively, one might argue for a given theory of grammar on the basis of consistency with psycholinguistic results concerning adult processing of language. However, the components of the theory need not be included in the bias of the language acquisition model unless they cannot be projected from the input data on the basis of a weaker bias. It is necessary to motivate a claim of irreducibility of this sort through independent (non-)learnability evidence for the conditions of the grammar.

[7] Achieving wide coverage descriptive adequacy for the generalizations of a theory of grammar is, in general, far more difficult than is often assumed. Postal (2004) shows that many (perhaps most) of the conditions and constraints that have been proposed at one time or another as elements of UG admit of substantial counter-evidence.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

3. Review of machine learning

Engineering-oriented computational linguistics is the study of methods for the construction or improvement of useful artifacts that manipulate natural language. Examples of such artifacts include speech recognizers, machine translation systems, natural-language interfaces to databases or software systems, and information retrieval systems.

A relatively standard approach has emerged for dealing with many of the problems that arise in this area. A phenomenon is characterized as a function from some inputs to outputs: speech signals to natural-language transcriptions (speech recognition), English sentences to their labeled bracketings (parsing), French sentences to English translations (machine translation), Hebrew text to Hebrew text annotated with vowel diacritics (diacritic restoration), and so on. A large sample (corpus) of the function is collected, consisting of sampled inputs with corresponding outputs. The corpus is taken as a benchmark of correctness for determining the function. It is divided into two subsets, a TRAINING SET and a TEST SET.

A parameterized model characterizing a HYPOTHESIS SPACE of possible functions is designed. A LEARNING ALGORITHM is applied to select one instance of this space by setting the parameters on the basis of the training set. The particular instance that is selected defines the learned function.[8]

The success of learning is measured by testing the function on the test set. To the extent that the outputs of the learned function resemble the test samples' annotated outputs, the learning algorithm has successfully generalized the training samples. Resemblance can be specified in various ways. If we require identity between the output of the learned function and the benchmark, the percentage of the test set for which identity does not hold is the ERROR RATE. This condition gives a particularly stringent metric of similarity. More liberal notions than strict identity are frequently invoked to obtain more realistic measures of the similarity between the learning algorithm's outputs and the training set gold standard. The details are inessential.

Suppose, for example, that the phenomenon to be characterized is English grammaticality. It can be characterized as a function from strings to an output bit that specifies whether the string is or is not grammatical. A corpus of such a function might, in general, be difficult to acquire, but a sample of positive instances can be easily generated by taking a corpus of English text. All the elements are strings that map to the 'grammatical' bit value. We use this corpus as the training set; other examples of grammatical and ungrammatical text can then serve as a test set.

[8] The model might incorporate weak or strong bias, and the algorithm might be quite general or highly task-specific, in theory. However, most work to date has proceeded conservatively along these lines, adding biasing structure only grudgingly and using standard statistical learning techniques, less for principled reasons than for algorithmic simplicity.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

There are many possible models and learning algorithms that we could use to learn this function. One particularly simple one is the n-th order markov model, discussed, inter alia, by Shannon (1948). A probability distribution over the strings of a language can be characterized as the product of conditional probabilities:

$$P(w_1 \cdots w_c) = \prod_{i=1}^{c} P(w_i | w_1 \cdots w_{i-1})$$

We can approximate these conditional probabilities by assuming that each word is conditionally independent of those occurring n or more words earlier given the intervening $(n-1)$-word context.

$$P(w_i | w_1 \cdots w_{i-1}) \approx P(w_i | w_{i-n+1} \cdots w_{i-1})$$

We then take the space of possible functions to be those characterized by a parameter for each such conditional probability, one for each n-gram (sequence of n words) in the language, together with a single threshold probability to differentiate strings as grammatical (above the threshold) or ungrammatical (below the threshold). In particular, we associate with each n-gram $w_1 \ldots w_n$ the probability that its final word w_n occurs in the context of the initial $n-1$ words. This probability $P(w_n \mid w_1 \ldots w_{n-1})$ can be estimated from the training corpus in many ways, most simply by the MAXIMUM LIKELIHOOD ESTIMATOR (MLE):

$$P(w_n | w_1 \cdots w_{n-1}) \approx \frac{c(w_1 \cdots w_n)}{c(w_1 \cdots w_{n-1})},$$

where $c(w_1 \ldots w_k)$ is the count in the training sample of occurrences of the k-gram $w_1 \ldots w_k$. The probability of a new sentence can then be estimated by multiplying these conditional probabilities for each of the overlapping n-grams in the sentence, and the output bit can be determined by thresholding this probability. Such a reduction of grammaticality to sufficient probability is appealing in its simplicity even if dramatic in its naivete.

An apparent problem with this approach, noted as early as *Syntactic structures* (Chomsky 1957), is that it inherently fails to distinguish between sentences that are ungrammatical and those that are infelicitous. The two strings

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless.

likely have identical estimated probabilities (or at least did before 1957) for any n-gram MLE estimated model. For $n=1$, they contain the same unigrams, hence the same product. For $n>1$, they each contain at least one n-gram (and perhaps more) that would be unattested in any (pre-1957) training corpus. The MLE estimate for the probability of such unattested n-grams is 0, making the product probability for both sentences 0 as well,

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

and therefore identical. Thus, there is no value for *n* for which such nonsense sentences and non-sentences can be distinguished.

This may be, and in Chomsky's case was, taken to be an argument against the possibility of statistical learning of grammaticality judgments. Chomsky (1957: 17, fn. 4) states:

> One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have obvious flaws. Notice, in particular, that for any *n*, we can find a string whose first *n* words may occur as the beginning of a grammatical sentence S_1 and whose last *n* words may occur as the ending of some grammatical sentence S_2, but where S_1 must be distinct from S_2. For example, consider the sequences of the form 'the man who … are here', where … may be a verb phrase of arbitrary length. Notice also that we can have new but perfectly grammatical sequences of word classes, e.g., a sequence of adjectives longer than any ever produced in the context 'I saw a – house'. Various attempts to explain the grammatical-ungrammatical distinction, as in the case of (1), (2), on the basis of frequency of sentence type, order of approximation of word class sequences, etc., will run afoul of numerous facts like these.

But in fact the argument shows no more than that the simple MLE *n*-gram approach to the task fails. It has long been known that MLE estimates can overfit, underestimating the probabilities of events rare or absent in the training data. Methods for overcoming the problem go under the rubric of SMOOTHING. Pereira (2000) points out that smoothing techniques, introduced by Good (1953), permit the assignment of probability values to unobserved linguistic events. When enriched with smoothing, statistical modeling of NL learning can deal effectively with sparse data. Pereira shows that a smoothed bigram model actually provides a FIVE ORDER OF MAGNITUDE difference in probability for these two sentences.

Importantly, Pereira's smoothed model has far fewer parameters than the MLE bigram model. Fewer parameters are needed in this case by virtue of the greater bias implicit in the model.[9] In particular, the model instantiates

[9] One must beware of simplistic conflations of fewer parameters, more bias, and easier learning. As we point out in section 5, the situation is quite a bit more complex, and the field of computational learning theory has made significant progress in formally capturing a notion of intrinsic difficulty of learning of different hypothesis spaces, for instance through notions like VC dimension. In this particular example, the two models (MLE bigram and class-based bigram) are sufficiently similar and the disparity in number so large that a crude counting of the parameters is reasonable, but in general more sophisticated analysis is necessary. Indeed, the characterization of intrinsic learning complexity for hypothesis spaces is one of the insights of computational learning theory that is of potential value to linguistic theory; we review it in section 5.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

the claim that the probability of a word depends not on the identity of the prior word as in a bigram model but on its CLASS, and, contrary to Chomsky's (1957) claim, this model does predict a significant distinction in probability between the two example sentences. Pereira uses the following alternative bigram approximation:

$$P(w_i|w_{i-1}) \approx \sum_c P(w_i|c)P(c|w_{i-1})$$

(Since instances of these further probabilities $P(w_i \mid c)$ and $P(c \mid w_i)$ are not directly observed, we can no longer use simple observed frequencies, but other general statistical procedures suffice for inducing these parameters.)

