The Significance of Errors to Parametric Models of Language Acquisition

Paula Buttery and **Ted Briscoe**
Natural Language and Information Processing Group
Computer Laboratory, University of Cambridge
JJ Thomson Avenue, Cambridge CB3 0FD, England
{paula.buttery, ted.briscoe}@cl.cam.ac.uk

Abstract

The aim of this research is to investigate the process of grammatical acquisition from real data. In this paper we address the issue of errors. We demonstrate by simulation how a learning system may be robust when statistical error handling methods are employed.

Classification of Input Data

A normal child becomes rapidly fluent in their native language despite an absence of any formal language teaching. The child is exposed to evidence of her target language that must exclusively belong to one of three possible classes: positive evidence is information that describes which utterances are allowed in the target language; negative evidence is information that describes which utterances are *not* allowed in the target language; errors are pieces of information that have been mistakenly classified as either positive or negative evidence.

Positive Evidence

Positive evidence can be presented to a child in the form of example utterances spoken by proficient members of her language community. A large proportion of the language a child is exposed to will be positive evidence. In fact Pinker (Pinker 1994) goes as far as saying that "effectively the input to the learner *only* includes grammatical sentences". Following Gold's paradigm (Gold 1967), a child hypothesises her language based on accumulated positive evidence; all previously heard utterances form a subset of the current hypothesised language. Learning is completed once the hypothesised language no longer needs to be updated.

Negative Evidence

Negative evidence might be provided by correcting a child when they produce an ungrammatical sentence. Evidence of this sort could be used to constrain the child from hypothesising a language that is a superset of the target language. A child that is only ever exposed to positive evidence can not be corrected if she hypothesises a language that is too large. However, in general children do not learn from correction (Brown & Hanlon 1970) this indicates that there must be some other mechanism for constraining the hypothesised language: possible solutions include MDL learning (Rissanen 1978), where the child only ever hypothesises the simplest language that describes the evidence seen so far, or an innately defined hypothesis space for language.

Errors

Lacking any discerning information, a child is likely to assume that all the utterances she hears are grammatical and therefore constitute positive evidence. However, spoken language can contain ungrammatical utterances, perhaps in the form of interruptions, lapses of concentration or slips-of-the-tongue. When a child mis-classifies such utterances as positive evidence, an error has occurred.

Situations also arise where entirely grammatical sentences can produce an error because of misclassification due to indeterminacy. For just one example consider the sentences "John kissed Kate" and "Kate was kissed by John"; young children are likely to misinterpret the subject of the passive sentence to be the agent, leading to a misclassification of word meaning. Also, indeterminacy of the input may lead to noise within the parameter settings of the universal grammar (Clark 1992); for instance, sentences of English (subject-verb-object ordering) can be misclassified as a subject-object-verb ordering with an active V2 (verb movement) parameter, as in German.

In general, any environment that contains ambiguity will produce errors. Often a child is exposed to input from more than one target language and yet manages to learn one (or more) consistent grammar(s), rather than a grammar that allows all possible combinations of the sampled input. Specific examples of this include diglossia (Kroch 1989) and language change (Lightfoot 1979). In such situations, misclassification of one of the input languages is an example of an error. Furthermore, there are documented situations where a conflicting and inconsistent input is "regularised" and fashioned into a single consistent generative grammar; as in the cases of rapid creolisation (Bickerton 1984) and the acquisition of sign language from a source of non-expert signers (the case of Simon (Newport 1999)).

Errors of these sorts are always accidental and lead to a false assignment of an utterance to the class of positive evidence. A child is somehow able to cope with such erroneous assignments.

