Probabilistic language acquisition: Theoretical, computational, and experimental analysis

Anne S. Hsu (ahsu@gatsby.ucl.ac.uk)
Division of Psychology and Language Sciences, 26 Bedford Way
London, WC1H 0AP

Nick Chater (n.chater@ucl.ac.uk)
Division of Psychology and Language Sciences, 26 Bedford Way
and Centre for Economic Learning and Social Evolution (ELSE)
London, WC1H 0AP

Abstract

There is much debate over the degree to which language learning is governed by innate language-specific biases, or acquired through cognition-general principles. Here we examine the probabilistic language acquisition hypothesis on three levels: We outline a theoretical result showing that probabilistic learning in the limit is possible for a very general class of languages. We then describe a practical computational framework, which can be used to quantify natural language learnability of a wide variety of linguistic constructions. Finally, we present an experiment which tests the learnability predictions for a variety of linguistic constructions, for which learnability has been much debated. We find that our results support the possibility that these linguistic constructions are acquired probabilistically from cognition-general principles.

Keywords: child language acquisition: Gold's theorem; poverty of the stimulus; probabilistic learning; simplicity principle; adult grammar judgments; natural language

Introduction

A central debate in cognitive science revolves around how children acquire their first language. A significant portion of this debate centers on how children learn complex linguistic structures, such as restrictions to general rules. An example restriction-rule can be seen in the contraction of '*going to*': '*I'm gonna leave'* is grammatical whereas '*I'm gonna the store'* is ungrammatical. Language communication requires the speaker to generalize from previously heard input. However, research shows children rarely receive feedback when they produce an over-general, ungrammatical sentence. Children also aren't explicitly told which generalizations are allowed and which are not (Bowerman, 1988). These observations evoke the question: how do children learn that certain overgeneralizations are ungrammatical without explicitly being told?

Traditionally, linguists have claimed that such learning is impossible without the aid of innate language-specific knowledge (Chomsky, 1975; Crain, 1991; Pinker, 1989; Theakston, 2004). However, recently, researchers have shown that statistical models are capable of learning restrictions to general rules from positive evidence only (Dowman, 2007; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Grünwald, 1994; Perfors, Regier, & Tenenbaum, 2006; Regier & Gahl, 2004).

Here we examine language acquisition from a probabilistic perspective on a theoretical, computational and experimental level. We first revisit Gold's theorem and show that language identification *is* possible from a probabilistic perspective. Next we mention a recently proposed, general framework which can quantify learnability of constructions in natural language. This flexible framework allows for predictions to be made concerning the natural language learnability of a wide variety of linguistic rules. Finally, we experimentally test the learnability predictions obtained from this framework by comparing these predictions with adult grammaticality judgments for a wide range of linguistic constructions.

Gold revisited: probabilistic language acquisition with a simplicity prior

Inherent in a simplicity-based approach to language acquisition is the trade-off between simpler vs. more complex grammars: Simpler, over-general grammars are easier to learn. However, because they are less accurate descriptions of actual language statistics, they result in inefficient encoding of language input, i.e. the language is represented using longer code lengths. More complex grammars (which enumerate linguistic restrictions) are more difficult to learn, but they better describe the language and result in a more efficient encoding of the language, i.e., language can be represented using shorter code lengths. Under simplicity models, language learning can be viewed in analogy to investments in energy-efficient, money-saving appliances. By investing in a more complicated grammar, e.g. one which contains a restriction on a construction, the language speaker obtains encoding savings every time the construction occurs. This is analogous to investing in an expensive but efficient appliance that saves money with each use. A linguistic restriction is learned when the relevant linguistic context occurs often enough that the accumulated savings makes the more complicated grammar worthwhile. Because complex grammars become worth while as linguistic constructions appear more often,

simplicity models are able to learn restrictions based on positive evidence alone (See Figure 1).