However, no stipulation is made as to the set of classes or the class of particular words; these are learned from the training data simultaneously with the class-based bigram parameters. The smoothed model manifests a linguistically motivated bias, namely the uncontroversial claim that WORDS FORM DISTRIBUTIONAL PATTERNS BY VIRTUE OF FALLING INTO CLASSES. The surprising fact is that such an impoverished bit of additional bias has such far-reaching consequences for a learning method to capture important linguistic phenomena.

There are three lessons here. First, the inapplicability of one weak-bias learning method does not rule out all such methods. Though MLE *n*-grams cannot distinguish the two sentences, little additional bias is needed for a learning method to achieve this effect, and that bias seems hardly language-specific.

Second, the applicability of a weak-bias learning method does not guarantee its general adequacy. To forestall confusion, we emphasize that thresholded smoothed *n*-gram models of the type Pereira uses in his example are completely implausible as a model of language on many grounds – empirical coverage, descriptive complexity, appropriateness for interpretation, among others. No one would argue that they can completely characterize grammaticality. The inadequacy of the model is immaterial to our point, which is simply to show that there is sufficient information in the corpus of data to learn a distinction between these sentences that are putatively indistinguishable without language-specific bias.We must look for evidence for the language structure for the sentences in question from some other source. But this case does indicate that information-theoretic approaches to modeling grammar are not vulnerable to the simple arguments for rich innate structure that have been widely accepted in the linguistics literature for the past forty-five years.

Finally, and most importantly, all machine learning methods introduce bias. Even the *n*-gram model, paradigmatic of a weak-bias language modeling method, makes the assumption that words more than *n* words apart are conditionally independent. Indeed, without bias, there is no generalization, and generalization is the OBJECT of learning.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

4. WEAKLY BIASED MACHINE LEARNING METHODS FOR PARSING

To illustrate our claim that machine learning methods can, in principle, provide linguistic insight, we consider parsing, the recovery of the hierarchical syntactic structure of a sentence. Discussing a specific natural-language task allows us to highlight the important properties of machine learning that bear on the issues at hand: the distinction between supervised and unsupervised methods, the implicit grounding of all methods in a linguistic theory however impoverished, and the explicit ability to determine appropriate generalization to new data.

To test if a system (including a person) can parse, we need to compare the performance of the system against correct parses of sentences. Of course, what constitutes a correct parse is a theory-relative notion. For concreteness, we use a proxy for real parsing, namely, replicating the parses of the Penn Treebank (Marcus 1993). Recall that machine learning methods involve inducing an approximation of a function. The function at hand here, then, is the function from a string to its syntactic analysis as given in the Treebank. We use the Treebank itself for both training and testing samples.

As we noted in section 3 we can use different similarity measures to evaluate the performance of a learning algorithm, in this case a grammar induction system for parsing, against a gold standard. Requiring only consistency of bracketing of the parser's output relative to the annotated structure of the corpus yields the NON-CROSSING-BRACKETS RATE metric. Many other possibilities have been proposed (Goodman 1996). We will measure the quality of the approximation via F-MEASURE, a weighted combination of precision and recall over constituents (labeled brackets).[10] Such a measure tracks how often the reconstructed brackets are correct and how often the correct brackets are reconstructed, with performance ranging from zero (worst) to one (best), and often expressed as a percentage.[11]

If the observations on which the system is trained include input (the sample strings) annotated with the intended parse (that is, their benchmark output structure), the learning task is SUPERVISED. (One can imagine a

[10] In particular, we report values of the F1-measure here, in which precision and recall are equally weighted.

[11] Some critics of the view that machine learning can provide a viable model of human grammar acquisition have argued that an ML system learns only from a corpus of grammatical sentences and so is limited to recognizing well-formed phrases. See, for example, Carson Schutze's contributions of 20 April and 5 May 2005, on the LINGUIST LIST, to the discussion of Richard Sproat and Shalom Lappin, 'A challenge to the Minimalist community', LINGUIST LIST, April 5, 2005. This claim is misconceived. The parser that an ML system produces can be engineered as a classifier to distinguish grammatical and ungrammatical strings, and it can identify the structures assigned to a string under which this distinction holds. Such a classifier can also be refined to provide error messages that identify those properties of an unsuccessful parse that cause a sentence to be identified as ungrammatical.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

supervisor or teacher telling the student the right answer for each training problem.) When the observations include only the input, but not the output, the learning is UNSUPERVISED.

4.1 *Supervised learning of parsing*

The space of all parameter settings in the model for grammar induction determines the hypothesis space of learnable parsers. This space determines the expressivity of the learned model and so specifies an important part of its bias. In the linguistics literature, the bias induced by the parameter structure of the grammar has been thought of categorically, that is, as governing what is or is not a possible language. It is also quite possible, and frequently beneficial, to introduce distributional bias into the model, that is, the specification of an a priori probability distribution over the hypothesis space, and hence over the possible languages. (We will make this notion of distributional bias more precise in section 5.4.)

The notion of distributional bias, in contrast to categorical bias, has not been widely used in linguistics. A distributional bias does not directly limit the class of languages that are learnable, but rather affects the likelihood with which a particular language will be chosen to generalize the observations (or the amount of evidence required to select a particular language). A simple example may clarify the idea. Suppose we want to learn how many sides are on a multi-sided sequentially numbered die based on observations of rolls. We select an infinite hypothesis space of possibilities in which all die sizes are possible. One model might assume that all numbers of sides are equally likely a priori; another might assume that dice tend to have about six sides, with more or fewer being less likely the farther from this norm. Then, after seeing a certain set of observations, say 2, 4, 1, 3, 2, the most likely number of sides under the first (uniform) model would be 4, but under the second (non-uniform model) it might be 5. Categorical bias can be seen as a limit case of distributional bias where the distribution assigns zero probability to certain elements of the hypothesis space.

Indeed, the two kinds of bias can be measured on the same scale, representing the dimensionality of the learning problem. Distributional bias can be effective in allowing learning, even without restricting the class of possible learned languages. This is in contrast to the assumption, sometimes found in the linguistics literature, that without a categorical constraint on the class of languages, perhaps even reducing it to a finite set, learning cannot succeed.

By way of example of the weak linguistically-oriented bias found in current engineering-oriented models, and the power of distributional bias, we consider the sequence of parsers built by Collins (1999). As described above, the problem at issue is the reconstruction of a syntactic analysis from samples. We must assume, of course, that strings have structure; this is

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

implicit in the statement of the problem. Beyond that, all else is bias. For instance, we can assume that the probabilities for the expansion of a node are completely random, making no assumptions as to their distribution. In that case, no generalization from the training data is possible. In order to get started, let us assume that the constituents that we see are representative of the function outputs in at least the following way: the phrase types (non-terminal labels N) are generated according to a fixed distribution $P(N)$. With this assumption, we can use the training data to learn an approximation of the distribution in a variety of ways. The maximum likelihood estimate is merely the observed probability in the training sample. Generating trees in this way corresponds to selecting a random tree structure and decorating it with labels according to this fixed distribution. This would undoubtedly work better than nothing at all, though not much.

The normal starting point for statistical parsing models, therefore, conditions the probability of a child non-terminal sequence on that of the parent. We have conditional probabilities of the form $P(X_1 \ldots X_n|N)$ for each non-terminal N and sequence $X_1 \ldots X_n$ of items from the vocabulary of the grammar (the non-terminals plus the words of the language). We also need a probability distribution over the label of the root of the tree $P_s(N)$. We can then learn these parameters from a training corpus, perhaps by MLE. By structuring the model in this way, we are assuming that the trees were generated by a probabilistic context-free grammar (PCFG). The conditional probabilities $P(X_1 \ldots X_n|N)$ correspond to probabilistic parameters that govern the expansion of a node in a parse tree according to a context-free rule $N \rightarrow X_1 \ldots X_n$. (Because the languages generable by context-free grammars are limited, the hypothesis space of this model must be limited as well; this part of the bias is categorical.) Such models work poorly, if passably, as statistical parsers. Systems built in this way show F-measures in the low 0.7 range, that is, there is a match of about 70% between the constituents of the gold standard and those of trees that the parser generates for the training data.

