revised version to appear in *Language Universals* (Oxford Univ Press),
edited by M. Christiansen, C. Collins, S. Edelman

Edward P. Stabler

Computational models of language universals: Expressiveness, learnability and consequences

Every linguist is struck by similarities among even the most different and most culturally isolated human languages. It is natural to assume that some of these common properties, these language universals, might reflect something about the way people can learn and use languages. In some relevant sense, some of these properties may arise and be maintained even in culturally isolated languages because of special restrictions on the range of structural options available for human language learners. A bolder idea is that some of these language universals may guarantee that the whole class of languages with such properties is 'learnable' in a relevant sense. While considerable progress has been made on finding ways to clearly articulate and assess possibilities of these sorts in precise computational models, there has also been a shift to more sophisticated versions of a long-standing traditional perspective: it may not be so much the formal structure of human languages, but the special kinds of fit between form and meaning that give human languages their most distinctive properties, in which case some early work on language acquisition may have characterized inappropriately difficult learning problems. A more reasonable perspective on the learners' predicament may recognize a certain *non-arbitrariness* in the relation between structures and their semantic values, so that only certain kinds of structures are expected to carry certain sorts of semantic values. This can allow semantic properties of expressions to provide clues about syntactic structure, and vice versa, enriching the evidence available to the learner. This paper will review some fundamental results in this line of inquiry, from universals formulated in terms of expressive power of grammars, to results on learnable subsets of the languages defined by those grammars, leading finally to recent views on semantically-characterized grammatical universals. Even restricting attention to hypotheses that are most empirically secure and independent of any particular choice among the major traditions in grammatical theory, the modern perspective is surprising in many respects and quite different from anything that could have been conceived at the 1961 Conference on Language Universals (Greenberg, 1963).

UCLA Department of Linguistics

1. Universals of language complexity

Chomsky and others in the 1950's noticed that languages can be classified by the kinds of grammars that generate them, and that a straightforward classification in terms of grammar also corresponds also to a classification of the kinds of resources needed to recognize those languages (Chomsky, 1956). This 'Chomsky hierarchy' has been considerably elaborated and integrated into the theory of automata and complexity (Hopcroft and Ullman, 1979). Finding the place of human languages in this hierarchy is of interest because it provides an indication of what resources (memory, time) are required to recognize and produce them. This may sound straightforward, but it actually requires some sophistication to understand the project. In the first place, human linguistic behavior is influenced by many things; we would like to abstract away from coughs, interruptions, and memory limitations of various sorts. We adopt similar abstractions when we say that a calculator computes the sum or product function on integers. Such a claim is not refuted by the behavior of the device when its power fails or when the inputs exceed the memory limitations of the device.[1] The motivation for these abstractions is not merely simplicity. Rather, as in any science, we hope to be factoring the explanation along lines that correspond to the real causal sources of the behavior. The mechanisms involved in coughing or in responding to interruptions are relevantly different from those involved in producing or perceiving a fluent utterance. Consequently, to place human languages in the Chomsky hierarchy is to adopt a certain kind of explanation of human linguistic behavior, and so controversy is expected even among the best-informed researchers.

There is another reason for interest in properties of human languages, regarded as sets of sequences. These sequences, as produced in context and subject to various kinds of 'noise', certainly comprise one of the most important sources of evidence available to language learners. We would like to understand how perceptible properties of these unanalyzed sequences shape early language acquisition. Grammatically sophisticated notions like 'subject', 'modifier', or 'verb phrase' are used in framing most familiar universals, but to understand the earliest stages of language acquisition it is useful to identify universals that can apply before such sophisticated analyses are available.[2]

A third reason for being interested in claims about language complexity is that it provides a common basis for comparing grammars of very different kinds. Linguists are often very concerned with the exact nature of the description they provide of linguistic structures, and this concern is completely reasonable. For one thing, given the complexity of the domain being described, the simplicity of our description is a practical concern. But this also leads to a proliferation of descriptive formalisms – several major, distinct traditions and many very significant variants in each tradition – which can be an obstacle to effective communication and critical assessment. In the great diversity of formal proposals, though, an astounding convergence

among a great range of independently proposed formalisms has been discovered.

In the work of Joshi, Vijay-Shanker, and Weir (1991), Seki et al. (1991), and Vijay-Shanker and Weir (1994) four independently proposed grammar formalisms are shown to define exactly the same languages: a kind of head-based phrase structure grammars (HGs), combinatory categorial grammars (CCGs), tree adjoining grammar (TAGs), and linear indexed grammars (LIGs). Furthermore, this class of languages is included in an infinite hierarchy of languages that are defined by multiple context free grammars (MCFG), multiple component tree adjoining grammars (MCTAGs), linear context free rewrite systems (LCFRSs), and other systems. Later, it was shown a certain kind of "minimalist grammar" (MG), a formulation of the core mechanisms of Chomskian syntax – using the operations merge, move, and a certain strict 'shortest move condition' – define exactly the same class of languages (Michaelis, 2001; Harkema, 2001; Michaelis, 1998). These classes of languages are positioned between the languages defined by context free grammars (CFGs) and the languages defined by context sensitive grammars (CSGs) like this,

$$(1)\quad \text{CFG} \subset \boxed{\text{TAG} \equiv \text{CCG}\ldots} \subset \boxed{\text{MCTAG} \equiv \text{MCFG} \equiv \text{MG}\ldots} \subset \text{CSG}$$

where $\subset$ indicates proper subset relations between the definable languages and $\equiv$ relates formalisms that define exactly the same languages. The equivalence $\equiv$ is often called 'weak' since it considers only the definable sequences and not the structures of derivations, but an inspection of the proofs of these weak equivalence results reveals that they are not very difficult. The proofs provide recipes for taking a grammar from one formalism and converting it into an exactly equivalent grammar in another formalism. The recipes are not difficult because, in an intuitive sense which has not yet been formally captured,[3] the recursive mechanisms of each of these formalisms are rather similar. Furthermore, unlike earlier very expressive grammar formalisms[4] it is known that the classes boxed in (1) can both be recognized feasibly, by 'polynomial time' computations.

It may be a universal structural fact about human languages that they are always included in one of the classes boxed in (1). Joshi (1985) proposes a slightly weaker hypothesis, namely that human languages are 'mildly context sensitive' (MCS) in the sense that they have (i) limited crossing dependencies, (ii) constant growth, and (iii) polynomial parsing complexity. A language is said to have "constant growth" if there is a bound k such that whenever two sentences have lengths that differ by more than k, there is a sentence of intermediate length. The intuition here is that sentences are built up by simple combinations of smaller constituents (and so, for example, they do not allow an operation of unbounded copying). The notion of polynomial recognizability is discussed in any standard introduction to formal languages and computing (Hopcroft, Motwani, and Ullman, 2000; Lewis and Papadimitriou, 1981, for example). Both TAG languages and MCFG languages are MCS in this sense, but other classes are too.

