BOUNDED CONTEXT PARSING AND EASY LEARNABILITY

Robert C. Berwick
Room 820, MIT Artificial Intelligence Lab
Cambridge, MA 02139

ABSTRACT

Natural languages are often assumed to be constrained so that they are either easily learnable or parsable, but few studies have investigated the connection between these two "functional" demands. Without a formal model of parsability or learnability, it is difficult to determine which is more "dominant" in fixing the properties of natural languages. In this paper we show that if we adopt one precise model of "easy" parsability, namely, that of *bounded context parsability*, and a precise model of "easy" learnability, namely, that of *degree 2 learnability*, then we can show that certain families of grammars that meet the bounded context parsability condition will also be degree 2 learnable. Some implications of this result for learning in other subsystems of linguistic knowledge are suggested.[1]

I INTRODUCTION

Natural languages are usually assumed to be constrained so that they are both learnable and parsable. But how are these two functional demands related computationally? With some exceptions,[2] there has been little or no work connecting these two key constraints on natural languages, even though linguistic researchers conventionally assume that learnability somehow plays a dominant role in "shaping" language, while computationalists usually assume that efficient processability is dominant. Can these two functional demands be reconciled? There is in fact no *a priori* reason to believe that the demands of learnability and parsability are necessarily compatible. After all, learnability has to do with the scattering of possible grammars with respect to evidence input to a learning procedure. This is a property of a *family* of grammars. Efficient parsability, on the other hand, is a property of a *single* grammar. A family of grammars could be easily learnable but not easily parsable, or vice-versa. It is easy to provide examples of both sorts. For example, there are finite collections of grammars generating non-recursive languages that are easily learnable (just use a disjoint vocabulary as triggering evidence to distinguish among them). Yet by definition these languages cannot be easily parsable. On the other hand as is well known even the class of all finite languages plus the universal infinite language covering them all is not learnable from just positive evidence (Gold 1967). Yet each of these languages is finite state and hence efficiently analyzable.

This paper establishes the first known results formally linking efficient parsability to efficient learnability. It connects a particular model of efficient parsing, namely, bounded context parsing with lookahead as developed by Marcus 1980, to a particular model of language acquisition, the Bounded Degree of Error (BDE) model of Wexler and Culicover 1980. The key result: bounded context parsability implies "easy" learnability. Here, "easily learnable" means "learnable from simple, positive (grammatical) sentences of bounded degree of embedding." In this case then, the constraints required to guarantee easy parsability, as enforced by the bounded context constraint, are at least as strong as those required for easy learnability. This means that if we have a language and associated grammar that is known to be parsable by a Marcus-type machine, then we already know that it meets the constraints of bounded degree learning, as defined by Wexler and Culicover.

A number of extensions to the learnability-parsability connection are also suggested. One is to apply the result to other linguistic subsystems, notably, morphological and phonological rule systems. Although these subsystems are finite state, this does not automatically imply easy learnability, as Gold (1967) shows. In fact, identification is still computationally intractable -- it is NP-hard (Gold 1978), taking an amount of evidence exponentially proportional to the number of states in the target finite state system. Since a given natural language could have a morphological system of a few hundred or even a few thousand states (Kimmo 1983, for Finnish), this is a serious problem. Thus we must find additional constraints to make natural morphological systems tractably learnable. An analog of the bounded context model for morphological systems may suffice. If we require that such systems be *k-reversible*, as defined by Angluin (in press), then an efficient polynomial time induction algorithm exists.

To summarize, what is the importance of this result for computational linguistics?

> o It shows for the first time that parsability is *stronger* constraint than learnability, at least given this particular way of defining the comparison. Thus computationalists may have been right in focusing on efficient parsability as a metric for comparing theories.

1. This work has been carried out at the MIT Artificial Intelligence Laboratory. Support for the Laboratory's artificial intelligence research is provided in part by the Defense Advanced Research Projects Agency.

2. See Berwick 1980 for a sketch of the connections between learnability and parsability.

o It provides an explicit criterion for learnability. This criterion can be tied to known grammar and language class results. For example, we can say that the language $a^n b^n c^n$ will be easily learnable, since it is bounded context parsable (in an extended sense).

o It formally connects the Marcus model for parsing to a model of acquisition. It pinpoints the relationship of the Marcus parser to the LR(k) and bounded context parsing models.

o It suggests criteria for the learnability of phonological and morphological systems. In particular, the notion of *k-reversibility*, the analog of bounded context parsability for finite state systems, may play a key role here. The reversibility constraint thus lends learnability support to computational frameworks that propose "reversible" rules (such as that of Koskenniemi 1983) versus those that do not (such as standard generative approaches).