A malicious error would occur if a deliberate attempt was made to confound the child's acquisition of language. An example might be if the child is corrected on her grammatically correct utterances or if she is deliberately exposed to utterances that are ungrammatical. Malicious errors are unlikely in spoken language but do occur in some very early child-directed-speech in the form of nonsense words. Such errors are likely to be presented to the child before she is acquiring lexical information and are therefore unlikely to be misclassified at all. In fact studies of parent-child interaction during a child's language learning period have tended to show that child-directed-speech is slow, clear, grammatical and quite repetitious (Snow 1977).[1]

To summarise: in an ideal learning situation a learner would have access to an oracle (Valiant 1984) that can correctly identify every utterance heard as either positive or negative evidence. However, a child learning its first language can not rely on receiving *any* negative evidence; at best she can hope to receive positive evidence or, more realistically, positive evidence that is interspersed with errors. The child would, of course, be unaware of when an error has been encountered. Any simulation or explanation of language acquisition should therefore attempt to learn from every utterance it encounters and should be unaffected by accidental erroneous utterances.

How do Errors effect Learning?

Consider a simplified learning problem, a game for two players: the first player, the exampler, thinks of a set of numbers that can be defined by a rule, such as multiples of two $\{x|x/2 \in Z\}$; the second player, the guesser, attempts to reproduce the set by discovering the rule which defines it. The only information available to the guesser is a continuous stream of examples provided by the exampler.

A possible scenario might be that the first two examples provided are 4 and 8. At this point the guesser may well hypothesise that the set contained multiples of four $\{x|x/4 \in Z\}$. The guesser doesn't need to revise this hypothesis until she encounters an example that breaks the rule. If the guesser ever arrives at the hypothesis that the set contains multiples of two she'll never have to revise her hypothesis again.[2]

Now if the same game was played in a noisy room or with a distracted exampler the guesser might receive erroneous examples. For instance, in attempting to guess the set $\{x|x/2 \in Z\}$, the guesser may have heard the examples 2, 4, 7, 8,... If the guesser classifies all the examples as positive evidence then there are two possible outcomes: either the guesser fails to find a rule or she hypothesises the wrong rule.

The guesser could only arrive at the correct hypothesis if she is aware that some of the examples may be erroneous. The guesser's best chance of winning is to figure out which hypothesis is *most likely* to be correct. Before the game begins the guesser will consider all hypotheses equally likely. As the game proceeds the working hypothesis is selected if it is the most likely given the accumulated evidence. In other words the guesser must adopt a statistical methodology to cope with the erroneous examples.

Introducing a hypothesis bias

Now, the interesting problem is: how many erroneous examples could the guesser encounter before she is completely unable to guess the rule. The answer lies in the type and frequency of the errors encountered as well as any bias the guesser may have towards certain hypotheses.

For example, consider the set $\{x|x/5 \in Z\}$. With no examples the guesser considers all hypotheses equally likely. After being exposed to the examples 15, 30, 45 the guesser has to consider the hypotheses $\{x|x/5 \in Z\}$ and $\{x|x/3 \in Z\}$ to be equally likely. Too many erroneous examples that happen to be multiples of three but not multiples of five may lead the guesser to eventually choose the later and incorrect hypothesis, $\{x|x/3 \in Z\}$. However, if the guesser had been initially biased towards the $\{x|x/5 \in Z\}$ hypothesis, perhaps because her favourite number is five, then she may have continued to select this hypothesis despite the accumulated evidence.

A Robust Learning System

Previous learning systems, ed. the Trigger Learning Algorithm (Gibson & Wexler 1994) and the Structural Trigger Learner (Fodor 1998), set parameters deterministically and as such rely on the input to the learning system being free from error. Considering the evidence above, this is obviously not realistic.

We demonstrate a learning system that is robust to the errors presented in real child-directed speech. We investigate two sources of error:

1. Errors introduced by indeterminacy of meaning of an utterance, i.e. misclassification of word meaning;
2. Errors introduced by indeterminacy in parameter setting.

In both cases the errors are successfully dealt with by using statistical methods.

System Overview

The learner is composed of three modules: a semantics learning module, syntax learning module, and a Universal Grammar module based on Chomsky's Principle and Parameter theory (Chomsky 1965).