To improve performance, we need to add either further expressivity or additional bias by effectively reducing the dimensionality of the problem, and so removing degrees of freedom in hypothesis selection. We will do both. Further expressivity is needed in part because the context-freeness assumption implicit in the model fails to hold – the probability of a child sequence is manifestly not independent of everything but the parent label. We can further articulate the distributions in various ways. Collins does so by conditioning on lexical material, introducing the notion of a head word, and replacing non-terminals with non-terminal/head pairs in the model. The probability distributions are now of the form $P_s(N/h)$ and $P(X_1/h_1 \ldots H/h \ldots X_n/h_n|N/h)$.

Because this dramatic increase in dimensionality makes the function more difficult to learn (prone to overfitting), we assume that these numerous

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

conditional probabilities themselves have further structure, in particular, that the head constituent is conditioned only on the parent, and each sibling of the head constituent is conditioned only on the head, the parent, and the side of the head on which it falls. These further independence assumptions allow the probabilities $P(X_1/h_1 \ldots H/h \ldots X_n/h_n|N/h)$ to be computed as the product of a far smaller and simpler set of conditional probabilities. In this way additional bias is introduced.

Bias could have been introduced through further categorical constraints, by allowing only certain context-free rules that follow from detailed linguistic analysis (equivalently, reducing the number of conditional probability parameters). Instead, we adopt a more conservative approach here (as is characteristic of engineering-oriented CL), allowing all such rules, but adjusting the prior distribution of probabilities over them through the structure of parameters of the previous paragraph.

By proceeding in this way it is possible to construct models with F-measure performance of approximately 88% (Collins 1999), with more recent work improving this further to approximately 91% (Charniak & Johnson 2005). This level of performance is quite impressive, if not yet at the level of human parsing. Nonetheless, one might wonder to what extent these models can be characterized as weak-bias methods. They certainly import structure consistent with our understanding of language. The design of the models they use encodes the requirements that sentences have hierarchical (tree) structure, that constituents have heads which select for their siblings, and that this selection is determined by the head words of the siblings. However, these conditions on language do not express the complex view of universal grammar assumed in much of current linguistic theory. One might have thought that it is necessary to posit a parameter governing whether complements precede or follow their heads in a given language. In fact, no such parameter is part of Collins' model, which still correctly generalizes the relevant phenomena, while, in principle, allowing analyses in which head-complement order varies in linguistically unlikely ways.

The point here is that by structuring a statistical parsing model, one is making claims with linguistic import. But current ML parsers demonstrate the fact that relatively weak claims may still yield surprisingly strong performance.

At this point one might object that supervised learning has limited relevance to the problem we are considering, given that the features and structures that the machine learner acquires are already marked in the training data. In what sense, then, can supervised learning claim to induce a classifier for the recognition of complex properties from data through weakly biased models if the data is already annotated with these properties? There are two points to make in reply to this objection. First, as we have already noted, recent psycholinguistics studies of child language acquisition suggest

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

that negative evidence plays a significant role in first language learning. This evidence turns the language learning problem into a species of supervised learning. Second, current research on unsupervised learning has achieved very promising results for parsing using unsupervised ML methods. We will briefly describe some of this work next.

4.2 *Unsupervised learning of parsing*

In unsupervised learning the training corpus is not annotated with the structures that the system is intended to acquire. Instead the learning algorithm uses distributional patterns and clustering properties of more basic features in the training data to project a classifier that recognizes the objects to be learned. This is a considerably more difficult task than supervised learning, as it requires identification of a class of target objects without the benefit of prior exposure to explicit pairings of feature patterns and target objects in the input data.

Advocates of a poverty-of-stimulus argument for a strong learning bias encoded in a rich set of parameterized linguistic principles have generally characterized language acquisition as an unsupervised learning problem of a particularly stringent variety. If it turns out that this problem can be solved by an ML model with comparatively weak bias, then this result would provide insight into the nature of the initial conditions (UG) that are sufficient for language acquisition, even if one discounts the negative evidence that may be available to the human learner.

Early experiments with unsupervised grammar induction (like those reported by Carroll & Charniak 1992) did not yield encouraging results. However, recent work has shown significant progress. The grammar induction system proposed by Klein & Manning (2002) is an unsupervised method that learns constituent structure from part of speech (POS) tagged input by assigning probability values to sequences of tagged elements as constituents in a tree. They bias their model to parse all sentences with binary branching trees, and they use an Expectation Maximization (EM) algorithm to identify the most likely tree structure for a sentence. Their method relies on recognizing (unlabeled) constituents through distributional clustering of corresponding sequences in the same contexts, where a tree structure is constrained by the requirement that sister constituents do not overlap (have non-null intersections of elements).

The Klein & Manning procedure achieves an F-measure of 71% on *Wall Street Journal*: (WSJ) text, using Penn Treebank parses as the standard of evaluation. This score is impressive when one considers a limitation that the evaluation procedure imposes on their system. The upper bound on a possible F-measure for their algorithm is 87% because the Penn Treebank assigns non-binary branching to many constituents. In fact, many of the system's 'errors' are linguistically viable parses that do not conform to analyses of the

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Penn Treebank. So, for example, the Treebank assigns flat structure to NPs, while the Klein & Manning procedure analyzes NPs as having iterated binary branching. Parses of the latter kind can be motivated on linguistic grounds.

One might object to the claim that Klein & Manning's parser is genuinely unsupervised on the grounds that it uses the POS tagging of the Penn Treebank as input. They run an experiment in which they apply their procedure to WSJ text annotated by an unsupervised tagger, and obtain an F-measure of 63.2%. However, as they point out, this tagger is not particularly reliable. Other unsupervised taggers, like the one that Clark (2003) describes, yield very encouraging results, and outputs of these taggers might well permit the parser to perform at a level comparable to that which it achieves with the Penn Treebank tags.

Klein & Manning (2004) describe a probabilistic model for unsupervised learning of lexicalized head dependency grammars. The system assigns probabilities to dependency structures for sentences by estimating the likelihood that each word in the sentence is a head that takes a specified sequence of words to its left and to its right as argument or adjunct dependents. The probabilities are computed on the basis of the context in which the head appears, where this context consists of the words (word classes) occurring immediately on either side of it. Like the constituent structure model, their dependency structure model imposes binary branching as a condition on trees. The procedure achieves an F-measure of 52.1% on Penn Treebank test data. This result underrates the success of the dependency model to the extent that it relies on strict evaluation of the parser's output against the dependency structures of the Penn Treebank, in which NPs are headed by Ns. Klein & Manning report that in many cases their dependency parser identifies the determiner as the head of the NP, and this analysis is, in fact, linguistically viable.

When the dependency system is combined with their unsupervised constituency grammar, the integrated model outperforms each of these systems. In the composite model the score for each tree is computed as the product of the individual models that the dependency grammar and the constituency structure grammar generate. This model uses both constituent clustering and the probability of head dependency relations to predict binary constituent parse structure. It yields an F-measure of 77.6% with Penn Treebank POS tagging. It also achieves an F-measure of 72.9% with an unsupervised tagger (Schütze 1995).

This work on unsupervised grammar induction indicates that it is possible to learn a grammar that identifies complex syntactic structure with a relatively high degree of accuracy using a model containing a weak bias, specifically the assumption of binary branching, a non-overlap constraint for constituents, and limited conditions on head argument/adjunct dependency relations.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

A criticism might be raised that unsupervised learning does not provide a credible model of human language acquisition because children acquire their language through semi-supervised learning in rich non-linguistic contexts. Neither supervised nor unsupervised learning from corpora expresses this task. While it is clearly true that information from extra-linguistic context plays a crucial role in human language acquisition, the fact that experiments on unsupervised grammar induction from corpora are beginning to achieve good results is highly relevant to the issue that we are addressing here. By considering a more restricted learning problem in which severer conditions are imposed on input and induction than apply in child language learning, these experiments establish the computational viability of weak-bias models for the human case, where additional sources and varieties of learning input are available.

Another criticism that might be raised is that anything short of 100 percent success fails to establish the credibility of either a supervised or an unsupervised ML grammar induction device taken as a model for human language acquisition. To achieve successful coverage of the data not handled by the system one might require entirely different sorts of learning principles than those that the system applies. Taken to its ultimate conclusion this objection also undermines poverty-of-stimulus arguments for a richly articulated UG. Unless a theory of strong-bias UG accounts for all of the properties of natural language and the acquisition process, then it is subject to the same skepticism. Linguistic theories that leave some of this data unexplained are in the same position as ML systems that do not parse all of the corpora on which they are tested. The objection invokes the possibility that grammar induction, and, in fact, language acquisition in general, is nonmonotonic in that it might consist of disjoint components defined by entirely distinct sets of learning methods. In the absence of significant evidence for this possibility, the objection is not compelling. As Hume would say, it admits of no answer, but produces no conviction.