The claims that human languages are definable by TAGs or MCFGs, or that they are MCS, are very strong claims with significant computational consequences. Mainstream work in linguistic theory can be seen as aiming to sharpen these results with more precise characterizations of the recursive mechanisms of grammar. But the basic claims mentioned here are also being challenged on empirical grounds. For example there are proposals to the effect that the grammars need certain kinds of copying mechanisms (Michaelis and Kracht, 1997; Stabler, 2004; Kobele, 2006), and this may require placing human languages in a slightly larger class. The 'parallel multiple context free grammars' (PMCFGs) defined by Seki et al. (1991) allow this kind of copying, and remain efficiently recognizable, but they lack the constant growth property. Many linguists like Joshi remain unpersuaded that anything like reduplication is needed anywhere in the syntax (cf. Pullum 2006). Other possible but less plausible threats to the MCS claims are more drastic; many seemingly minor variations on MCS grammars yield systems that can define any 'recursively enumerable' language (Gärtner and Michaelis, 2005; Kobele and Michaelis, 2005; Kobele, 2005, for example), in which case the mechanisms of grammar would tell us essentially nothing about human languages beyond the fact that they are finitely representable. But many linguists feel that even the strong claim that human languages are universally in the classes boxed in (1) is actually rather weak. They think this because, in terms of the sorts of things linguists describe in human languages, these computational claims tell us little about what human languages are like.

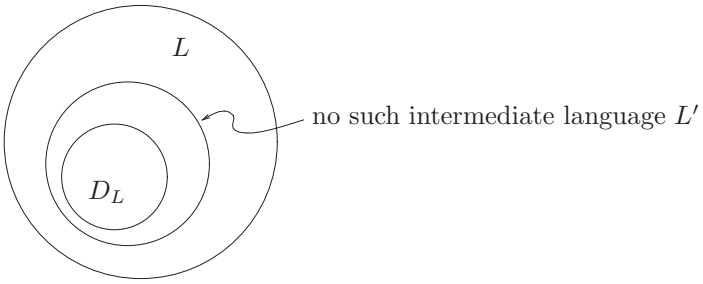
2. Learnable syntactic patterns: Gold

Perhaps stronger universal claims about language structure will come from computational models of learning. Some basic syntactic universals may reflect properties of the language learning mechanism, and it might even be the case that some of these properties guarantee the 'learnability' of human languages, in some relevant sense.

One framework for addressing these issues is provided by Gold and others (Gold, 1967; Jain et al., 1999). Noting that human learners seem to succeed without explicit instruction or feedback (Braine, 1971; Bowerman, 1988), one model of the evidence available to a learner is a 'positive text'. Given any language L, a text for that language is an infinite sequence containing all and only strings of the language. A learner can then be regarded as a function from longer and longer finite initial sequences of such a text to grammars, guesses about the language of the text. We say the learner converges if on some initial sequence of the text the learner makes a guess that does not change with any longer initial sequence. We say the learner successfully learns the text if the learner converges on a grammar that generates the language of the text. The learner is said to be able to learn the language L if the learner learns every text for that language. And finally a learner can learn a class of languages $\mathcal{L}$ if and only if it learns every language in the class.

Obviously, this is not meant to provide a realistic picture of human learning, but the framework is of interest for the insight it provides into the conditions in which a generalizing learner can be guaranteed to succeed in this simple sense of correctly identifying a text. A precise characterization of the classes of languages which can be learned from positive text, in the sense just defined, was provided by Angluin's (1980) subset theorem, which can be formulated this way:

> A collection $\mathcal{L}$ of (recursively enumerable) languages is learnable just in case, for every language L in the collection, you can find a finite subset D_L such that no language L' in the collection includes D_L and is properly included in L.

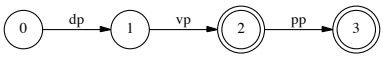


Using this theorem, it is easy to see that no learner can learn any class $\mathcal{L}$ that contains all the finite languages and also one or more infinite languages. For consider any one of the infinite languages L. Every one of the finite subsets $F \subset L$ is in $\mathcal{L}$, and so is the larger set L' that results from adding one more element $x \in L$ to F. So then for every finite F we have a situation where $F \subset L' \subset L$. That is, L has no finite distinguished subset of the sort required by the theorem, and so $\mathcal{L}$ is not learnable. The intuition behind this demonstration is clear: roughly, to conclude that a finite sample of data indicates an infinite pattern is to conclude that the data is not simply a finite stipulation that does not generalize; but the class of all finite languages is one where every finite set of data might be a stipulation. From this result, it follows that the none of classes indicated in (1) are learnable since they too contain all the finite languages together with some infinite ones.

We can also use the subset theorem to show that any finite class $\mathcal{L}$ is learnable. Any such finite class of languages can be listed $L_1, L_2, \ldots, L_n$ in such a way that if $j < i$, then $L_i \not\subseteq L_j$. Now, considering each L_i in turn and each earlier language, L_j, $j < i$, let let $x_{i,j}$ be some element that is in L_i but not in L_j. Then define $D_{L_i} = \{x_{i,j} |\ j < i\}$. It is easy to check that each D_{L_i} defined in this way is a 'distinguishing subset' satisfying the requirements of the subset theorem, and so $\mathcal{L}$ is learnable. This example is not very interesting, because the learnability of the class does not depend on any 'structural property' in any intuitive sense. That is, the learner needs no structural analysis of the expressions of the languages; all that matters is the presence of some string not found in 'earlier' languages.

It is worth briefly sketching a more interesting example to illustrate the kind of result we would like to obtain for human languages, a class that is learnable because of some interesting universal structural property. A finite state language can be defined by a machine like the following,

(2)



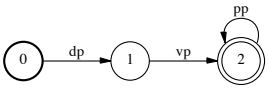
The states of the machine are circled; 0 is the initial state; the final states are doubly circled; and a string is in the language defined by the machine just in case that it labels a path along the arcs from the initial state to a final state.[5] So the machine in (2) defines a language containing just two strings, namely

(3) dp vp
dp vp pp

This finite state machine is deterministic in the sense that (i) it has at most 1 initial state, and (ii) no two arcs leaving any state have the same label. It is not hard to show that no deterministic machine with fewer states can define this same language.