For each utterance the learner is exposed to, we simulate the output from two cognitive systems; a *Speech Perception system* and a *Cognitive system* (Siskind 1996). Thus, the learner receives input as two streams; a stream of word symbols and a stream of semantic hypotheses (Buttery 2003b).

[1] It should be noted that this type of deliberate parent-child interaction is not considered to be a requirement for learning since there are societies where children are rarely directly spoken to before they can speak back (Pinker 1994). However, it is very likely to be helpful.

[2] The situation where the guesser hypothesises a rule that produces a superset of the original set is not discussed here. This situation would be avoided by allowing the exampler to provide negative evidence (i.e. examples of numbers not in the set) or by constraining the hypothesis space.

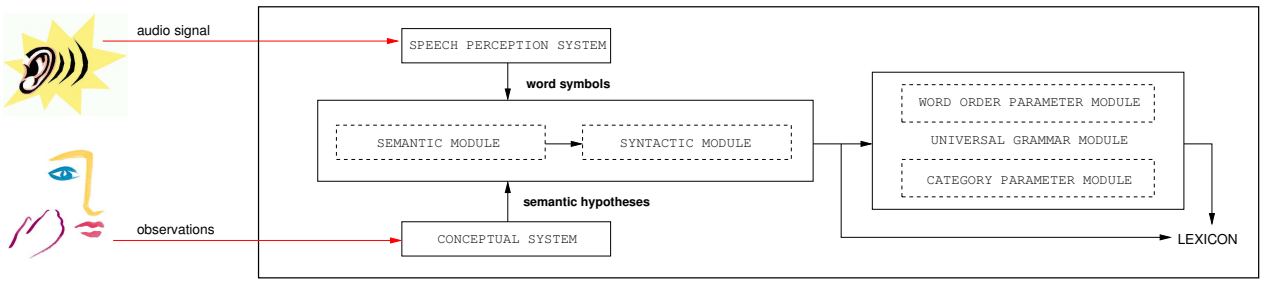

Figure 1: The Learning System: the interaction of the semantic, syntactic and UG modules

.

The semantics learning module receives these input streams attempts to learn the mapping between words and meanings, building a lexicon containing the meaning associated with each word sense. Subsequently, the syntax learning module tries to learn possible syntactic categories for a word sense, based initially on the semantic predicates associated with each word in the utterance. Finally, any syntactic information acquired is used to set the parameters of the Universal Grammar (Villavicencio 2002).

The Semantics Learning Module

The semantic module attempts to learn the mapping between word symbols and semantic symbols, building a lexicon containing the meaning associated with each word sense. This is achieved by analysing each input utterance and its associated semantic hypotheses using cross-situational techniques (following Siskind (1996)).

An example of one such technique is *Constraining Hypotheses with Partial Knowledge* which deals with the situation where the learner hypotheses more than one meaning for a given utterance. The method reduces the number of meanings by removing all those which are impossible given what has already been learnt. To show how this works, imagine the learner has heard the utterance "Mice like cheese" and hypothesised the following semantic expressions:

$$\mathbf{like}(\mathbf{mice}, \mathbf{cheese}) \quad (1)$$

$$\mathbf{madeOf}(\mathbf{moon}, \mathbf{cheese}) \quad (2)$$

$$\mathbf{madeOf}(\mathbf{moon}, \mathbf{cake}) \quad (3)$$

If the learner has already established that *cheese* maps to **cheese** then 3 can be ruled out as a possible meaning since it doesn't contain the necessary semantic expression. Hypothesis 2, however, can not be ruled out. If all semantic hypotheses are ruled out then the learner might assumes that one of the words in the utterance has multiple senses.

The Syntactic Learning Module

The learning system links the semantic module and syntactic module by using the typing assumption: *the semantic arity of a word is usually the same as its number of syntactic arguments.*

This idea is related to the Projection Principle (Chomsky 1981) which states that the selectional requirements of a word are projected onto every level of syntactic representation.