This paper is organized as follows. Section 1 reviews the basic definitions of the bounded context model for parsing and the bounded degree of error model for learning. Section 2 sketches the main result, leaving aside the details of certain lemmas. Section 3 extends the bounded context--bounded degree of error model to morphological and phonological systems, and advances the notion of *k-reversibility* as the analog of bounded context parsability for such finite state systems.

II BOUNDED CONTEXT PARSABILITY AND BOUNDED DEGREE OF ERROR LEARNING

To begin, we define the models of parsing and learning that will be used in the sequel. The parsing model is a variant of the Marcus parser. The learning theory is the Degree 2 theory of Wexler and Culicover (1980). The Marcus parser defines a class of languages (and associated grammars) that are easily parsable; Degree 2 theory, a class of languages (and associated grammars) that is easily learnable.

To begin our comparison, we must say what class of "easily learnable" languages Degree 2 theory defines. The aim of the theory is to define constraints such that a family of transformational grammars will be learnable from "simple" data; the learning procedure can get positive (grammatical) example sentences of depth of embedding of two or less (sentences up to two embedded sentences, but no more). The key property of the transformational family that establishes learnability is dubbed *Bounded Degree of Error*. Roughly and intuitively, BDE is a property related to the "separability" of languages and grammars given simple data: if there is a way for the learner to tell that a currently hypothesized language (and grammar) is incorrect, then there must be some simple sentence that reveals this -- all languages in the family must be separable by simple sentences.

The way that the learner can tell that a currently hypothesized grammar is wrong given some sample sentence is by trying to see whether the current grammar can map from a deep structure for the sentence to the observed sample sentence. That is, we imagine the learner being fed with a series of base (deep structure)-surface sentence (denoted "b, s") pairs. (See Wexler and Culicover 1980 for details and justification of this approach, as well as a weakening of the requirement that base structures be available; see Berwick 1980 1982 for an independently developed computational version.) If the learner's current transformational component, T_j, can map from b to s, then all is well. If not, and $T_j(b) = s^*$ does not equal s, then a *detectable error* has been uncovered.

With this background we can provide a precise definition of the BDE property:

> A family of transformationally-generated languages L possesses the BDE property iff for any base grammar B (for languages in L) there exists a finite integer U, such that for any possible adult transformational component A and learner component C, if A and C disagree on any phrase-marker b generated by B, then they disagree on some phrase-marker b' generated by B, with b' of degree at most U. Wexler and Culicover 1980 page 108.

If we substitute 2 for U in the theorem, we get the Degree 2 constraint.

Once BDE is established for some family of languages, then convergence of a learning procedure is easy to proved. Wexler and Culicover 1980 have the details, but the key insight is that the number of possible errors is now bounded from above.

The BDE property can be defined in any grammatical framework, and this is what we shall do here. We retain the idea of mapping from some underlying "base" structure to the surface sentence. (If we are parsing, we must map from the surface sentence to this underlying structure.) The mapping is not necessarily transformational, however; for example, a set of context-free rules could carry it out. In this paper we assume that the mapping from surface sentences to underlying structures is carried out by a Marcus-type parser. The mapping from structure to sentence is then defined by the inverse of the operation of this machine. This fixes one possible target language. (The full version of this paper defines this mapping in full.)

Note further that the BDE property is defined not just with respect to possible adult target languages, but also with respect to the distribution of the learner's possible guesses. So for example, even if there were just ten target languages (defining 10 underlying grammars), the BDE property must hold with respect to those languages and any intervening learner languages (grammars). So we must also define a *family* of languages to be acquired. This is done in the next section.

BDE, then, is our criterial property for easy learnability. Just those families of grammars that possess the BDE property (with respect to a learner's guesses) are easily learnable.

Now let us turn to bounded context parsability (BCP). The definition of BCP used here an extension of the standard definition as in Aho and Ullman 1972 p. 427. Intuitively, a grammar is BCP if it is "backwards deterministic" given a radius of k tokens around

every parsing decision. That is, it is possible to find deterministically the production that applied at a given step in a derivation by examining just a bounded number of tokens (fixed in advance) to the left and right at that point in the derivation. Following Aho and Ullman we have this definition for *bounded right-context grammars*:

> G is bounded right-context if the following four conditions:
>
> (1) $S \Rightarrow \alpha A \omega \Rightarrow \alpha\beta\omega$ and
> (2) $S \Rightarrow \gamma B \chi \Rightarrow \gamma\delta\chi = \alpha'\beta\psi$
> are rightmost derivations in the grammar;
>
> (3) the length of χ is less than or equal to the length of ψ and
>
> (4) the last m symbols of α and α' coincide, and the first n symbols of ω and ψ coincide
>
> imply that $A = B$, $\alpha' = \gamma$, and $\psi = \chi$.