We reverse a machine like the one shown above by (i) changing start states to final states, (ii) changing final states to start states, and (iii) reversing every arc. It is clear that the reverse of the machine shown above is not deterministic, since the reverse has two initial states. Now following Angluin (1982), define a finite state language L as reversible just in case the result of reversing the smallest deterministic finite state machine for L yields another deterministic machine. Clearly then, every language consisting of a single string is reversible, and so the class of reversible languages is infinite. But example (2) shows that the class of reversible languages does not include every finite language. And it is easy to see that the class includes infinitely many infinite languages, like the one defined by this machine,

(4)



Because of the loop on the final state, this machine defines the language containing sentences with dp vp followed by 0 or more pp's. In fact, Angluin proves that this infinite language is the smallest reversible language that contains the two sentences in (3). In other words, if a learner knows that the target language is reversible, and sees the two strings (3) then the most conservative guess the learner can make is that the target language is the infinite language defined by (4). It turns out that given any sample of input strings, the smallest reversible language containing that sample can

be efficiently computed, and a learner that always guesses this language will successfully learn any reversible language.

Do human languages have a universal structural property that similarly guarantees the learnability of human languages? There are two important points to make here. The first is that the grammars (or machines) in the examples above generate the data available to the learner. But in traditional approaches to human language we factor the grammar into parts. The syntax may determine the order of morphemes, but morphological and phonological processes also have an influence on what is available to the learner. In particular, notice that the definition of reversible explicitly depends on the identities of the elements labeling each arc, requiring a kind of 'forward and backward' non-ambiguity. All interesting positive learning results are like this: the learner must be able to figure out the language structure from the identities and positions of the elements in the data. So obviously, in human languages, we can expect structural universals to emerge from learning only when the data available to the learner is reflecting structural properties. Most structural properties would be hidden if every morpheme were silent, or if every morpheme sounded exactly like every other. So already we have a preliminary problem. Human languages allow homophony of various kinds, and there is no apparent fixed, finite bound to the extent of homophony. There are patterns of systematic homophony (syncretism) found in human languages, and there is also some 'random' accidental homophony (Williams, 1994; Bobaljik, 2002; Pertsova, 2006), and we would like to specify these things in such a way that we could determine the sorts of structural properties which should be visible nevertheless.

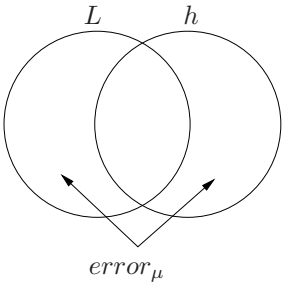
For the moment, the standard move is to adopt a linguistically non-standard understanding of the grammar and of what we mean by 'structural property', extending these notions down to the identities of perceived forms, e.g. morpheme sequences. And we adopt a psychologically non-standard view of the data available to the learner: morpheme sequences. We would like to remove these simplifications eventually, but they provide a preliminary way to return to our question: Do human languages have a universal structural property that guarantees the learnability of human languages? Recent work suggests that some phonotactic domains may have a basic property that guarantees learnability (Heinz, 2006), but for syntax (extended in the way just suggested to define languages of morpheme sequences), no such property is known.[6] For example, in reversible languages, if a word can be added to the end of a sentence, that word can be iterated any number of times, but this does not hold in human languages. For example, in English, while sentence-final modifiers might be iterable, optional final words cannot always be iterated:

(5)
I see
I see it
* I see it it

To determine how humans will generalize, what constructions can iterated or extracted from, it seems we need to be able to identify things like modifiers, arguments, and predicates. The way the learner generalizes must, it seems, be based on an analysis of the input in terms of this kind. How can such analyses be learned? The standard response is that we require semantic information to obtain such analyses, and the evidence for this suggestion is that terms like 'modifier', 'argument', and 'predicate' are semantically loaded. But it is quite possible for items with distinctive semantic properties to also have distinctive syntactic ones. We return to this matter in §4 below.

3. Learnable syntactic patterns: PAC

A different idea about a shortcoming of the Gold framework for learning sketched above is that it does not accommodate 'noise' of any kind (coughs, slips, false starts, intrusions of other languages), and the exact identification criterion of success is too strict. We might get a rather different picture of what is required for learning by adopting a probabilistic criterion of success. One proposal of this kind is presented by Valiant (1984).[7] Suppose that a learner is presented with expressions according to some probability distribution μ, where each expression is categorized as either being in the target language L or not. In this setting, we can quantify the degree to which the learner's hypothesis h misses the target by letting it be the probability of expressions in $L - h$ and $h - L$.



As before the learner is a function from samples to hypotheses h, but now the samples are drawn according to some arbitrary probability μ and classified according to whether they are in the target language or not. We say a class of languages (or 'concepts') is learnable if the learner will always be 'probably approximately correct' (PAC) after some number m of examples, where m can depend on how probably δ we want to be approximately ϵ correct. A class $\mathcal{L}$ is 'PAC learnable' if and only if there is a learner and a function m such that for all probability distributions μ, for every language L in $\mathcal{L}$, for every level of confidence $0 < \delta < 1$ and every margin of error $0 < \epsilon < 1$, the learner's guess after $m(\epsilon, \delta)$ samples will be a hypothesis h, where the probability that the hypothesis is within ϵ of the target is at least $1 - \delta$:

$$\mu(error_\mu \leq \epsilon) \geq (1 - \delta).$$

Mastering this success criterion takes some study, but it has the very nice property that a learner can be counted as successful even when some extremely rare expressions would be misclassified. And some classes of languages can be PAC learned by learners that only revise their hypotheses in response to 'positive data', data that is classified as being in the target language. Furthermore, the PAC criterion has been shown to be the discrete analog of a standard criterion for the consistent statistical approximation of real-valued functions by 'empirical risk minimization' and related methods.[8] The classes $\mathcal{L}$ that are learnable in this sense turn out to have an elegant combinatorial characterization: a class $\mathcal{L}$ is PAC learnable if and only if it has finite 'VC dimension'.[9]

This convergence of results, the coincidence of independently proposed criteria of success on such a simple combinatorial bound, suggests that this work has in fact identified a robust and natural notion. So it is perhaps no surprise that some researchers have proposed

> Applying this approach to natural language... one concludes that the family of learnable grammars must have a finite Vapnik Chervonenkis (VC) dimension. (Niyogi 2004, p. 941; cf. also Poggio et al. 2004)

This would be a very significant restriction on the class of available languages. But the proposal requires an important qualification.