The module attempts to create valid parse trees starting from the syntactic information already assumed by the typing assumption (Buttery 2003a). A valid parse is one that is composed purely from legal syntactic category types; legal types are defined by the Universal Grammar.

The Universal Grammar Module

The Universal Grammar has been constructed according to Chomsky's Principle and Parameter theory (Chomsky 1965): *Principles* are characteristics that are common across languages (in this work the principles are represented by an underspecified inheritance hierarchy); *Parameters* are points of variation that will be fixed upon exposure to the linguistic environment. Two types of parameters have been used here:

1. *Categorial Parameters* determine whether a category from the principles hierarchy is in use within the learners current model of the input language. There are 60 categorial parameters; one associated with each legal syntactic category of the language. Setting these parameters places a constraint on the syntactic categories that may be hypothesised by the syntax learning module. The parameters themselves are set by the *Category Parameter Setting Module*.

2. *Word Order Parameters* determine the underlying order in which constituents (such as the subject or direct object) occur. There are 18 word order parameters that may be set to indicate whether constituents are generally located to the right or left of the constituent in question. These parameters are set by the *Word Order Parameter Setting Module*. An example of a word order parameter is the *subject direction parameter* which is used to indicate whether the subject is generally located to the left of the verb (SVO, SOV) or to the the right (OVS, VSO).

The Universal Grammar module is consulted whenever the syntactic learner returns a legal syntactic category for every word in the current utterance. The values of the parameters are then updated as appropriate, according to the syntactic

properties of the utterance.[3]

Experiment 1: Error introduced by indeterminacy of meaning

If a child is to ever use language in a purposeful manner she must not only determine which utterances belong to the language but also determine what they mean. In an ideal situation a child will hypothesise the correct meaning for every utterance heard. However, assigning a meaning is not straight-forward; there is unlikely to be only one obvious candidate. When the correct meaning is not hypothesised, error has been introduced.

An error of this type introduces a false association between word and meaning within the semantics learning module of the learner. To combat this problem a statistical error handler was adopted:

A confidence score was assigned to word meanings according to their frequency and consistency. Word meanings whose confidence scores fell below a threshold value were systematically pruned. We investigated the robustness of our learner under increasing levels of indeterminacy of meaning. This was achieved by associating utterances from the corpus with a set of possible semantic meanings (rather than just their single correct meaning). Indeterminacy could be increased by increasing the size of this associated meaning set and also by not including the actual meaning within the set.

Experiment 1a: The learner was run with increasing numbers of semantic hypotheses per utterance. The extra hypotheses were chosen randomly and the correct semantic expression was always present in the set. Hypotheses sets of sizes 2, 3, 5, 10 and 20 were used.

Experiment 1b: The learner was run with some utterances being completely mismatched with semantic hypotheses (i.e. the correct hypothesis was not present amongst the set).

Experiment 2: Error due to indeterminacy of parameter setting

It is claimed that children rarely mis-set syntactic parameters (Wexler 1998). However, proposed methods of parameter setting, such as the Trigger Learning Algorithm (TLA) (Gibson & Wexler 1994) or the Structural Trigger Learner (STL) (Fodor 1998), allow a parameter to be updated only once during acquisition. This type of approach can only work if a child exclusively receives positive evidence; if an error causes a parameter to be set incorrectly there is never a chance to reset it and consequently the correct grammar cannot be learnt. The problem is even worse for this learning system, where parameters are defined by an inheritance based partial ordering, since setting a parameter in the hierarchy results in a non-monotonic refinement of the language hypothesis space.