We will use the term "bounded context" instead of "bounded right-context." To extend the definition we drop the requirement that the derivation is rightmost and use instead non-canonical derivation sequences as defined by Szymanski and Williams (1976). This model corresponds to Marcus's (1980) use of *attention shifts* to postpone parsing decisions until more right context is examined. The effect is to have a lookahead that can include nonterminal names like NP or VP. For example, in order to successfully parse *Have the students take the exam*, the Marcus parser must delay analyzing *have* until the full NP *the students* is processed. Thus a canonical (rightmost) parse is not produced, and the lookahead for the parser includes the sequence *NP--take*, successfully distinguishing this parse from the *NP--taken* sequence for a yes-no question. This extension was first proposed by Knuth (1965) and developed by Szymanski and Williams (1976). In this model we can postpone a canonical rightmost derivation some fixed number of times t. This corresponds to building t complete subtrees and making these part of the lookahead before we return to the postponed analysis.

The Marcus machine (and the model we adopt here) is not as general as an LR(k) type parser in one key respect. An LR(k) parser can use the *entire* left context in making its parsing decisions. (It also uses a bounded right context, its lookahead.) The LR(k) machine can do this because the entire left context can be stored as a regular set in the finite control of the parsing machine (see Knuth 1965). That is, LR(k) parsers make use of an encoding of the left context in order to keep track of what to do. The Marcus machine is much more limited than this. Local parsing decisions are made by examining strictly *literal* contexts around the current locus of parsing contexts. A finite state encoding of left context is not permitted.

The BCP class also makes sense as a proxy for "efficiently parsable" because all its members are analyzable in time linear in the length of their input sentences, at least if the associated grammars are context-free. If the grammars are not context-free, then BCP members are parsable in at worst quadratic (n squared) time. (See Szymanski and Williams 1976 for proofs of these results.)

III CONNECTING PARSABILITY AND LEARNABILITY

We can now at least formalize our problem of comparing learnability and parsability. The question now becomes: What is the relationship between the BDE property and the BCP property? Intuitively, a grammar is BCP if we can always tell which of two rules applied in a given bounded context. Also intuitively, a family of grammars is BDE if, given any two grammars in the family G and G' with different rules R and R' say, we can tell which rule is the correct one by looking at two derivations of bounded degree, with R applying in one and yielding surface string s, and R' applying in the other yielding surface string s', with s not equal to s'. This property must hold with respect to all possible adult and learner grammars. So a space of possible target grammars must be considered. The way we do this is by considering some "fixed" grammar G and possible variants of G formed by substituting the production rules in G with hypothesized alternatives.

The theorem we want to now prove is:

> If the grammars formed by augmenting G with possible hypothesized grammar rules are BCP, then that family is also BDE.

The theorem is established by using the BCP property to directly construct a small-degree phrase marker that meets the BDE condition. We select two grammars G, G' from the family of grammars. Both are BCP, by definition. By assumption, there is a detectable error that distinguishes G with rule R from G' with rule R'. Let us say that Rule R is of the form $A \Rightarrow \alpha$; R' is $B \Rightarrow \alpha'$.

Since R' determines a detectable error, there must be a derivation with a common sentential form Φ such that R applies to Φ and eventually derives sentence s, while R' applies to Φ and eventually derives s' different from s. The number of steps in the derivation of the the two sentences may be arbitrary, however. What we must show is that there are two derivations bounded in advance by some constant that yield two different sentences.

The BCP conditions state that identical (m,n) contexts imply that A and B are equal. Taking the contrapositive, if A and B are unequal, then the (m,n) context must be nonidentical. This establishes that BCP implies (m,n) context error detectability.[3]

We are not yet done though. An (m,n) context detectable error could consist of terminal *and* nonterminal elements, not just terminals (words) as required by the detectable error condition. We must show that we can extend such a detectable error to a surface sentence detectable error with an underlying structure of bounded degree. An easy lemma establishes this.

> If R' is an (m,n) context detectable error, then R' is bounded degree of error detectable.