5. Insights from computational learning theory

We have argued that empirical results in machine learning applied to natural language can provide linguistic insight. In this section, we describe how results from the theoretical branch of machine learning, COMPUTATIONAL LEARNING THEORY, can inform linguistic theorizing as well.

Computational learning theory provides precise mathematical frameworks for the analysis of learning problems. The development of this theory has been central in the analysis of existing algorithms used in machine learning and in the development of new learning algorithms. Learning theory has provided profound insights into what can and cannot be learned, and the resources (for example, in terms of computation or the number of training examples) required for learning under different assumptions.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Arguments of a learning-theoretic nature are sometimes seen in linguistic theorizing, as for instance claims that restricting the set of possible languages to a finite number allows language to be learned, or that eliminating a linguistic parameter makes the system more learnable, or that the early learning theory results of Gold (1967), showing the difficulty of learning languages under certain assumptions, demonstrate the need for strong-bias learning.

In this section we argue that recent results from learning theory are highly relevant to language acquisition questions in general, and cast light on these claims in particular. Specifically, we describe how learning theory suggests alternatives to the P&P model of UG. We will emphasize the following conclusions that can be drawn from our current understanding of computational learning theory:

- More recent learning frameworks – PAC/VC, online, and Bayesian learning – should be considered as serious alternatives to Gold's framework. These methods typically consider rates of convergence, not just convergence in the limit. They allow models of learning in the presence of noise. They have been far more successful in practice when analyzing and developing machine learning methods within applied fields.
- Naive counting of parameters is not a useful guide to difficulty of learning. It is possible to learn in infinite dimensional spaces, given the right learning bias or parameter estimation method. Conversely, some finite spaces are too complex (large) for learning with a reasonable number of training examples.
- Similarly, even for finite-dimensional spaces, comparing the number of parameters is not a reasonable method for evaluating the difficulty of learning.
- Learning bias can be embodied both in the definition of the hypothesis space and in the learning algorithm. Linguistic arguments have traditionally concentrated on the former alone.
- There are non-trivial issues in learning of finite hypothesis spaces: for example, how rates of convergence depend on the size of the hypothesis space, and how a distributional prior can be defined over elements of the hypothesis space.

In order to provide backing for these points, we introduce computational learning theory approaches, describe PAC/VC learning in finite and infinite hypothesis spaces, and discuss the crucial role of distributional bias in learning.

5.1 *Computational learning theory approaches*

Recall that we characterize a phenomenon to be learned as a function from inputs to outputs. A learning method proposes a hypothesis that is thus such

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

a function. As outlined earlier, a learning approach consists of two high-level components:

- A HYPOTHESIS SPACE: The set of functions that the learning approach will consider. For example, in the P&P framework, the hypothesis space would be the set of all possible grammars under UG. If there are n binary parameters, and these parameters can be set independently from each other, the hypothesis space is finite and is of size 2^n. In the remainder of this paper we will use $\mathscr{H}$ to denote a hypothesis space.
- A LEARNING FUNCTION: A function that maps a sequence of training examples to a member of the hypothesis space $\mathscr{H}$. We will use LEARN to denote this function.

We can imagine that the learner proceeds by observing a training example o_1 and applying the LEARN function to it, generating a first hypothesis $h_1 = \text{LEARN}(o_1)$. Upon seeing the next observation o_2, the learner hypothesizes $h_2 = \text{LEARN}(o_1, o_2)$. In general, after n observations, the learner hypothesizes $h_n = \text{LEARN}(o_1,\ldots,o_n)$. Each such hypothesis is a function drawn from the hypothesis space $\mathscr{H}$, and represented by a setting of the parameters.

Given these definitions, the following issues are crucial in determining whether learning is successful under a particular choice of $\mathscr{H}$ and LEARN:

- CONVERGENCE IN THE LIMIT: In the limit, as the number of training examples goes to infinity, will the learning method 'converge' to the correct member of $\mathscr{H}$?
- RATE OF CONVERGENCE: If the method does converge, how quickly will it converge, in terms of the number of training examples required, before the optimal member of $\mathscr{H}$ is settled on?

We argue that both of these questions are critical when assessing the plausibility of a proposal for language acquisition. Convergence in the limit is a minimal requirement for any learning method. The rate of convergence is also important, as it is uncontroversially the case that the number of utterances available to a human language learner is limited.

The convergence results for a learning approach depend on (in addition to the definitions of $\mathscr{H}$ and LEARN) two further definitions: (i) a precise definition of convergence; and (ii) the assumptions that are made about the process generating the training examples. Learning in the limit (Gold 1967) was one of the earliest proposals for how to frame a theory of learning that follows this general approach. Since then several other frameworks have been introduced, some prominent ones being VC theory (Vapnik & Chervonenkis 1971); PAC learning (Valiant 1984); online learning (see Blum 1996 for a survey); and Bayesian methods (see Berger 1985 for a survey). These methods differ significantly from Gold's framework, both in terms of their definitions of convergence and the assumptions they make about how

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

training examples are generated. In practice this has led to the development of definitions of $\mathscr{H}$ and LEARN significantly different from those considered within Gold's approach.

In the framework of Gold (1967), the sequence of training examples can be any sequence defined by a Turing machine – this sequence is unknown to the learner. Convergence is defined as the very strict criterion that at some point in any such sequence LEARN must converge to a state where it continually predicts the correct member of $\mathscr{H}$.

In PAC/VC learning, by contrast, quite different definitions of convergence and data generation are used. A central assumption is that there is some probability distribution $\mathscr{D}$ over examples in the domain, and the training set is a sample of points drawn randomly from this distribution. A notion of convergence parameterized by two factors, numbers ε and δ between 0 and 1, is defined with respect to this distribution. Informally speaking (ε, δ)–convergence means that with high probability $(1-\delta)$, the learning method generates a hypothesis that is close to (within ε of) the ideal, that is, the hypothesis is PROBABLY APPROXIMATELY CORRECT (PAC).[12] Results from the literature in PAC/VC learning give bounds on the number of training examples required for (ε, δ)–convergence, or show that (ε, δ)–convergence is impossible with a finite training sample for certain definitions of $\mathscr{H}$ and/or LEARN.

The PAC notion of convergence is quite natural in the context of language learning. Neither a guarantee of convergence ($\delta = 0$) nor exact convergence ($\varepsilon = 0$) are necessary or even plausible for human language acquisition. The noisiness of actual linguistic behavior, especially at the margins, means that distinguishing between convergence to an ε of zero versus non-zero may not even be empirically possible.

Whether convergence is guaranteed within the PAC/VC framework will depend critically on the choice of the hypothesis space $\mathscr{H}$ and the learning function LEARN. In effect, the choices of $\mathscr{H}$ and LEARN implement a substantial bias in the definition of a learning method. In the next sections we give a survey of different proposals for the definition of $\mathscr{H}$ and LEARN, in each case discussing the consequences for learnability within the PAC/VC framework.

[12] More formally, for any member h of $\mathscr{H}$, we define ERROR(h) to be the probability that h makes an error on a randomly drawn example from the distribution $\mathscr{D}$. Define L^* to be the lowest value of ERROR(h) for any member of $\mathscr{H}$. We would like LEARN to return a member of $\mathscr{H}$ whose value for ERROR(h) is as close to L^* as possible. Under these definitions, (ε, δ)–convergence means that with probability at least $1-\delta$ the hypothesis returned by LEARN will have ERROR(h) $\leqslant L^* + \varepsilon$. The statement 'with probability at least $1-\delta$' refers to probability with respect to the random choice of training examples from $\mathscr{D}$. Intuitively, this statement means that there is a small probability δ, that the training sample will be a 'bad' one where LEARN fails to return a good member of $\mathscr{H}$.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

5.2 *PAC/VC learning with finite hypothesis spaces*

First, we consider results for PAC/VC learning when the hypothesis space $\mathscr{H}$ is finite. For now we assume a particularly simple definition of LEARN: We take LEARN to return the member of $\mathscr{H}$ that makes the smallest number of errors on the training sample. This definition of LEARN is often referred to as EMPIRICAL RISK MINIMIZATION (ERM; Vapnik 1998).