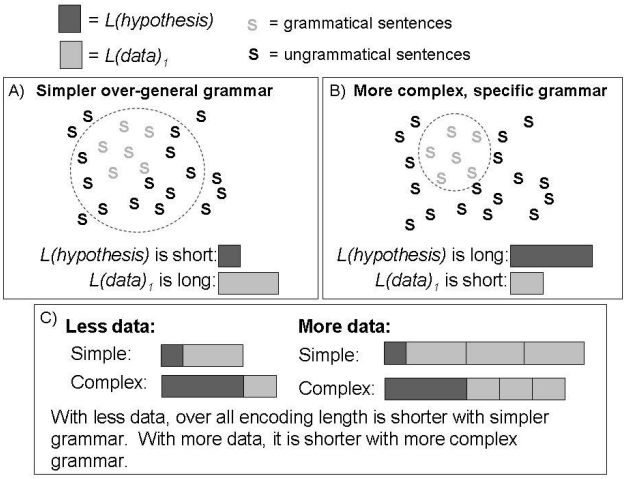

Figure 1: MDL simple grammar vs. efficient language encoding trade off. A) A simpler grammar is often over-general, i.e., allows for ungrammatical sentences as well as grammatical ones. Such an over-general grammar may be easy to describe (i.e., short grammar encoding length), but results in less efficient (longer) encoding of the language data. B) A more complex grammar may capture the language more accurately, i.e., allows only for grammatical sentences and doesn't allow for ungrammatical sentences. This more complex grammar may be more difficult to describe (i.e., longer grammar encoding length), but will provide a shorter encoding of language data. C) Initially, with limited language data, the shorter grammar yields a shorter coding length over-all, and is preferred under MDL. However, with more language input data, the savings accumulated from having a more efficient encoding of language data correctly favour the more complex grammar.

A central theoretical question is: given sufficient exposure to the language, can the learner recover a perfectly accurate description of that language? Gold (1967) famously showed that, under certain assumptions, this is not possible. However, a range of more positive results have since been derived, e.g., (J. A. Feldman et al 1969; Chater & Vitányi 2007). Here we show that under a simplicity-based probabilistic formulation, a new and strong positive result can be derived.

Suppose that the learner encounters sentences, s, which are independently sampled generated from a computable probability distribution, $C_P(s)$, which has Kolmogorov complexity $K(C_P)$. Here we will define learning a language as the process of identifying this distribution. $C_P(s)$ generates a corpus $S_n = s_1, s_2,...s_n,...$ which continues indefinitely. We assume that $C_P(s)$ allows all and only grammatical sentences in language L. That is, the probability of generating all sentences s, that are grammatical in L, is greater than zero, $C_P(s) > 0$; and conversely, if the probability of a sentence being generated is greater than zero, then it is grammatical according to L.

There is one additional mild constraint that we need to impose: that $C_P(s)$ has a finite entropy, i.e.,

$$H(C_p) \propto \sum_{j=1}^{\infty} C_p(s_j) \log\left(\frac{1}{C_p(s_j)}\right) < \infty$$

This is a modest constraint, because it follows from the assumption that the mean sentence length under distribution $C_P(s)$ is finite, which is clearly true for natural language.

The learning problem proceeds as follows: A learner is given an initial sample of the corpus S_n. The question then is: how should the learner assign probabilities to the various possible computable distributions C_Q that might have generated the corpus? This is equivalent to learning: $Pr(C_Q|S_n) \propto Pr(S_n|C_Q)Pr(C_Q)$

Also, we ask how these probabilities change as the corpus grows arbitrarily long, i.e., as n tends to infinity? In particular, can the learner identify the true probability distribution, C_P, in the limit?

Intriguingly, it turns out that this is possible – and indeed that an ideal learner (Chater & Vitányi 2007)) will 'converge' on the true probability distribution, C_P, with probability of measure 1, given a sufficiently large corpus. Suppose, for concreteness, that the learner "announces" its current most probable generating distribution each time a new sentence i arrives, based on the i sentences that he has received so far $S_i = \{s_1, s_2,..., s_i\}$. More formally, the following theorem holds: Consider any computable probability distribution C_P, from which samples, s_i, are drawn independently to generate a semi-infinite corpus S. Let m' be the number of initial items of S so that $S_{m'}$ is a "prefix" of S (i.e., a corpus consisting of the first m' items of s). With probability greater than $1-\varepsilon$, for any $\varepsilon > 0$, there is an m such that, under the simplicity principle, for all $m' \geq m$, the most probable C_Q, given $S_{m'}$ is the generating distribution C_P, i.e., $\text{argmax}(Pr(C_Q|S_{m'}))=C_P$.