The proof (by induction) is omitted: only a sketch will be given here. Intuitively, the reason is that we can extend any nonterminals in the error-detectable (m,n) context to some valid surface sentence and bound this derivation by some constant fixed in advance and depending only on the grammar. This is because unbounded derivations are possible only by the repetition of nonterminals via recursion; since there are only a finite number of distinct nonterminals, it is only via recursion that we can obtain a derivation chain that is arbitrarily deep. But, as is well known (compare the proof of the pumping lemma for context-free grammars), any such arbitrarily deep derivation producing a valid surface sentence also has an associated truncated derivation, bounded by a constant

dependent on the grammar, that yields a valid sentence of the language. Thus we can convert any (m,n) context detectable error to a bounded degree of error sentence. This proves the basic result.

As an application, consider the strictly context-sensitive language $a^n b^n c^n$. This language has a grammar that is BCP in the extended sense (Szymanski and Williams 1976). The family of grammars obtained by replacing the rules of this BCP grammar by alternative rules that are also BCP (including the original grammar) meets the BDE condition. This result was established independently by Wexler 1982.

IV EXTENSIONS OF THE BASIC RESULT

In the domain of syntax, we have seen that constraints ensuring efficient parsability also guarantee easy learnability. This result suggests an extension to other domains of linguistic knowledge. Consider morphological rule systems. Several recent models suggest finite state transducers as a way to pair lexical (surface) and underlying forms of words (Koskenniemi 1983; Kaplan and Kay 1983). While such systems may well be efficiently analyzable, it is not so well known that easy learnability does not follow directly from this adopted formalism. To learn even a finite state system one must examine all possible state-transition combinations. This is combinatorially explosive, as Gold 1978 proves. Without additional constraints, finite transducer induction is intractable.

What is needed is some way to localize errors; this is what the bounded degree of error condition does.

Is there an analog of the the BCP condition for finite state systems that also implies easy learnability? The answer is yes. The essence of BCP is that derivations are backwards and forwards deterministic within local (m,n) contexts. But this is precisely the notion of *k-reversibility*, as defined by Angluin (in press). Angluin shows that k-reversible automata have polynomial time induction algorithms, in contrast to the result for general finite state automata. It then becomes important to see if k-reversibility holds for current theories of morphological rule systems. The full paper analyzes both "classical" generative theories (that do not seem to meet the test of reversibility) and recent transducer theories. Since k-reversibility is a sufficient, but evidently not a necessary constraint for learnability, there could be other conditions guaranteeing the learnability of finite state systems. For instance, One of these, the strict cycle condition in phonology, is also examined in the full paper. We show that the strict cycle also suffices to meet the BDE condition.

In short, it appears that at least in terms of one framework in which a formal comparison can be made, the same constraints that forge efficient parsability also ensure easy learnability.

3 One of the other three BCP conditions could also be violated, but these are assumed true by assumption. We assume the existence of derivations meeting conditions (1) and (2) in the extended sense, as well as condition (3).

V REFERENCES

Aho, J. and Ullman, J. 1972. *The Theory of Parsing, Translation, and Compiling*, vol. 1., Englewood-Cliffs, NJ: Prentice-Hall.

Angluin, D. 1982. Induction of k-reversible languages. In press, JACM.

Berwick, R. 1980. Computational analogs of constraints on grammars. Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics.

Berwick, R. 1982. Locality Principles and the Acquisition of Syntactic Knowledge, PhD dissertation, MIT Department of Electrical Engineering and Computer Science.

Gold, E. 1967. Language identification in the limit. *Information and Control*, 10.

Gold, E. 1978. On the complexity of minimum inference of regular sets. *Information and Control*, 39, 337-350.

Kaplan, R. and Kay, M. 1983. Word recognition. Xerox Palo Alto Research Center.

Koskenniemi, K. 1983. Two-Level Morphology: A General Computational Model for Word Form Recognition and Production, PhD dissertation, University of Helsinki.

Knuth, D. 1965. On the translation of languages from left to right. *Information and Control*, 8.

Marcus, M. 1980. *A Model of Syntactic Recognition for Natural Language*, Cambridge MA: MIT Press.

Szymanski, T. and Williams, J. 1976. Noncanonical extensions of bottomup parsing techniques. *SIAM J. Computing*, 5.

Wexler, K. 1982. Some issues in the formal theory of learnability, in C. Baker and J. McCarthy (eds.), *The Logical Problem of Language Acquisition.*

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