Many classes with infinite VC dimension are efficiently learnable in other looser but still reasonable senses. For example, consider the problem of learning conjunctions of positive or negated atomic propositions (these conjunctions are often called 'monomials') from a sample of the situations of models that makes them true, in a propositional calculus with infinitely many propositional symbols. This space has infinite VC dimension, but if we 'parameterize' the space by the number n of proposition symbols used, then the complexity of the learning problem grows only polynomially with respect to n, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ (Kearns and Vazirani, 1994, Thm.1.2). When we consider language-oriented problems like learning reversible languages, we find that the space of reversible languages has infinite VC dimension.[10] But in this case, it has been difficult to find a way to parameterize the problem to appropriately reveal its efficiency.[11]

In sum, to expect finite VC dimension for the available human languages when we do not find it for the monomials or the reversible languages seems unreasonable. A possible response is to say that, in a clear sense, the *whole class* of monomials and the *whole class* of reversible languages is not efficiently learnable. That's true, but in the first place, on reasonably sized reversible language learning problems, Angluin's learner is efficient. And in the second place, there seems no principled (linguistic or cognitive) dividing line between the 'reasonably sized' problems that we are likely to encounter and the rest.

A more important point about this direction of research is this: the adoption of the PAC success criterion or something similar obviously does not address the main concern mentioned at the end of the previous section.

That is, we have not discovered how to define the kinds of generalizations made by human learners, and our universals of language complexity in §1 were rather weak, so these models do not yet explain the sorts of similarities across languages noticed by linguists.

4. Syntactic-semantic relations: languages as logics

The common descriptions of language are all semantically-laden.[12] Subjects, objects, predicates, modifiers, names, anaphors, etc. – these are all traditionally identified with criteria that are at least in part semantic. The typological universals identified by Greenberg and others in the 1960's, are all expressed in such terms, as are more recent proposals in that tradition (Hawkins, 2005, for example). Much of recent syntactic theory is so semantically-laden that the distinctions between semantic and syntactic arguments can be difficult to discern. Furthermore, psychological studies of acquisition confirm the commonsense idea that children and other language learners use multiple cues to figure out what is meant by utterances. For example, in one recent paper we find this suggestion:

> ... the learning procedure in some way makes joint use of the structures and situations that cooccur with verbs so as to converge on their meanings. Neither source of evidence is strong or stable enough by itself, but taken together they significantly narrow the search space. (Lidz, Gleitman, and Gleitman, 2004)

Can we provide computational models of how this works? There has been much activity in this area – much of it focused on making sense of the Augustinian (398) idea that the meaning of a word like 'cat' might be determined in part by noticing a common element in many of the situations where that word is used. But here we will very briefly discuss two fundamental questions about the potential and limits of such learning strategies: What is the nature of the fit between syntax and semantics such that a learner could expect to find semantic evidence of syntactic structure, and vice versa? And what kind of compositional structure do we find in human languages?

4.1. The syntactic/semantic fit and 'bootstrapping'

The fundamental approaches to learning discussed in §§2-3 extend immediately to language learning situations where the target is a grammar that defines form-meaning associations, and where the samples available to the learner are (at least sometimes) of this form too. It is completely clear that the availability of both forms and meanings in the data completely changes the situation! For example, in the Gold paradigm, it is obvious that some (sentence,meaning) texts are learnable where the text of sentences alone is not. We can prove this with a simple example. Consider any class of languages containing all the finite languages and an infinite language – known to be unlearnable in the Gold sense, as discussed in §2. Now pair each

sentence with a meaning in the following way: let the expressions in the finite languages all have distinct meanings (e.g. let them each denote distinct numbers), but let all the expressions in the infinite language all denote the same thing (e.g. the number 1). Then, after seeing any 2 different (sentence,meaning) pairs, the learner is in a position to know whether the target language is infinite or not, and in either case, the learner has a strategy for successful identification. It is also easy to define unlearnable classes of (sentence,meaning) languages where the corresponding sentence-only languages are easily identifiable. So we conclude immediately: when paired semantic and syntactic information is available, the nature of the learning problem varies fundamentally with the nature of the syntax-semantics relation.

Simple, artificial logics provide some useful examples of languages where the meanings of expressions is not arbitrary. A logic is typically given by a syntax that defines a set of expressions, a semantics that associates these sequences with semantic values, and an inference relation that is defined on expressions but which also preserves some semantic property. In systems like this, there is a fit between syntactic and semantic properties; for example, expressions that semantically denote binary functions on truth values are syntactically elements that combine with two sentential expressions. More interestingly, when there is a syntactic restriction on the number of elements that play a certain syntactic role, the elements with that syntactic role typically denote in semantic domains that are similarly restricted.

This perspective is extended to human languages by Keenan and Stabler (2003) with particular attention to the extreme case of the 'syntactic constants', elements that play unique roles in the grammar. While one proper name can typically be replaced by any another without changing structure in any human language, the syntactic constants are those words with unique roles, elements that cannot be replaced by any other, in the expressions of the language. In every sentence of standard English, for example, we can replace the name *Bill* by *Sam* without affecting structural properties, but there is no other element that can replace every occurrence of the infinitival *to*; no other element can replace the auxiliary *be*; and so on for many other elements. These syntactic constants, 'grammatical words', have a semantic distinction too: on any reasonable approach, they do not denote the same kinds of things that things like names or transitive verbs denote. Rather, they tend to denote 'semantic constants', that is, semantic values that are constant in the sense (roughly) that they do not depend on which individuals have which properties.[13]

This is of particular interest in the present context for two reasons. In the first place, it defines a setting in which various kinds of syntactic evidence could bear on semantic properties, and vice versa. Clearly, in this kind of setting, it is possible to get evidence about the semantic values of elements that could not be learned with the Augustinian method of correlating utterances with the situations of utterance. If the learner has access to both syntax and to syntactically characterizable relations of plausible inference,

then obviously the bearing on semantic hypotheses can be even more direct (Stabler, 2005).

A second reason to take note of this kind of fit between syntax, semantics and inference is that very prominent directions in current syntactic research suggest that human languages may tie semantic value and syntactic properties together very tightly across languages. For example, Szabolcsi has proposed in a series of papers that quantifiers of various kinds occupy distinct syntactic positions across languages (Szabolcsi, 1996; Szabolcsi and Brody, 2003). And Cinque has proposed in a series of works that, across languages, adverbial elements appear in a fixed order (Cinque, 1999; Cinque, 2001). These are but two examples from an enormous range of proposals that share the idea that the ties between syntactic role and semantic values may be very rich indeed.

4.2. Compositional structure

Although it is now a commonplace that complex expressions take their semantic values as a function of the semantic values of their parts, it is still difficult to formulate this idea in a fully general, precise and substantial way, so that it can underpin substantial linguistic universals, and so that we can properly understand its relation to language acquisition and use.