Why is this true? A full proof is beyond the scope of this paper (see Chater & Hsu, in preparation); but the essence of the argument is the following. We know that almost all random samples from P will be incompressible (i.e., n sentences generated by the true generative model P will have no shorter description than the entropy $nH(P)$). This implies that, for *typical* data generated by P (which have summed probability arbitrarily close to 1), $K(P)+nH(P) \geq K(S_n) \geq nH(P)$. Now for each S_n, consider the set of probability distributions Q which satisfy this criterion: $K(Q)+nH(Q) \geq K(S_n) \geq nH(Q)$. For each n, there will be finitely many such Q; and, by our argument above, these will include the true distribution P. Now, for each n, the learner "announces" the simplest Q', i.e., the Q' such that for all Q, $K(Q) \geq K(Q')$. We know that P will always be in this set, by the argument above. However, there are only finitely many Q that are simpler than P. Once these simpler Q have been eliminated, then P will be the shortest element in the set, and will be announced indefinitely thereafter. We know that each of this finite set will be eliminated for

sufficiently large n, because the expected excess cost of encoding data generated by P with distribution Q is $n\mathrm{D}(Q||P)$, where $\mathrm{D}(Q||P) > 0$ unless $Q=P$; this excess cost tends to infinity as n tends to infinity. Hence, for some $n' > n$, after all probability distributions Q with shorter codes than P have been eliminated, P will be announced indefinitely.

Practical framework for quantifying learnability

The positive learnability results indicate that the probabilistic approach can be practically applied to the problem of language acquisition. Recently, researchers have used probabilistic models to show that many complex linguistic rules can be acquired by directly learning the probability distribution of grammatical sentence structures in language. These models learn this probability distribution under a cognition general prior for simplicity (Dowman, 2007; Foraker et al., 2009; Grünwald, 1994; Perfors et al., 2006; Regier & Gahl, 2004). Many of these studies used restricted language sets. In the context of natural language, a few studies have addressed specific linguistic cases such as anaphoric one (Foraker et al., 2009) and hierarchical phrase structure (Perfors et al., 2006).

Recently, a *general quantitative framework* has been proposed which can be used to assess the learnability of any given *specific linguistic restriction* in the context of real language, using positive evidence and language statistics alone (Hsu & Chater, 2010). This framework built upon previous probabilistic modeling approaches to develop a method that is generally applicable to any given construction in natural language. This new tool can be used to explicitly explore the learnability in a corpus relative to well-known information theoretic principles given a grammatical description. When using this framework to analyze learnability of a linguistic construction, there are two main assumptions: 1) The description of the grammatical rule for the construction to be learned. 2) The choice of corpus which approximates the learner's input. Given these two assumptions, the framework provides a method for evaluating whether a construction is present with adequate frequency to make it learnable from language statistics. The framework allows for comparison of different learnability results which arise from varying these two main assumptions. By making these assumptions explicit, a common forum is provided for quantifying and discussing language learnability.

Minimum Description Length hypothesis

Because this framework is detailed elsewhere (Hsu & Chater 2010), we will only provide a brief overview here. Learnability evaluations under a simplicity prior can be instantiated through the principle of minimum description length (MDL). MDL is a computational tool that can be used to quantify the information available in the input to an idealized statistical learner of language as well as of general cognitive domains (Jacob Feldman, 2000). When MDL is applied to language, grammars can be represented as a set of rules, such as that of a probabilistic context free grammar (PCFG) (Grünwald, 1994). An information-theoretic cost can then be assigned to encoding the grammar rules as well as to encoding the language under those rules.