Within the PAC/VC framework, for finite hypothesis spaces the ERM method always leads to a learning method that converges in the limit. The rate of convergence depends directly on the size of the hypothesis space $\mathscr{H}$. More precisely, for fixed ε and δ, the number of training examples required scales linearly with the LOGARITHM of the number of functions within $\mathscr{H}$. Within the P&P framework, assuming n binary parameters, the size of the hypothesis space is $|\mathscr{H}| = 2^n$; the number of training examples required therefore scales linearly with $\log |\mathscr{H}| = n\log 2$.[13] The linear dependence on $\log |\mathscr{H}|$ as opposed to $|\mathscr{H}|$ has important consequences. If the number of training examples required scaled linearly with 2^n then learning within the P&P approach for any significantly large value of n would be completely infeasible.

The most important point here is that the SIZE OF THE HYPOTHESIS SPACE is the central factor underlying rate of convergence of the learning method. Different hypothesis spaces that have the same size should require roughly the same number of training examples for learning. This is a very useful guideline when considering the viability of alternatives to the P&P approach. As one example, in Optimality Theory each hypothesis is defined by a ranking of c constraints. The size of the hypothesis space in this case is $|\mathscr{H}| = c!$; this leads to the bound $\log |\mathscr{H}| \leqslant c\log c$. When comparing a proposal within the P&P framework with n parameters to an OT proposal with c constraints, we would expect the complexity of learning to be roughly equivalent in the two cases if $n\log 2 \approx c\log c$. In practice, $\log c$ grows slowly as a function of c, so that n and c can be the same order of magnitude.

As a less familiar example, albeit one with clear potential applications, consider SPARSE PARAMETER REPRESENTATIONS. As in the P&P approach, assume that each hypothesis is defined by the setting of n binary parameters, but where only a small number, k, of these parameters are ever set to 1 rather than 0 in the hypothesis returned by LEARN. (Equivalently, we could assume that the remaining $n-k$ parameters take some default value.) In this case the size of the hypothesis space is $\frac{n!}{(n-k)!\,k!} \leqslant n^k$. The number of training examples required scales linearly with $\log |\mathscr{H}| \leqslant k\log n$. The striking property of this representation is that the number of training examples, while linear in k, depends linearly on $\log n$ rather than n. This means that learning is feasible for very large values for n, assuming that k is relatively small. To illustrate

[13] To be exact, results from PAC learnng suggest that at most $m = \frac{\log |\mathscr{H}| - \log \delta}{\varepsilon^2}$ training examples are required for (ε, δ)–convergence for any finite hypothesis space $\mathscr{H}$.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

this, consider a P&P model with $n = 200$ parameters, and a hypothesis space which is therefore of size 2^{200}. In the sparse representation model, values of $k = 10$ and $n = 2^{20} \approx 1{,}000{,}000$ lead to a hypothesis space of size less than 2^{200}. That is, learning in a space defined by 1,000,000 parameters is no harder than one defined by 200 parameters, if in the former case we know that only 10 of the parameters take non-default values. Sparse parameter representations thus provide a simple example showing that naive counting of parameters is not a reasonable measure of the complexity of learning.

5.3 *PAC/VC learning with infinite hypothesis spaces*

Now consider the case where the hypothesis space $\mathscr{H}$ is infinite. In particular, we will assume that each member of $\mathscr{H}$ is associated with n REAL-VALUED parameters. Thus the hypothesis space is uncountably infinite. In this case, the results of the previous section are clearly not applicable. We will again consider convergence results for the ERM learning method, where LEARN simply picks the hypothesis h with the smallest number of errors on the training sample.

In the ERM method for infinite hypothesis spaces, a measure called the VC-DIMENSION (Vapnik 1998) is critical in determining whether the learning method converges in the limit, and also in determining the rate of convergence. The VC-dimension of a hypothesis space $\mathscr{H}$ is defined as the largest value of m such that there is a training sample of size m that is SHATTERED by $\mathscr{H}$. A training sample is shattered by a hypothesis space $\mathscr{H}$ if for each of the 2^m possible labelings of the sample, there is a hypothesis in $\mathscr{H}$ that assigns that labeling.[14] For example, suppose that the function to be learned maps points in two-dimensional space onto 0 and 1. Such a function assigns a single bit to each point in the real plane. A hypothesis space is a subset of all such possible functions. It might, for instance, contain just those functions that assign 1s to all points 'northeast' or 'southwest' of a designated point p in the plane, and 0s to all points in the other two quadrants. Each element of this 'northeast-southwest' hypothesis space is characterized by two real-valued parameters, the two coordinates of the designated point p that picks out the junction between the 1s quadrants and the 0s quadrants. This hypothesis space (by virtue of its impoverished expressivity) has a VC-dimension of 2, because it shatters two points, but not three, as depicted in figure 1.[15]

A hypothesis space $\mathscr{H}$ has infinite VC-dimension if for any value of m, there is some training sample of size m that is shattered by $\mathscr{H}$. Given these

[14] We assume here that the function to be learned has a range of 2, that is, it assigns a single bit to each input.

[15] As an exercise, the reader might want to verify that a 'northeast' hypothesis space, which labels points as 1s only in the northeast quadrant, has VC-dimension of 1.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

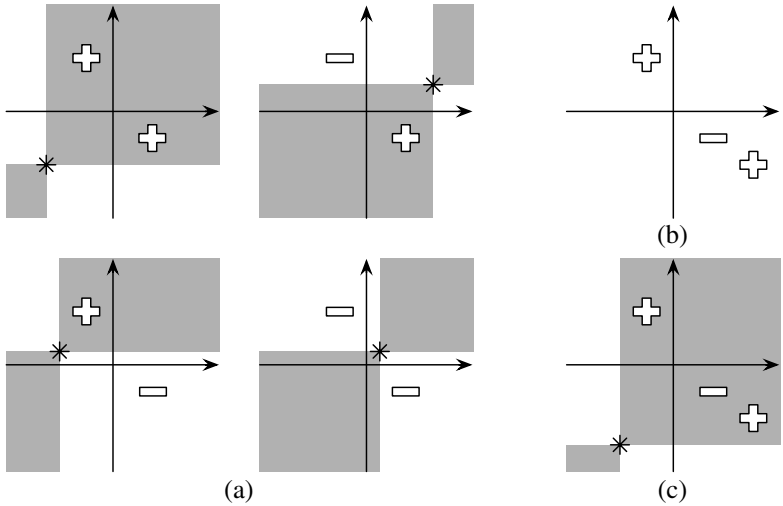


Figure 1
The 'northeast-southwest' hypothesis space labels points in the northeast and southwest (gray) quadrants defined by its single point parameter (shown as '*') as 1, and 0 in the other two (white) quadrants. (a) Two points in the two-dimensional space can be labeled with all four possible labelings. (Here, we depict a labeling of a point as a 1 or 0 by showing the point as a '+' symbol or '–' symbol, respectively.) Adding a third point in the configuration (b) cannot be labeled with the depicted labeling (a 0 surrounded by two 1s), because (c) any hypothesis that labels the outside points properly captures the inside point. It can be verified that no set of three points can be shattered by this hypothesis space. The hypothesis space thus shatters two points, but not three. The VC-dimension of the space is therefore 2

definitions, the following results hold: (i) the ERM method converges in the limit if and only if the VC-dimension of the hypothesis space is finite; and (ii) the number of training examples required is roughly linear in the VC-dimension of the hypothesis space.

There are a number of important points here. First, learning in infinite hypothesis spaces with the ERM method is possible in some cases, more precisely, in those cases where the hypothesis space has finite VC-dimension. This in itself is a surprising result: intuition would suggest that infinite hypothesis spaces might be too complex for learning to be successful. Second, the VC-dimension will be highly sensitive to the precise definition of the hypothesis space (in contrast to the results for finite hypothesis spaces, where merely knowing the size of $\mathscr{H}$ was sufficient to characterize learnability). For example, even though the 'northeast-southwest' hypothesis space is infinite, learning in this space converges quickly because its VC-dimension is low. A simple linear (hyperplane) classifier with n parameters has VC-dimension equal to n. The number of training examples will then scale linearly with the number of parameters in the model. In contrast, it is possible to define hypothesis spaces that have a single real-valued parameter, and yet have

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

infinite VC-dimension and are therefore not learnable through the ERM method. (Vapnik 1998 provides one such example.)