Suppose we think of a human language as the set of expressions generated from a lexicon by some structure building rules (again setting aside the worry that we want to factor the grammar into a syntax and some kind of morphology, or other parts). To allow for structural ambiguity, let's regard the semantics as assigning semantic values to derivations. Then a simple idea about compositional structure is this: the lexical elements have meanings, and with each way of composing the parts is associated a function from the meanings of the parts to the meanings of the resulting complex. The language learner can master the whole language by identifying the meanings of the parts and the semantic significance of the ways of combining expressions. This simple picture cannot be right. It raises a puzzle about how the language learner could proceed: What evidence from situations could lead the language learner to the meanings of each component of that phrase? And we would like a solution to this puzzle that is compatible with the fact that human languages have so many idioms, so many complex expressions with meanings that seem idiosyncratic – not just phrasal idioms (*kick the bucket, pop the question, chew the fat,...*) but also idiomatic compounds and fixed phrases (*by and large, in short, every which way, do away with, spick and span, break a leg, monkey wrench, sunflower, traffic light, deadline, ...*), and special verb-particle constructions (*take up/down/in/out/back/over, turn up/down/over/in/out/around/off, hold up/down/off/out/over,...*).

It is reasonable to suppose that, at least to a good first approximation, the learner's first evidence about what the morphemes of a language are is not semantic but combinatorial (Harris, 1957; Brent, 1999; Goldsmith, 2006). In that case, language learners may sometimes realize that an expression is complex and sometimes even have a good idea from situational

cues what it means, but not know the meanings of the parts. Under what circumstances can such a learner proceed from information about the meanings of sentences to an idea about the meanings of the parts of those sentences? One approach to this idea has been inspired by a simple proposal from (Frege, 1884, §60): "It is enough if the sentence as a whole has meaning; it is this that confers on its parts also their content."

Following Hodges (2001) and Westerståhl (2004), suppose a language is given by a lexicon $\{a, b, c, \ldots, a_1, b_1, \ldots\}$ together with some rules $\{f_1, f_2, \ldots f_n\}$ for building complexes. A derivation can be given as a function expression like $f_1(a, b)$, and the semantics μ can be a partial function from these derivations into some semantic domain. Let's say that two derivations d_1, d_2 are synonymous with this semantics, $d_1 \equiv_\mu d_2$, just in case they have the same meaning $\mu(d_1) = \mu(d_2)$. Then, the language is compositional if for every rule f, $f(a_1, \ldots, a_n) \equiv_\mu f(b_1, \ldots, b_n)$ whenever for each i between 1 and n, $\mu(a_i) = \mu(b_i)$. In other words, the semantics is compositional if substituting one synonymous element for another in a derivation leaves the meaning of the complex unchanged. In this setting, suppose that a language learner has acquired a fragment of English with a total compositional semantics, and then the speaker hears an idiom like *let the cat out of the bag* for the first time, with some evidence that it means something like *reveal the secret*. Now it is clear that there are several ways to provide a compositional extension of the language that accommodates this. One can maintain compositionality by assuming the rules assembling the idiom are different from the usual ones, leaving us free to set the interpretation of the idiom even when the parts have exactly their usual meanings. But a more appealing extension introduces new senses for some of the words – for example *cat* could be interpreted as the 'secret', and *bag* might even be interpreted as 'concealment', in which case the complex could be interpreted compositionally. But for idioms lacking compositionally interpretable parts (like perhaps *kick the bucket*), new senses could be introduced with no meanings specified for them.

In sum, a simple version of compositionality appears to be compatible with a reasonable range of proposals about how a language learner might handle idioms and collocations of various kinds. Something like this seems to represent a consensus now in the field,[14] and while much remains mysterious in our models of language recognition and production, this picture seems compatible with results of acquisition research.[15] One hopes that this recent work may lead to models of language acquisition and use that will properly predict the range of idiomatic constructions found across languages, and be part of a picture in which we can make sense of the way language learners identify semantic properties of linguistic elements. These preliminaries appear to be essential first steps towards a realistic conception of how semantically characterized elements appear in the configurations we find across languages.

5. Conclusions

The fact that humans notice certain kinds of patterns in small samples of sentences, patterns that extend well beyond the sample, has the consequence that many languages cannot be learned (as we see, in different ways, in the formal results of Gold, Angluin, Valiant). It is natural to assume that the human way of doing this will determine some structural universals of human languages. Our understanding of this matter has been shaped by two rather recent developments. First, there has been an astounding convergence among grammar formalisms on a certain 'mildly context sensitive' (MCS) level of combinatorial complexity. There are still many controversies in this area, but it appears that Joshi's hypothesis that human languages are MCS may be right or very close to right. Second, there has been a remarkable convergence among independent characterizations of the learnable patterns. Again, there are controversies, but it is clear that VC dimension provides a relevant measure of learning complexity. With these two major convergences, our understanding of language structure and learning complexity has advanced considerably.

We began by asking: Do nontrivial universal properties of language structure reflect important properties of human language learning? And do some of these structural properties guarantee that the class of all languages with those properties is a 'learnable' class in some relevant sense? This last question certainly must be answered negatively if by 'structural' we refer to purely syntactic structure. Since purely structural properties typically determine neither the sequences of perceptible forms nor their semantic properties, structure alone is not enough to determine the learning problem. Human language learning depends, for example, on the fact that the pronunciation *dog* does not change arbitrarily in every utterance, and on the fact that many utterances of *dog* are perceptibly different from at least many other morphemes, but these are not matters of syntactic structure. So to set the stage for learning structural properties of expressions, we need to worry about the bounds or pressures limiting syncretism and accidental homophony. Standardly, one proceeds by adopting simplifying assumptions, assuming that such bounds will be forthcoming. Still, our initial questions remain largely open.

The standard idea about why these questions remain open is that we must first bridge the gulf between the perceptible properties of unanalyzed linguistic input and the terms of linguistic analysis (cf. note 2). In computational models, we can stipulate that one or another expression is a subject or a modifier or whatever, but this does not take us toward an explanation until the stipulations are understood and hence removable. To get to such terms without stipulation, it is typically assumed that semantic cues may be essential, as discussed at the end of §2 and in §4, but even this is not established. In particular, the role of grammatical constants as indicators of structure has not been fully explored. When the computational bases of

traditional analyses are better understood, then can hope for explanations of how traditional universals emerge.

Obtaining clear open questions is an achievement, and the terms for addressing them must derive from secure foundations in learning and complexity theory. The computational foundations reviewed here are almost entirely new since the 1961 meeting on universals. It is certain that the next 40 years will yield a much deeper and more comprehensive understanding of language universals and how they emerge.