Hsu & Chater (2010) used an instantiation known as 2-part MDL, which we will refer to as just MDL for brevity. In the context of language acquisition, the first part of MDL uses probabilistic grammatical rules to define a probability distribution over linguistic constructions, which combine to form sentences. Note that these probabilities are not necessarily the real probabilities of sentences in language, but the probabilities as specified under the current hypothesized grammar. The second part of MDL consists of the encoded representation of all the sentences that a child has heard so far. MDL selects the grammar that minimizes the *total* encoding length (measured in bits) of both the grammatical description and the encoded language length[1].

According to information theory, the most efficient encoding occurs when each data element is assigned a code of length equal to the smallest integer greater than or equal to $-\log_2(p_n)$ bits, where p_n is the probability of the nth element in the data. For our purposes, these elements are different grammar rules. The probabilities of these grammar rules are defined by the grammatical description in the first part of MDL. Because efficient encoding results from knowing the correct probabilities of occurrence, the more accurately the probabilities defined in the grammar match the actual probabilities in language, the more efficient this grammar will be.

Under MDL, the grammatical description is updated to be the most efficient one each time more data input is obtained. Savings occur because certain grammatical descriptions result in a more efficient (shorter) encoding of the language data. In general, more complex (i.e., more expensive) grammatical descriptions allow for more efficient encoding of the language data. Because savings accumulate as constructions appear more often, more complex grammars are learned (i.e., become worth investing in) when constructions occur often enough to accumulate a sufficient amount of savings. If there is little language data (i.e., a person has not been exposed to much language) a more efficient encoding of the language does not produce a big increase in savings. Thus, when there is less language data, it is better to make a cheaper investment in a simpler grammar as there is not as much savings to be made. When there is more language data, investment in a more costly, complicated grammar becomes worthwhile. This characteristic of MDL learning can explain the early overgeneralizations followed by retreat to the correct

[1] The MDL framework can also be expressed as a corresponding Bayesian model with a particular prior (Chater, 1996; MacKay, 2003; Vitányi & Li, 2000). Here, code length of the model (i.e., grammar) and code length of data under the model (i.e., the encoded language) in MDL correspond to prior probabilities and likelihood terms respectively in the Bayesian framework.

Table 1: Grammatical and ungrammatical sentences used in experiment.

Construction	*Grammatical usage*	*Ungrammatical usage*
is	She's as tall as he is.	She is as tall as he's.
arrive	The train arrived.	He arrived the train.
come	The train came.	I came the train.
donate	He donated some money to the charity.	He donated the charity some money.
fall	The ornament fell.	He fell the ornament.
disappear	The rabbit disappeared.	He disappeared the rabbit.
what is	What's it for?	What's it?
shout	I shouted the news to her.	I shouted her the news.
pour	I poured the pebbles into the tank.	I poured the tank with pebbles.
vanish	The rabbit vanished.	He vanished the rabbit.
whisper	I whispered the secret to her.	I whispered her the secret.
create	I created a sculpture for her.	I created her a sculpture.
who is	Who's it for?	Who's it?
going to	I'm gonna faint.	I'm gonna the store.
suggest	I suggested the idea to her.	I suggested her the idea.
that	Who do you think that she called?	Who do you think that called her?
want to	Which team do you wanna beat?	Which team do you wanna win?

grammar that has been observed in children's speech (Bowerman, 1988). The output of the framework described in Hsu & Chater (2010) results in an estimated number of occurrences needed for a specific linguistic rule to be learned and corpus analysis is then used to assess how many years on average are needed for the sufficient number of occurrences. The general applicability of this framework and its ability to produce clear learnability predictions allow us to take the crucial next step in addressing the language acquisition problem: experimentally assessing whether language might actually be probabilistically acquired.

Testing learnability predictions

Hsu & Chater (2010) used the above framework to assess language learnability of constructions, whose learnability have been commonly debated. These all involve restrictions on a general linguistic rule, which was described using PCFG's. Predictions for learnability in terms of years needed was made for constructions whose learnability have been commonly debated in the language acquisition field. These included restrictions on the following 17 constructions[2]: contractions of *want to, going to, is, what is and who is*; the optionality of *that* reduction; dative alternation for the verbs *donate, whisper, shout, suggest, create, pour*; transitivity for the verbs, *disappear, vanish, arrive, come, fall.* See Hsu & Chater (2010) for the explicit grammar descriptions of linguistic rules to be learned. The results showed a large spread in learnability. Some constructions appeared readily learnable within just a few years whereas other constructions required years that far outnumbered human life spans. Hsu & Chater (2010) compared predicted MDL learnablity with child grammar judgments of constructions for which there was data collected from previous experimental work (Ambridge, Pine, Rowland, & Young, 2008; Theakston, 2004). It was found that child grammar judgments for the constructions were more correlated with learnability than frequency counts (the entrenchment hypothesis (Theakston, 2004)).