The solution is to use a statistical method that tracks relative frequencies of parameter-setting-utterances in the input: we use the Bayesian Incremental Parameter Setting (BIPS) algorithm to set parameters of the universal grammar (Briscoe 2000). Such an approach sets the parameters to the values that are most likely given all the accumulated evidence. An advantage of using a Bayesian approach is that if the parameter's default value is strongly biased against the accumulated evidence then it will not be reset. Also, we no longer need to worry about indeterminacy of parameter-setting-utterances (Clark 1992) (Fodor 1998): the Bayesian approach allows us to solve this problem "for free" since indeterminacy just becomes another case of error due to misclassification of input data.

For this learning system the parameters of the universal grammar are all ternary valued (active, inactive, unset). To start with, all parameters may be unset (assigned probability 0.5) or set to an initial bias (for instance assigned probability 0.8, making the parameter active). Evidence from input utterances will either enforce the current parameter settings or negate them. Either way, there is re-estimation of the parametric probabilities. Parameters become set (considered active or inactive) as soon sufficient evidence has been accumulated, i.e. once the assigned probability reaches a threshold value. By employing this method, it becomes unlikely for parameters to switch between settings as the consequence of an erroneous utterance (unlike in the deterministic methods mentioned above).

In this work we investigated the mis-setting of word order parameters due to misclassification of thematic role. For illustration consider the following example utterance from the Sachs corpus: "he likes fish". A child listening to this utterance may be unsure who is doing the liking and who is being liked. Semantically, she could consider there to be two possibilities $\mathbf{LIKES}(\mathbf{he}, \mathbf{fish})$ or $\mathbf{LIKES}(\mathbf{fish}, \mathbf{he})$. In the first case we have the subject on the left of the verb and the object on the right; this is an SVO language like English. In the second case we have the reverse; an OVS language. An error occurs when an English (SVO) speaking child hypothesises the OVS interpretation of the utterance.

For this work, the order of constituents is recorded in the word order parameters; the subject direction in the *subject direction parameter* and the object direction in the *verb-argument direction parameter*.

For this experiment the learner was exposed to increasing amounts of misinterpreted thematic roles (from 0% up to 50% of all occurrences). This was achieved by randomly selecting the appropriate number of utterances and reversing the roles in their semantic representation.

Results

The learner received input from a real corpus of child-directed sentences. This corpus, the annotated Sachs corpus, simulates to some extent the environment to which a child is exposed. The corpus has been modelled as a categorial grammar by Villavicencio (2002) and we use her model as a reference against which to evaluate our acquired grammar. We also have an annotated version of the corpus that associates utterances with their correct semantic representation(s).

[3] See Buttery(2003b) or Villavicencio(2002) for a more detailed explanation of the parameters and parameter setting module.

Experiment 1a:

Input utterances were associated with semantic hypotheses sets rather than just the correct meaning. The extra hypotheses were chosen randomly and the correct semantic expression was always present in the set. Hypotheses sets of sizes 2, 3, 5, 10 and 20 were used. The learner was run several times for each size of set. Recall remained fairly constant regardless of the number of hypotheses. The precision also remained very high moving only as low as 93% for one of the runs with a hypothesis set of size 20.

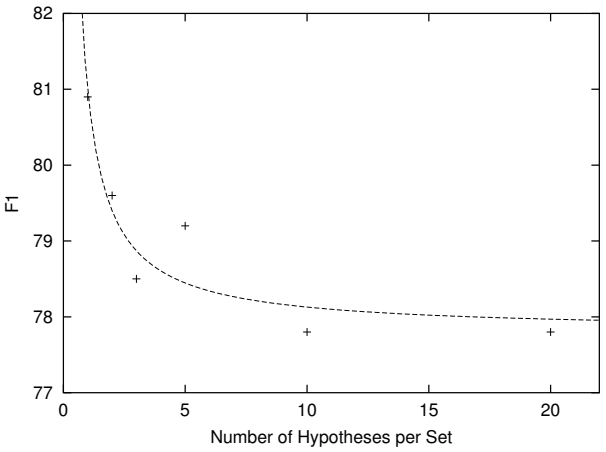

Figure 2: Size of Hypothesis Set vs. F_1

From Figure 2: as the number of hypotheses increases the F_1 of the learner decreases, tending towards a steady state. The results can be interpreted as follows:

The confidence values associated with the word-meaning pairs of extremely common words in the corpus (such as "kitty") become very high, very quickly. This means that many incorrect hypotheses in the hypothesis set can be ruled out (using the method of constraining hypotheses explained above). Once this starts to happen the problem rapidly reduces to that of one with a smaller starting hypothesis set. This accounts for the levelling off in F_1.