5.4 *PAC/VC learning and distributional bias*

Thus far we have considered a relatively simple definition of LEARN, based on the ERM principle, where LEARN simply returns the member of $\mathscr{H}$ with the smallest number of errors on the training sample. In this approach, the only form of learning bias comes in the choice of $\mathscr{H}$. We now consider a powerful generalization of this method, where a DISTRIBUTIONAL BIAS over the members of $\mathscr{H}$ can be used by LEARN.

We provide an example of this method for finite hypothesis spaces. Consider the following definition of LEARN. In addition to $\mathscr{H}$, we assume that the learning algorithm has access to some probability distribution $P(h)$ over the members of $\mathscr{H}$. The distribution $P(h)$ can be arbitrary, the only restriction being that it must be chosen independently of the training examples. We now consider a modified definition of LEARN. The function LEARN returns a hypothesis h in $\mathscr{H}$ that attempts to simultaneously minimize two factors: the number of errors that the function makes on the training sample (as for the ERM method), which we call ERROR$'(h)$, and the negative log probability $-\log P(h)$ of the hypothesis. The former quantity is, as before, a measure of how well the hypothesis fits the training sample. The latter quantity is a measure of the 'complexity' of the hypothesis h, phrased in terms of a penalty for choosing an unlikely h. When $P(h)$ is 1, the penalty is 0 (since $\log 1 = 0$). As $P(h)$ decreases to 0, the penalty increases to infinity.[16]

Intuitively, $P(h)$ will assign an a priori bias towards choosing or not choosing h in the learning process. In particular, a higher value of $P(h)$ reflects a stronger bias towards choosing h as the output of LEARN. As for the ERM method, it can be shown that the method described above converges in the limit. However, the analysis of the RATE OF CONVERGENCE is different from that of the ERM method. If the correct hypothesis h has a relatively high value for $P(h)$ – in other words, if the a priori bias is well-tuned to the learning problem in that it places high weight on the correct member of $\mathscr{H}$ – then convergence can be substantially faster than that for the ERM method with the same hypothesis space $\mathscr{H}$. The choice of $P(h)$ gives an additional degree of flexibility in defining a learning bias for the function LEARN WITHOUT CHANGING THE HYPOTHESIS SPACE $\mathscr{H}$ AT ALL, and this bias can

[16] As one example of minimizing both of these factors, LEARN might return the member of $\mathscr{H}$ that minimizes

$$\text{ERROR}'(h) + \sqrt{\frac{-\log P(h)}{m}},$$

where m is the number of training examples.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

substantially increase the rate of convergence of the method, if the values for $P(h)$ are well chosen.

Similar methods can be used for infinite hypothesis spaces. In this case we again identify each $h \in \mathscr{H}$ with a vector of n real-valued parameters $h_1, \ldots, h_n$. Learning theory results for neural networks (Bartlett 1998) and support vector machines (SVMs) (Cortes & Vapnik 1995) essentially suggest replacing the factor $-\log P(h)$ with the quantity $\sum_{i=1}^{n} h_i^2$, which is the Euclidean norm $||h||^2$ of the vector h. The function LEARN again picks a hypothesis that minimizes a combination of two terms, but in this case the second term is a function that penalizes high values for $||h||^2$.

A remarkable property of this method (at least for certain hypothesis spaces, such as SVMs or neural networks) is that the number of training examples required is linear in $||h||^2$, but is otherwise INDEPENDENT OF THE NUMBER OF PARAMETERS n. This method thus takes advantage of distributional bias in such a way that it is the LIKELIHOOD of the hypothesis (as defined by $||h||^2$) that determines the difficulty of learning it, not the complexity of its representation.

5.5 *PAC/VC learnability and distributions on training samples*

While imposing a prior distributional bias on $\mathscr{H}$ can improve convergence rates in PAC/VC-learning, it does not address an important limitation of this framework with respect to natural-language grammar induction.[17] Nowak et al. (2002) point out that the set of finite and the set of regular languages are not PAC/VC-learnable due to the fact that they have infinite VC-dimension. This result follows straightforwardly for both cases. For the set of finite languages, all possible assignments of Boolean values to the members of any sample set of k strings (1 for membership in the language and 0 for exclusion) will be covered by elements of the set of finite sets of strings. Hence any finite sample string is shattered by the elements of $\mathscr{H}$ for finite languages. Similarly, if $\mathscr{H}$ consists of the infinite set of possible Finite State Automata (equivalently, the set of Finite State Grammars), then any sample set of regular strings will be shattered by members of $\mathscr{H}$. Therefore, the set of finite languages and the set of regular languages are not PAC/VC-learnable. At first blush, this may seem surprising, but with so few constraints from the hypothesis space, generalization is simply impossible.

A possible solution to this difficulty is to impose an upper bound on the size of finite languages and on the number of states in a possible FSA (rules in a possible FSG). Nowak et al. (2002) observe that for any specified value n, the set of finite languages of size n and the set of regular languages generated by FSAs with n states (FSGs with n rules) have finite VC dimension

[17] We are grateful to Alex Clark for invaluable discussion and advice on the issues dealt with in this section.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

and so are PAC/VC-learnable. In fact, Shinohara (1994) shows that the set of languages generated by context-sensitive grammars with n productions is learnable in the limit in Gold's sense.[18]

We can even choose the bound as an appropriate function of the number of training examples. In a framework in which we attempt to converge on a hypothesis as the number of training examples grows, this allows convergence in the limit for arbitrary languages from these otherwise problematic classes. A similar effect can be obtained by placing a prior probability over the bound, effectively a prior on the size of the grammar.

Another natural way of dealing with this problem is to impose constraints on the distribution of the samples to which the learner is exposed. On this view the data available to a learner exhibits a distributional pattern determined by a grammar. Although the samples to which a learner is exposed are randomly selected, they are taken from a particular probability distribution on the strings of the target language. Clark & Thollard (2004) propose a distributional sample bias for PAC learning of regular languages. They show that if the data string distributions for regular languages are determined by probabilistic FSAs corresponding to the target grammars that generate these languages, then the set of regular languages is PAC-learnable. It seems reasonable to generalize this approach to other types of grammars, like context-free grammars (CFGs) and (mildly) context sensitive grammars (CSGs).[19] We would obtain the distributions on the training data necessary to render the sets of languages corresponding to these grammars PAC-learnable, from probabilistic versions of the target grammars. The advantage of this approach is that it achieves PAC learnability not by imposing an arbitrary cardinality restriction on the number of rules in the grammars of $\mathscr{H}$, but through constraints on possible distributions of the training samples that bear a direct relation to the grammars that generate this data.

5.6 *Discussion*

While we have concentrated on the PAC/VC framework, online learning (Blum 1996) and Bayesian methods (Berger 1985) have many similarities. Online learning does not make use of the assumption of a distribution $\mathscr{D}$ over training and test examples, but nevertheless shares close connections with

[18] Scholz & Pullum (2002) cite Shinohara's result to support their claim that natural-language grammars are learnable even within the narrow strictures of Gold's learning theory, given a sufficiently large upper bound on the number of rules in the set of possible CSGs for natural languages.

[19] See Clark (2006) for an application of this approach to a particular subclass of CFGs, the class corresponding to the non-terminally separated languages.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

PAC/VC learning. Bayesian methods do not make use of the frequentist style of analysis in the PAC/VC approach, but nevertheless share several concepts – for example, the idea of a hypothesis or parameter space; the idea of a distributional bias of members of the hypothesis space; and the idea of a definition of LEARN that balances how well a hypothesis fits the training data against some prior probability of that hypothesis.