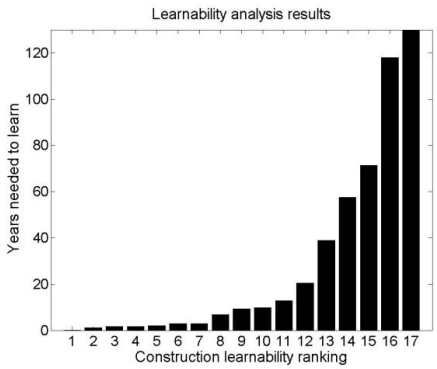

Figure 2: Estimated years required to learn construction. The constructions are sorted according to learnability: 1) *is* 2) *arrive* 3) *come* 4) *donate* 5) *fall* 6) *disappear* 7) *what is* 8) *shout* 9) *pour* 10) *vanish* 11) *whisper* 12) *create* 13) *who is* 14) *going to* 15) *suggest* 16) *that* 17) **want to.* *Predicted years for learning *want to* is 3,800years.

Here we propose that construction learnability should also correlate with adult grammaticality judgments: The more difficult a construction is to learn, the greater the difference

[2] Hsu & Chater (2010) also included analysis of two more linguistic rules concerning the necessary transitivity of the verbs *hit* and *strike*. Though these verbs are traditionally known to be transitive, in colloquial speech they have evolved to have a ambitransitive usage: e.g. *The storm hit. Lightening struck.* In COCA there are 3678 and 1961 intransitive occurrences of *hit* and *strike* respectively. Thus we did not assess rules regarding the intransitivity of these verbs in our experiment.

should be between judgments of the ungrammatical vs. grammatical uses of the construction.

Model Predictions

We conducted our learnability analysis using the full Corpus of Contemporary American English (COCA), which contains 385 million words (90% written, 10% spoken). We believe this is a reasonable representation of the distributional language information that native English language speakers receive. Learnability results using the British National Corpus were similar to that from COCA (Hsu & Chater, 2010). Figure 2 shows the estimated number years required to learn the 17 constructions. We quantified learnability as $log(1/N_{years})$, where N_{years} was the number of estimated years needed to learn a construction (Hsu & Chater, 2010).

Learnability vs. entrenchment To verify that our experimental results are not also trivially explained by a simpler hypothesis, we will also compare experimental results with the predictions of entrenchment theory. Entrenchment is the hypothesis that the likelihood of a child over-generalizing a construction is related to the construction's input occurrence frequency. There is some relation between learnability and entrenchment predictions because high construction occurrence frequencies do aid learnability. However, learnability differs from mere frequency counts because MDL also takes into account the complexity of the grammatical rule that governs the construction to be learned. Additionally, learnability is influenced by whether the restricted form would be commonly or uncommonly expected, if it were grammatically allowed. Here, we propose that under entrenchment hypothesis, the relative grammar judgment difference should be related to the construction's input occurrence frequency. (Frequencies estimated from COCA).

Experimental method

Participants 105 participants were recruited for an online grammar judgment study (age range: 16-75 years, mean=34 years). Results were included in the analysis only for participants who answered that they were native English speakers (97 out of 105 participants). The majority (74%) of our participants learned English in the United States. Other countries included the UK (14%), Canada (5%), Australia (4%). The rest learned English in either Ireland or New Zealand.