The precision remains high for all hypothesis set sizes due to the statistical nature of the confidence factor; since the surplus hypotheses (those extra to the correct hypothesis) were always chosen at random, the real meaning of the word eventually emerges as the most likely.

Experiment 1b:

The learner was run with some utterances being completely mismatched with semantic hypotheses (i.e. the correct hypothesis was not present amongst the set). This is analogous to the case where the child was not able to understand the meaning of the utterance from its observations. The results were found to be highly dependent on the utterances that were chosen to be mismatched. If many utterances were chosen that contained infrequently occurring words then the recall would plummet. There is a clear reason for this result. The distribution of words in the corpus is Zipfian. Most words appear very infrequently (over 250 words appear just once and more than 125 appear twice). In the original experiment (where only the correct hypothesis was paired with the meaning) 36% of words could be learnt with only one exposure. This capability is useless if a word that appears only once in the corpus is paired with an incorrect hypothesis. In such a situation the word will never be learnt. This highlights the obvious issue that statistical error handling methods are not much use when data is extremely sparse.

Experiment 2

The learner was run with increasing amounts of error due to misclassification of thematic role. Such an investigation has an obvious effect on the *subject direction parameter* and *verb-argument direction parameter*; with misclassification at 0% the parameters converges rapidly, while at 50% they struggle to converge at all.[4] It is more interesting, however, to consider the effect upon the other parameters since there is certainly interaction between them; mis-setting one parameter due to error, can lead the learner to mis-set others in a chain of consequence. Consider the example given earlier of an SVO language (like English) being confused with a SOV language with V2 (like German). If the *verb-argument direction parameter* gets set incorrectly due to a misclassification of thematic roles, then the parameter specifying V2 is likely to get incorrectly activated as a consequence. With regard to the learning system described here: the current settings of the Universal Grammar module are placing constraints on the legal set of syntactic categories that may be hypothesised by the Syntax Learning module; the Universal Grammar module is only activated when the Syntax Learning module has successfully allocated a legal syntactic category to every word in the current utterance; consequently the mis-setting of a parameter could have disastrous and self perpetuating effects.

Using a Bayesian approach to set parameters we are extremely unlikely to set a parameter in error and thus unlikely to encounter such problems. Indeed, results showed that, for all degrees of misclassification of thematic role, the overall performance of Universal Grammar module was constant (confirming previous results by Villavicencio 2002). Thematic role misclassification varied between 0% to 50% at 10% intervals (0%, 10%, 20%, 30%, 40%, 50%); for all degrees of misclassification an average of 9 word order parameters were set correctly resulting in 13.5 being correct according to the target (due to the inheritance of default values from super-type parameters).

Intuition suggests that the time taken for parameter convergence will increase with increasing numbers of error; a wider legal syntactic category set is hypothesised for longer. Villavicencio (2002) has confirmed this by showing that there is a difference of 45% in the speed of convergance between the error-free and the maximum thematic-role-error cases.

[4]Note that the *subject direction parameter* will eventually become set due to the presence of intransitive verbs.

Conclusions

Our results have shown that by using statistical error handling a learning system (learning from real data) may be robust to errors introduced by indeterminacy of meaning, and to errors introduced by indeterminacy in parameter settings.

In general, deterministic methods of learning seem to be unfeasible: A deterministic learning system that has been exposed to an error is doomed. The system must find a solution that incorporates that error as well as all positive evidence. When a statistical method is adopted errors cannot affect convergence to the target only the speed of that convergence.