We are by no means the first authors to suggest that Gold's framework is problematic, or that other frameworks such as PAC/VC learning offer a preferable alternative. Johnson (2004) discusses problems with Gold's theorem. Clark (2004) also points out problems with some of the assumptions underlying Gold's approach when applied to grammar induction for natural languages. Nowak et al. (2002) consider the use of PAC/VC analysis instead of Gold's framework; Poggio et al. (2004) propose that the definition of LEARN can employ a distributional bias that may be useful in language acquisition. Pereira (2000) makes the point that in some cases infinite hypothesis spaces may be learnable, whereas in some cases finite hypothesis spaces are too complex. However, he discusses computational complexity of learning, which we have not discussed here, concentrating instead on complexity in terms of the number of training examples required. Niyogi (2006) provides an excellent source for theory in this area.

We are also not the first to note that sample complexity (the number of samples required for effective learning) is an important issue. Several authors have argued for learning mechanisms within particular (typically strong-bias) theories on the basis of low sample complexity, at least informally construed. The work of Niyogi & Berwick (1996) and Niyogi (2006) is a central example; this work and related efforts are important for addressing questions of plausibility of P&P models from a learning perspective. We merely point out that strong bias is neither necessary nor sufficient for low sample complexity.

Most importantly, PAC/VC learning, online learning, and Bayesian methods all have a very clear notion of learning bias, which is instantiated within the choices of $\mathscr{H}$ and LEARN. Indeed, results concerning convergence in the limit, or rate of convergence, give strong evidence for the need for such a bias, and the relationship between the character of this bias and the feasibility of learning. This runs counter to the perception – which in our view is misguided – that 'statistical' or 'empiricist' approaches to learning do not employ bias.

6. Different concepts of parameters in grammar

There are important distinctions between the notion of parameter invoked within the P&P view of grammar and the one employed to specify probabilistic language models. It is useful to identify these differences and consider their implications for the issues that we are addressing here.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

6.1 *Parameters in a P&P theory of UG*

Within the P&P framework a parameter is an underspecified element within one of the conditions that defines UG. Advocates of this framework argue that parameterized principles of UG provide an effective formal mechanism for explaining both the facts of language acquisition and of language variation. We have already addressed the problematic nature of the claim that a small number of parameters is necessitated by learnability considerations. Another advantage claimed for this approach is the possibility of organizing parameters in implicational relations that permit one to express observed dependencies of grammatical properties across languages. So, for example, a positive value for the pro-drop parameter is taken to imply selection of rich verbal agreement morphology.

It is important to note that this approach seeks to maximize the mutual dependence of parameters through a subsumption relation that supports type inference. One would expect proponents of this framework to develop detailed proposals for parameter hierarchies that specify the type structure which constitutes the core of UG. In fact, as Newmeyer (2004) observes, after twenty-five years of work within the P&P framework the only instance of such a parametric type system that has been suggested to date is Baker's (2001) parameter hierarchy (PH).

Newmeyer shows that the typological predictions that PH makes are incorrect for a large class of cases. So, for example, PH entails that head-final languages exhibit morphological case marking. However, 36% of such languages do not have case, while significant percentages of non-head-final languages do (42% of V-initial and 30% of V-medial languages). Newmeyer also shows that arguments for clustering of properties within a language due to a particular parameter setting do not, in general, hold. He illustrates this claim with a systematic study of the relation between null subjects, subject inversion, and *that*-trace violations, which indicates that none of the purported clustering effects among these phenomena are sustained cross-linguistically.

The potential explanatory significance of parameters has been significantly weakened by their recent relocation from general principles and constraints of grammar to lexical items, specifically functional heads, within the minimalist program (Chomsky 1995) version of the P&P project. As Newmeyer points out, on this view parametric values are set not for an entire language but for individual lexical items and categories. Lexical parameters become a descriptive device for capturing the facts of word order (and other phenomena) on virtually a language by language basis. They have little if any predictive content.[20]

[20] This point is also argued by Johnson & Lappin (1999: 83f.) and Bouchard (2005).

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Newmeyer argues that a rule-based theory of UG is a better alternative to the P&P framework. Rules are required in addition to parameters in a P&P theory in any case, and parameters are, when correctly formulated, fully equivalent to alternative rules of a particular kind.[21]

He concludes that discarding parameters in UG has no direct bearing on the issue of innateness. This is clearly true, as prior to the emergence of the P&P approach of Chomsky (1981), Chomsky (1965) described UG as the schema of a grammar that defines the set of possible rules for each component of the grammar.

However, giving up the P&P framework seriously undermines the advantages that its advocates have claimed for a rich theory of UG as the basis for an explanation of language acquisition, as well as an account of language variation. Specifically, language acquisition can no longer be reduced to the process of selecting among a small number of possible values for a finite set of parameters. Instead, a separate learning theory is required to explain how a schema of grammar can provide the basis for extracting a particular grammar from linguistic data. The probabilistic language models that ML employs can, in principle, fill this gap. But then the burden of innateness claims falls upon the learning theory that ML provides. To the extent that a weak-bias model of the sort that we have been suggesting is adequate to derive the observed facts of grammar, the task-specific content of UG becomes correspondingly minimal.

6.2 *Parameters in probabilistic language models*

In contrast to the P&P view of UG, probabilistic language models used in the natural-language engineering community tend to minimize the number of parameters and to maximize their relative independence in order to facilitate the computation of the model. Moreover, these parameters specify the basic elements of underlying structure in the data, rather than underspecified features of principles that make up a grammar.

Consider, for example, a simple non-lexicalized probabilistic context-free grammar G of the sort briefly sketched in section 4.1. We take G as a model of the language L (as described, for example, by Jurafsky & Martin 2000). Its parameters are (N, T, P, S, D), where N is the set of non-terminal symbols, T is the set of terminals (the lexicon), S is the start symbol of G (corresponding to the root node of a sentence), R is the set of productions (CFG rules), and D is a function that assigns probabilities to the elements of P. When values

[21] Roberts & Holmberg (2005) challenge Newmeyer's critique by citing evidence of parametric clustering in several Scandinavian languages. Newmeyer (2006) responds to these claims effectively by pointing out that (i) the data can be as easily accommodated on a rule-based account, and, more seriously, (ii) the parameters that Roberts & Holmberg propose do not generalize across genetically unrelated languages.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

are specified for all of these parameters, G provides a model of L that determines a probability for every sentence s relative to the parses of s that G produces.

If we have a corpus of parse-annotated sentences of L that provides a gold standard for the model, then we can set the parameters of G straightforwardly. N, T, P, and S are extracted directly from the parse annotations, and the value of D for each element $A \rightarrow \beta_1 \ldots \beta k$ of P can be estimated by the MLE formula

$$\frac{c(A \rightarrow \beta_1 \cdots \beta_k)}{c(A \rightarrow \gamma)}$$

However, if we have only an unannotated corpus, then it is necessary to estimate the values of the parameters of G. This involves using statistical learning algorithms to identify structure in the data corresponding to each of the parameters. The values of N (non-terminals) and T (terminals) of G can be identified by unsupervised learning through clustering techniques. Once these parameters have been specified, S and the elements of P are effectively given. It is then necessary to estimate D by computing the possible parses of the sentences in the corpus using a procedure like the inside-outside algorithm (as described, for example, by Manning & Schütze 1999).

Notice that while the parameters of G jointly define the search space for constructing a grammar they do not have to be taken as irreducibly given, in the sense that the sets of their possible values are pre-specified, as is the case for the parameters of a UG within the P&P framework. The possible values for parameters of a language model like G can be built up on the basis of antecedent applications of probabilistic learning algorithms that supply more basic structures. The vocabulary of G can be obtained by unsupervised morphological learning (Goldsmith 2001, Schone & Jurafsky 2001). The preterminal lexical part of N can be acquired by unsupervised POS tagging (Clark 2003). The remaining non-terminals and the CFG rules can be identified through unsupervised grammar induction techniques (Klein & Manning 2002, 2004). Each of the learning algorithms used to obtain values for these parameters will in turn require a model with parameters whose possible values are constrained in non-trivial ways. Some of these parameters may be further reducible to structures supplied by more basic unsupervised learning procedures.

7. Conclusion

Recent research on unsupervised machine learning of grammar offers support for the view that knowledge of language can be achieved through general machine learning methods with a minimal set of initial settings for possible linguistic categories and rule hypotheses. This work also suggests a

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

sequenced bootstrap model of language learning in which each level of structure acquired provides the input to a higher successor component of grammar. In at least some cases both the basic categories and the hypothesis space might be derived from more general cognitive processing patterns (like the binary branching trees and non-overlap constraint that Klein & Manning use to bias their models).