Acknowledgments. Parts of this work were presented at Cornell University, Massachusetts Institute of Technology, University of Chicago, and UCLA, where audience suggestions were helpful. Thanks especially to Jeff Heinz. Aravind Joshi, Ed Keenan, Greg Kobele, Marcus Kracht, Tony Kroch, Lillian Lee, Partha Niyogi, Katya Pertsova, Stuart Shieber, and Sarah VanWagenen.

Notes

[1]This is standardly recognized in the literature. For example, in Chomsky's 1956 paper we find,

> We might avoid this consequence by an arbitrary decree that there is a finite upper limit to sentence length in English. This would serve no useful purpose, however... (Chomsky, 1956, p.115)

And in recent introductions to formal languages and computation, we find remarks like this:

> Viewing the computer as a finite state system... is not satisfying mathematically or realistically. It places an artificial limit on memory capacity, thereby failing to capture the real essence of computation. (Hopcroft and Ullman, 1979, p.14)

The suggestion is not that the limitations are unimportant or uninteresting, but just that we may get a more revealing and accurate understanding from an account that factors the definition of grammatical patterns away from the interfering factors.

[2]This point has been noted often before in both linguistic work and studies of language acquisition. For example,

> In short, a problem that is central to understanding the learning of syntax is that of arriving at a theory of how the child determines appropriate base structures for the types of sentences that appear in the corpus. However, the peculiarly abstract relation between base structures and sentences unfits any of the usual learning mechanisms for explaining their assimilation. (Fodor, 1966, p.113)

> ... in the case of Universal Grammar... we want the primitives to be concepts that can plausibly be assumed to provide a preliminary, prelinguistic analysis of a reasonable selection of presented data. it would be unreasonable to incorporate, for example, such notions as subject of a sentence or other grammatical notions, since it is unreasonable to suppose that these notions can be directly applied to linguistically unanalyzed data. (Chomsky, 1981, p.10)

> The problem with almost every nonsemantic property that I have heard proposed as inductive biases is that the property is itself defined over abstract symbols that are part of the child's input, that themselves have

> to be learned. For example, some informal proposals I have heard start from the assumption that the child knows the geometry of the phrase structure tree of a sentence, or, even worse, the syntactic categories and grammatical relations of phrases. (Pinker, 1984, p.51)

[3]One promising approach to formalizing the similarity among the derivations of different formalisms uses tree transducers: similar derivations may be related by relatively weak kinds of transducers. Cf., e.g. Shieber (2005); Engelfriet and Vogler (1985).

[4]Among the most important results on earlier extremely expressive formalisms are the Peters and Ritchie (1973) result on a certain 'standard theory' transformational grammar, and the series of results on 'unification grammar' formalisms (Berwick, 1981; Johnson, 1988; Trautwein, 1995; Torenvliet and Trautwein, 1995).

[5]This same machine is represented by the following rewrite grammar:

0→dp 1	2→ ϵ
1→vp 2	3→ ϵ
2→pp 3	

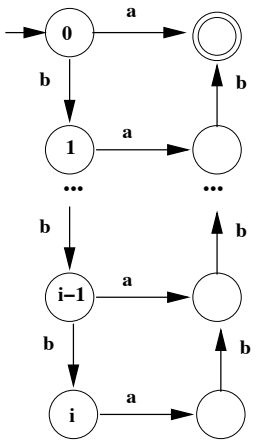
[6]The seminal work of Wexler and Culicover (1980) defines a learnable class and proposes a number of interesting structural properties for human syntax, but it is based on an early grammatical framework; a penetrating analysis of this work is provided by Osherson, Weinstein, and Stob (1986), showing that it allows only a finite range of grammars. There are many other less interesting ways to restrict the range human grammars to a finite set (Chomsky, 1981; Osherson, Weinstein, and Stob, 1984; Pinker, 1982); one can simply stipulate that only a finite set of (unspecified) properties are really relevant 'core' properties; and of course many non-linguistic factors (lifespan, attention span) conspire to make our concerns finite. After one such proposal it is appropriately remarked "As with many learnability presentations, the account just given has an air of science fiction about it" (Pinker, 1982, p.675). Compare also the remarks in note 1. The interesting question is whether there really are principled, empirically motivated, syntactic properties of language that guarantee learnability, properties which would explain the sorts of properties traditionally noticed by linguists (Greenberg, 1966, for example). As suggested at the outset, what we really want to know is how human learners generalize from the finite data they are given, such that we end up with languages like the ones we have.

[7] This proposal and important variants are now introduced in several good texts. See for example, Kearns and Vazirani (1994), Alpaydin (2004), Anthony and Biggs (1992).

[8]Alon et al. (1997), Mendelson (2003), Mukherjee et al. (2004), Poggio et al. (2004).

[9]This result is from Blumer et al. (1989), building on Vapnik and Chervonenkis (1971). 'VC dimension' is introduced and carefully discussed in the texts listed in note 7, and in the Blumer et al. (1989) paper. The definition is very simple. Given an arbitrary subset S of the domain, if $\{L \cap S |\ L \in \mathcal{L}\}$ is the set of all subsets of S, then we say S is 'shattered' by the class $\mathcal{L}$. The VC dimension of $\mathcal{L}$ is the size of the largest set Y that is shattered by $\mathcal{L}$.

[10]It is obvious that the space $\mathcal{L}_{rev}$ of reversible languages has infinite VC dimension, but I have not seen this in the literature before, so I sketch a proof. We need to show that there is no finite bound on the size of the sets that are shattered by $\mathcal{L}_{rev}$. For any finite $k \geq 0$, let the language $L_k = \{b_i a b_i |\ 0 \leq i \leq k\}$. It is clear that every such L_k is reversible:



Furthermore, it's clear that every subset of this language is reversible, since it will be defined by the result of deleting any number of the a-arcs (and then removing any states and arcs that are not on a path from the start state to the final state). Clearly the size of L_k grows with k, and every such set can be shattered by $\mathcal{L}_{rev}$ since every subset of L_k is also reversible.

[11]Cf. Pitt (1989), Freund et al. (1997), Head, Kobayashi, and Yokomori (1998), Yokomori (2003).

[12]Jakobson dismisses attempts to separate syntactic and semantic components of language at the 1961 Conference on Universals with the amusing remark, "Fortunately, in his quest for universals of grammar Greenberg does not share the whimsical prejudice against 'semantics oriented definitions,' which, strange as it seems, may have filtered even into our Conference on Language Universals" (Jakobson, 1963, p.271).