The idea of statistical learning is further endorsed by an investigation undertaken by Yang (2002). He has provided evidence that a statistical approach (a distribution over a grammar space) must be adopted if the productions made by children during acquisition are to be explained. Furthermore, it has been suggested that some parameters may have a default initial value (Chomsky 1981) and that these defaults prevent the learner from converging upon a superset of her language (Lightfoot 1991). Using a Bayesian methodology (such as the BIPS algorithm used here) prior probabilities can be assigned to parameters, biasing the learner. This incorporates the idea of a default value while still allowing the learner to re-assess the parameter in accordance with the evidence; after all, if a parameter is initialised with a default that is unchangeable that parameter becomes part of the definition of the principles framework.

The key feature of statistical approaches to parametric learning models is that a parameter is only set once enough evidence is accumulated; when there is not enough evidence the parameter remains unset. Such a feature allows a learner to "keep her options open" on features that are undetermined in her language, perhaps due to a process of language change. For instance, if the English language really was SVO half the time and VSO the other (which would be approximately equivalent to 50% thermatic role misclassification in experiment 2) it would not be sensible for the learner to set the *subject direction parameter* at all; she would be unable to parse half the utterances she heard!

Our illustration of a simple learning problem (the number game) demonstrated that errors can be dealt with if the learner is allowed to choose the most likely hypothesis in preference to one that incorporates all the evidence seen so far; meaning that some statistical retention of the data is necessary if a learner is going to cope with errors. This is not to suggest that a child should consciously be counting events, perhaps rather that the brain has some capacity to store this information without any effort on behalf of the learner.

References

Bickerton, D. 1984. The language bioprogram hypothesis. *The Behavioral and Brain Sciences* 7(2):173–222.

Brown, R., and Hanlon, C. 1970. Derivational complexity and the order of acquisition of child speech. In Hayes, J., ed., *Cognition and the Development of Language*. New York: Wiley.

Buttery, P. 2003a. A computational model for first language acquisition. In *CLUK-6*.

Buttery, P. 2003b. Language acquisiton and the universal grammar. 119. Glasgow: AMLaP.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. 1981. *Lectures on Government and Binding*. Foris Publications.

Clark, R. 1992. The selection of syntactic knowledge. *Language Acquisition* 2(2):83–149.

Fodor, J. 1998. Unambiguous triggers. *Linguistic Inquiry* 29(1):1–36.

Gibson, E., and Wexler, K. 1994. Triggers. *Linguistic Inquiry* 25(3):407–454.

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

Kroch, A. 1989. Reflexes of grammar in patterns of language change. *Journal of Language Variation and Change* 1:199–244.

Lightfoot, D. 1979. *Principles of Diachronic Syntax*. Cambridge University Press.

Lightfoot, D. 1991. *How to set parameters: Arguments from language change*. Cambridge, MA: MIT Press.

Newport, E. 1999. Reduced input in the acquisition of signed languages: contributions to the study of creolization. In DeGraff, M., ed., *Language Creation and Language Change*. Cambridge, Ma: MIT Press. 161–178.

Pinker, S. 1994. *The Language Instinct: How the Mind Creates Language*. New York: Harper Collins.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Siskind, J. 1996. A computational study of cross situational techniques for learning word-to-meaning mappings. *Cognition* 61(1-2):39–91.

Snow, C. 1977. Mothers' speech research: From input to interaction. In Snow, C., and Ferguson, C., eds., *Talking to children: language input and acquisition*. Cambridge, MA: Cambridge University Press.

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

Villavicencio, A. 2002. *The acquisition of a unification-based generalised categorial grammar*. Ph.D. Dissertation, University of Cambridge. Thesis published as Technical Report UCAM-CL-TR-533.

Wexler, K. 1998. Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua* 106:23–79.

Yang, C. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press.