Machine learning experiments on grammar induction, particularly those involving unsupervised learning, can contribute important insights into the necessary conditions for language acquisition, at the least by vitiating poverty-of-stimulus arguments. They do not, of course, show us anything about the processes that human learners actually apply in acquiring natural language. This is the proper concern of research in psychology and neuroscience. These experiments can, however, demonstrate the viability of particular language models as learning mechanisms. To the extent that the bias of a successful model is defined by a comparatively weak set of language-specific conditions, we can rely more heavily on task-general machine learning methods to explain the possibility of acquiring linguistic knowledge, in lieu of psychological evidence that supports an alternative view.

REFERENCES

Baker, Mark. 2001. *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

Bartlett, Peter L. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* **44**. 525–536.

Berger, James O. 1985. *Statistical decision theory and Bayesian analysis*, 2nd edn. New York: Springer.

Blum, Avrim. 1996. On-line algorithms in machine learning. *The Dagstuhl Workshop on Online Algorithms*, 306–325. Dagstuhl.

Bouchard, Denis. 2005. Exaption and linguistic explanation. *Lingua* **115**. 1685–1696.

Carroll, Glenn & Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In Carl Weir, Stephen Abney, Ralph Grishman & Ralph Weischedel (eds.), *Working notes of the Workshop on Statistically-based NLP Techniques*, 1–13. Menlo Park, CA: AAAI Press.

Charniak, Eugene & Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. *43rd Annual Meeting of the Association for Computational Linguistics* (*ACL '05*), 173–180. Ann Arbor, MI: Association for Computational Linguistics. http://www.aclweb.org/anthology/P/P05/P05-1022 (6 March 2007).

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, Noam. 1995. *The Minimalist program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2000. *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Chouinard, Michelle M. & Eve V. Clark 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* **30**. 637–669.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. *10th Annual Meeting of the European Association for Computational Linguistics*, 59–66. Budapest.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Clark, Alexander. 2004. Grammatical inference and the argument from the poverty of the stimulus. *AAAI Spring Symposium on Interdisciplinary Approaches to Language Learning*. Stanford, CA.
Clark, Alexander. 2006. Pac-learning unambiguous NTS languages. *8th International Colloquium on Grammatical Inference*, 59–71. Berlin: Springer.
Clark, Alexander & Remi Eyraud 2006. Learning auxiliary fronting with grammatical inference. *8th Conference on Computational Language Learning* (*conll-x*), 125–132. New York.
Clark, Alexander & Franck Thollard. 2004. Partially distribution-free learning of regular languages from positive samples. *COLING 2004*, 85–91. Geneva.
Collins, Michael. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. dissertation, University of Pennsylvania.
Cortes, Corinna & Vladimir Vapnik 1995. Support-vector networks. *Machine Learning* **20**. 273–297.
Crain, Stephen. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences* **14**. 597–650.
Fodor, Jerry & Zenon Pylyshyn. 1988. Connectionism and cognitive architecture: A critique. *Cognition* **28**. 3–71.
Gold, E. Mark. 1967. Language identification in the limit. *Information and Control* **10**. 447–474.
Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* **27**. 153–198.
Good, I. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* **40**. 237–264.
Goodman, Joshua. 1996. Parsing algorithms and metrics. In Arivind Joshi & Martha Palmer (eds.), *The Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, 177–183. San Francisco: Morgan Kaufmann Publishers.
Johnson, David E. & Shalom Lappin. 1999. *Local constraints vs. economy* (Stanford Monographs in Linguistics). Stanford, CA: CSLI.
Johnson, Kent. 2004. Gold's theorem and cognitive science. *Philosophy of Science* **71**. 571–592.
Jurafsky, Daniel & James Martin. 2000. *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.
Klein, Dan & Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. *40th Annual Meeting of the Association for Computational Linguistics*, 128–135. Philadelphia, PA.
Klein, Dan & Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. *42th Annual Meeting of the Association for Computational Linguistics*, 479–486. Barcelona.
Lappin, Shalom. 2005. Machine learning and the cognitive basis of natural language. *Computational Linguistics in the Netherlands 2004*, 1–11. Leiden.
Legate, Julie A. & Charles D. Yang. 2002. Empirical reassessment of stimulus poverty arguments. *The Linguistic Review* **19**. 151–162.
Lewis, John & Jeffrey Elman. 2002. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *26th Annual Boston University Conference on Language Development*, 359–370. Somerville, MA.
MacWhinney, Brian. 1995. *The CHILDES project: Tools for analyzing talk*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.
Manning, Christopher & Hinrich Schütze. 1999. *Foundations of statistical language processing*. Cambridge, MA: MIT Press.
Marcus, Mitchell. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**. 313–330.
Newmeyer, Frederick J. 2004. Against a parameter setting approach to typological variation. *Linguistic Variation Year Book* **4**, 181–234. Amsterdam & Philadelphia: John Benjamins.
Newmeyer, Frederick J. 2006. A rejoinder to 'On the role of parameters in Universal Grammar: A reply to Newmeyer' by Ian Roberts & Anders Holmberg. Ms., Department of Linguistics, University of Washington, Seattle, WA.
Niyogi, Partha. 2006. *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.
Niyogi, Partha & Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* **61**. 161–193.

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press

Nowak, Martin A., Natalia L. Komarova & Partha Niyogi. 2002. Computational and evolutionary aspects of language. *Nature* **417**. 611–617.
Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 1239–1253. London: Royal Society.
Perfors, Amy, Joshua B. Tenenbaum & Terry Regier. 2006. Poverty of the stimulus? A rational approach. *28th Annual Conference of the Cognitive Science Society*, 663–668. Vancouver.
Poggio, Tomaso, Ryan Rifkin, Sayan Mukherjee & Partha Niyogi. 2004. General conditions for predictivity in learning theory. *Nature* **428**. 419–422.
Postal, Paul M. 2004. *Sceptical linguistic essays*. Oxford: Oxford University Press.
Pullum, Geoffrey K. & Barbara Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* **19**. 9–50.
Roberts, Ian & Anders Holmberg. 2005. On the role of parameters in Universal Grammar: A reply to Newmeyer. In Hans Broekhuis, Norbert Corver, Riny Huybregts, Ursula Kleinhenz & Jan Koster (eds.), *Organizing grammar: Linguistic studies in honor of Henk van Riemsdijk*, 538–553. Berlin: Mouton de Gruyter.
Saxton, Matthew, Carmel Houston-Price & Natasha Dawson. 2005. The prompt hypothesis: Clarification requests as corrective input for grammatical errors. *Applied Psycholinguistics* **26**. 393–414.
Scholz, Barbara & Geoffrey K. Pullum. 2002. Searching for arguments to support linguistic nativism. *The Linguistic Review* **19**. 185–223.
Scholz, Barbara & Geoffrey K. Pullum. 2006. Irrational nativist exuberance. In Rob Stainton (ed.), *Contemporary debates in cognitive science*, 59–80. Oxford: Blackwell.
Schone, Patrick & Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. *The North American Chapter of the Association for Computational Linguistics* (*NAACL-2001*). Pittsburgh, PA.
Schütze, Hinrich. 1995. Distributional part-of-speech tagging. *The European Chapter of the Association for Computational Linguistics* (*EACL* **7**), 141–148. Dublin.
Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* **27**. 379–423. 623–656.
Shinohara, Takeshi. 1994. Rich classes inferable from positive data: Length-bounded elementary formal systems. *Information and Computation* **108**. 175–186.
Valiant, Leslie G. 1984. A theory of the learnable. *The Sixteenth Annual ACM Symposium on Theory of Computing* (*STOC '84*), 436–445. New York: ACM Press.
Vapnik, Vladimir. 1998. *Statistical learning theory*. New York: Wiley.
Vapnik, Vladimir & Alexey Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* **16**. 264–280.

Authors' addresses: (*Lappin*)
Department of Philosophy, King's College London, The Strand,
London WC2R 2LS, U.K.
E-mail: shalom.lappin@kcl.ac.uk

(*Shieber*)
Division of Engineering and Applied Sciences, Harvard University,
33 Oxford Street – 245, Cambridge, MA 02138, U.S.A.
E-mail: shieber@deas.harvard.edu

https://doi.org/10.1017/S0022226707004628 Published online by Cambridge University Press