[13]These notions are defined precisely in Keenan and Stabler (2003). What we here call 'semantic constancy' is sometimes called 'isomorphism invariance' or 'permutation invariance'. This is a familiar notion in semantics (van Benthem, 1986, §1.7, for example), closely following the classical notions of invariance from Klein (1893) and Tarski (1986).

[14]Cf., e.g., Westerståhl (2004), Hodges (2001), McGinnis (2002), Nunberg, Wasow, and Sag (1994), Keenan and Stabler (2003), Kracht (1998). Contrasting views can be found in earlier work like Jackendoff (1997), Di Sciullo and Williams (1987).

[15]In particular, some studies indicate that children do, in fact, look for compositional analyses of new phrases (Gibbs, 1991; Gibbs, 1994).

References

Alon, Noga, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. 1997. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the Association for Computing Machinery*, 44(4):615–631.

Alpaydin, Ethem. 2004. *Introduction to Machine Learning*. MIT Press, Cambridge, Massachusetts.

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

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

Anthony, Martin and Norman Biggs. 1992. *Computational Learning Theory*. Cambridge University Press, NY.

Augustine. 398. *Confessions*. Reprinted with commentary by J.J. O'Donnell. NY: Oxford University Press, 1992.

Berwick, Robert C. 1981. Computational complexity of lexical functional grammar. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics, ACL'81*, pages 7–12.

Blumer, Anselm, Andrzei Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36:929–965.

Bobaljik, Jonathan David. 2002. Syncretism without paradigms: Remarks on Williams 1981, 1994. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 2001*. Kluwer, Dordrecht, pages 53–85.

Bowerman, Melissa. 1988. The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J.A. Hawkins, editor, *Explaining Language Universals*. Blackwell, Oxford.

Braine, Martin. 1971. On two types of models of the internalization of grammars. In D.I. Slobin, editor, *The Ontogenesis of Grammar: A Theoretical Symposium*. Academic Press, NY.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, IT-2:113–124.

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

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads : A Cross-Linguistic Perspective*. Oxford University Press, Oxford.

Cinque, Guglielmo. 2001. The status of 'mobile' suffixes. In Walter Bisang, editor, *Aspects of Typology and Universals*. Akademie Verlag, Berlin, pages 13–19.

Di Sciullo, Anna Maria and Edwin Williams. 1987. *On the definition of word*. MIT Press, Cambridge, Massachusetts.

Engelfriet, Joost and Heiko Vogler. 1985. Macro tree transducers. *Journal of Computer and System Sciences*, 31(1):71–146.

Fodor, Jerry A. 1966. How to learn to talk: some simple ways. In Frank Smith and G.A. Miller, editors, *The Genesis of Language: A Psycholinguistic Approach*. MIT Press, Cambridge, Massachusetts, pages 105–122.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik*. Koebner, Breslau. J.L. Austin's translation available as *The Foundations of Arithmetic*, Evanston, Illinois: Northwestern University Press, 1980.

Freund, Yoav, Michael Kearns, Dana Ron, Ronnit Rubinfeld, Robert E Schapire, and Linda Sellie. 1997. Efficient learning of typical finite automata from random walks. *Information and Computation*, 138:23–48.

Gärtner, Hans-Martin and Jens Michaelis. 2005. A note on the complexity of constraint interaction. In *Logical Aspects of Computational Linguistics, LACL'05*, Lecture Notes in Artificial Intelligence LNCS-3492. Springer, NY, pages 114–130.

Gibbs, Raymond W. 1991. Semantic analyzability in children's understanding of idioms. *Journal of Speech and Hearing Research*, 35:613–620.

Gibbs, Raymond W. 1994. *The Poetics of Mind*. Cambridge University Press, NY.

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

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12.

Greenberg, Joseph. 1963. *Universals of Language: report of a conference held at Dobbs Ferry, New York, April 13-15, 1961*. MIT Press, Cambridge, Massachusetts.

Greenberg, Joseph. 1966. Some universals of language with particular attention to the order of meaningful elements. In *Universals of Human Language*. MIT Press, Cambridge, Massachusetts.

Harkema, Henk. 2001. A characterization of minimalist languages. In Philippe de Groote, Glyn Morrill, and Christian Retoré, editors, *Logical Aspects of Computational Linguistics*, Lecture Notes in Artificial Intelligence, No. 2099, pages 193–211, NY. Springer.

Harris, Zellig S. 1957. Cooccurrence and transformations in linguistic structure. *Language*, 33:283–340. Reprinted in H. Hiz, ed., *Papers on Syntax*, Reidel: Boston, 1981.

Hawkins, John A. 2005. *Efficiency and Complexity in Grammars*. Oxford University Press, NY.

Head, Tom, Satoshi Kobayashi, and Takashi Yokomori. 1998. Locality, reversibility, and beyond: learning languages from positive data. In M.M. Richter, C.H. Smith, R. Wiehagen, and T. Zeugmann, editors, *Proceedings of the 9th International Conference on Algorithmic Learning Theory, ALT'98*, volume 1501 of *Lecture Notes in Artificial Intelligence*, pages 191–204. Springer-Verlag.

Heinz, Jeffrey. 2006. Learning quantity insensitive stress systems via local inference. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group in Computational Phonology at HLT-NAACL*, pages 21–30. NY.

Hodges, Wilfrid. 2001. Formal features of compositionality. *Journal of Logic, Language and Information*, 10:7–28.

Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman. 2000. *Introduction to Automata Theory, Languages and Computation (2nd Edition)*. Addison-Wesley, Reading, Massachusetts.

Hopcroft, John E. and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, Massachusetts.

Jackendoff, Ray S. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, Massachusetts.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems that Learn: An Introduction to Learning Theory (second edition)*. MIT Press, Cambridge, Massachusetts.

Jakobson, Roman. 1963. Implications of language universals for linguistics. In Joseph Greenberg, editor, *Universals of Language: report of a conference held at Dobbs Ferry, New York, April 13-15, 1961*. MIT Press, Cambridge, Massachusetts.

Johnson, Mark. 1988. *Attribute Value Logic and The Theory of Grammar*. Number 16 in CSLI Lecture Notes Series. CSLI Publications, Chicago.

Joshi, Aravind. 1985. How much context-sensitivity is necessary for characterizing structural descriptions. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Processing: Theoretical, Computational and Psychological Perspectives*. Cambridge University Press, NY, pages 206–250.

Joshi, Aravind K., K. Vijay-Shanker, and David Weir. 1991. The convergence of mildly context sensitive grammar formalisms. In Peter Sells, Stuart Shieber, and Thomas Wasow, editors, *Foundational Issues in Natural Language Processing*. MIT Press, Cambridge, Massachusetts, pages 31–81.

Kearns, Michael J. and Umesh V. Vazirani. 1994. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts.

Keenan, Edward L. and Edward P. Stabler. 2003. *Bare Grammar*. CSLI Publications, Stanford, California.

Klein, Felix. 1893. A comparative review of recent researches in geometry. *Bulletin of the New York Mathematical Society*, 2:215–249. Translation by M.W. Haskell of the original October 1872 publication, with a prefatory note by the author.

Kobele, Gregory M. 2005. Features moving madly: A note on the complexity of an extension to MGs. *Research on Language and Computation*, 3(4):391–410.

Kobele, Gregory M. 2006. *Generating Copies: An investigation into structural identity in language and grammar*. Ph.D. thesis, UCLA.

Kobele, Gregory M. and Jens Michaelis. 2005. Two type 0 variants of minimalist grammars. In *Proceedings of the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language, FGMOL05*.

Kracht, Marcus. 1998. Strict compositionality and literal movement grammars. In *Proceedings, Logical Aspects of Computational Linguistics, LACL'98*. Springer, NY, pages 126–142.

Lewis, H.R. and C.H. Papadimitriou. 1981. *Elements of the Theory of Computation*. Prentice-Hall, Englewood Cliffs, New Jersey.

Lidz, Jeffrey, Henry Gleitman, and Lila R. Gleitman. 2004. Kidz in the 'hood: Syntactic bootstrapping and the mental lexicon. In D.G. Hall and S.R. Waxman, editors, *Weaving a Lexicon*. MIT Press, Cambridge, Massachusetts, pages 603–636.

McGinnis, Martha. 2002. On the systematic aspect of idioms. *Linguistic Inquiry*, 33(4):665–672.

Mendelson, Shahar. 2003. Geometric parameters in learning theory. Manuscript, Research School of Information Sciences and Engineering, Australian National University, submitted.

Michaelis, Jens. 1998. Derivational minimalism is mildly context-sensitive. In *Proceedings, Logical Aspects of Computational Linguistics, LACL'98*, NY. Springer.

Michaelis, Jens. 2001. Transforming linear context free rewriting systems into minimalist grammars. In Philippe de Groote, Glyn Morrill, and Christian Retoré, editors, *Logical Aspects of Computational Linguistics*, Lecture Notes in Artificial Intelligence, No. 2099, pages 228–244, NY. Springer.

Michaelis, Jens and Marcus Kracht. 1997. Semilinearity as a syntactic invariant. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, pages 37–40, NY. Springer-Verlag (Lecture Notes in Computer Science 1328).

Mukherjee, Sayan, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. 2004. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of Empirical Risk Minimization. *Advances in Computational Mathematics*. forthcoming.

Nunberg, Geoffrey, Thomas Wasow, and Ivan A. Sag. 1994. Idioms. *Language*, 70(3):491–538.

Osherson, Daniel, Scott Weinstein, and Michael Stob. 1984. Learning theory and natural language. *Cognition*, 17:1–28.

Osherson, Daniel, Scott Weinstein, and Michael Stob. 1986. An analysis of a learning paradigm. In William Demopoulos and Ausonio Marras, editors, *Language Learning and Concept Acquisition*. Ablex, Norwood, NJ, pages 103–116.

Pertsova, Katya. 2006. Lexical meanings of morphemes. 80th Annual Meeting of the LSA, Albuquerque, New Mexico.

Peters, P. Stanley and R. W. Ritchie. 1973. On the generative power of transformational grammar. *Information Sciences*, 6:49–83.

Pinker, Steven. 1982. A theory of the acquisition of lexical interpretive grammars. In Joan Bresnan, editor, *The mental representation of grammatical relations*. MIT Press, Cambridge, Massachusetts.

Pinker, Steven. 1984. *Language Learnability and Language Development*. Harvard University Press, Cambridge, Massachusetts.

Pitt, Leonard. 1989. Inductive inference, DFAs, and computational complexity. In K.P. Jantke, editor, *Workshop on Analogical and Inductive Inference*, volume 397 of *Lecture Notes in Artificial Intelligence*, pages 18–44, Berlin. Springer Verlag.

Poggio, Tomaso, Ryan Rifkin, Partha Niyogi, and Sayan Mukherjee. 2004. General conditions for predictivity in learning theory. *Nature*, 428:419–422.

Pullum, Geoffrey K. 2006. On syntactically mandated phrase reduplication. Presented at MIT.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229.

Shieber, Stuart M. 2005. Synchronous grammars as tree transducers. In *Proceedings of the Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+ 7)*.

Stabler, Edward P. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*. Springer-Verlag (Lecture Notes in Computer Science 1328), NY, pages 68–95.

Stabler, Edward P. 2004. Varieties of crossing dependencies: Structure dependence and mild context sensitivity. *Cognitive Science*, 93(5):699–720.

Stabler, Edward P. 2005. Natural logic in linguistic theory. Proof Theory at the Syntax/Semantics Interface, LSA Institute Workshop. Available at `http://wintermute.linguistics.ucla.edu/prooftheory`. Revised version forthcoming.

Szabolcsi, Anna. 1996. Strategies for scope-taking. In Anna Szabolcsi, editor, *Ways of Scope Taking*. Kluwer, Boston.

Szabolcsi, Anna and Michael Brody. 2003. Overt scope in Hungarian. *Syntax*, 6:19–51.

Tarski, Alfred. 1986. What are logical notions? *History and Philosophy of Logic*, 7:143–154.

Torenvliet, Leen and Marten Trautwein. 1995. A note on the complexity of restricted attribute-value grammars. In *Proceedings of Computational Linguistics In the Netherlands, CLIN5*, pages 145–164.

Trautwein, Marten. 1995. The complexity of structure-sharing in unification-based grammars. In *Proceedings of Computational Linguistics In the Netherlands, CLIN5*, pages 165–180.

Valiant, Leslie. 1984. A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142.

van Benthem, Johan. 1986. *Essays in Logical Semantics*. Reidel, Dordrecht.

Vapnik, V.N. and A.Y. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.

Vijay-Shanker, K. and David Weir. 1994. The equivalence of four extensions of context free grammar formalisms. *Mathematical Systems Theory*, 27:511–545.

Westerståhl, Dag. 2004. On the compositional extension problem. *Journal of Philosophical Logic*, 33(6):549–582.

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

Williams, Edwin. 1994. Remarks on lexical knowledge. *Lingua*, 92:7–34. Reprinted in Lila Gleitman and Barbara Landau, eds., *The Acquisition of the Lexicon*, MIT Press, 1994.

Yokomori, Takashi. 2003. Polynomial-time identification of very simple grammars from positive data. *Theoretical Computer Science*, 298:179–206.