Procedure Participants were asked to rate the grammaticality of grammatical and ungrammatical sentences using the 17 constructions whose learnability were quantified above. These sentences (34 total) are shown in Table 1. Grammar judgments ranged from 1-5: 1) Sounds completely fine (Definitely grammatical) 2) Probably grammatical (Sounds mostly fine) 3) Sounds barely passable (Neutral) 4) Sounds kind of odd (probably ungrammatical) 5) Sounds extremely odd (Definitely ungrammatical).

Results

Results show a strong correlation between averaged relative grammaticality vs. log learnability as predicted by MDL, r=.35; p=.0045 (see Figure 3). Relative grammaticality for a given linguistic construction is the grammatical rating for the ungrammatical sentence subtracted by the rating for the grammatical sentence. Note that 4 is the maximum possible relative grammaticality because the lowest ungrammatical rating is 5 and the highest grammatical rating is 1. In contrast, there is no correlation between relative grammaticality and construction occurrence frequency, as would be predicted by entrenchment (see Figure 4).

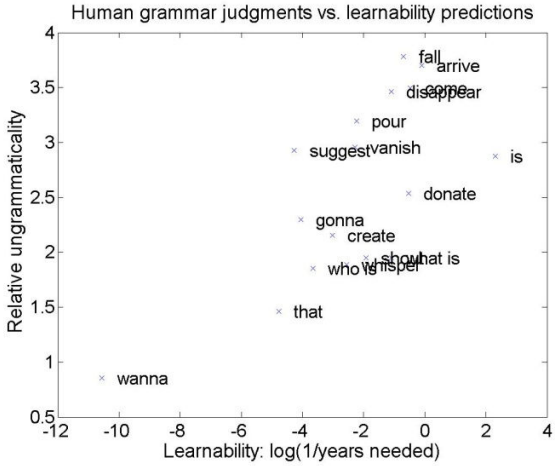

Figure 3: Human grammar judgments vs. learnability analysis. Learnability is log of the inverse of the number of estimated years needed to learn the construction. Correlation values: r=.35; p=.0045

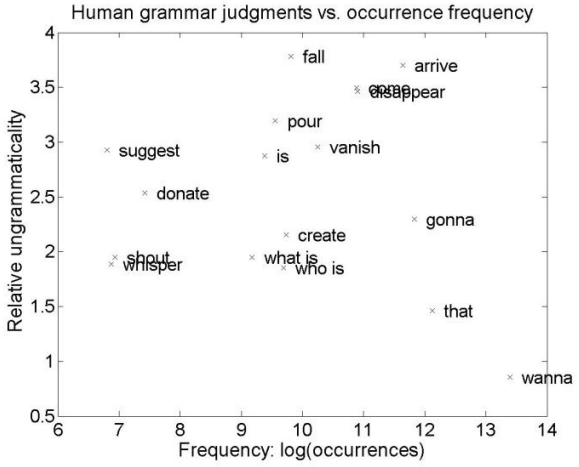

Figure 4: Human grammar judgments vs. log of occurrence frequency. Frequencies were estimated using Corpus of Contemporary American English.

Summary and Conclusions

This presented work helps evaluate how much of first language is probabilistically acquired from exposure. We show that, despite Gold's theorem, language is identifiable with a cognition general prior of simplicity under fairly general assumptions. We then describe a recently formulated framework which allows probabilistic learnability to be quantified in the context of natural language. This framework makes concrete predictions in terms of years needed to learn particular linguistic rules, given an assumed formulation of the rules to be learned and the corpus which represents a learner's language input.

There has now been a substantial body of work showing that probabilistic language learning *is theoretically and computationally possible.* The important next step in research on language acquisition is to assess whether probabilistic learning actually occurs in practice. Here we make the supposition that if language is probabilistically acquired, then there should be evidence of this in adult grammar judgments. There is a subtle leap of logic in this supposition. MDL learnability assumes that a grammar is learned in an absolute sense: once a grammar is chosen under MDL, that is the one used and there is no gradation of knowledge. However, here we are conjecturing that learnability should not only correlate with how long it takes for linguistic rule to be acquired, but also with how certain is one's knowledge of that rule. The more certain one is of a grammatical rule, the greater the difference should be one's acceptability rating of the ungrammatical form relative to the grammatical form. Experimental results show that predicted learnability correlates well with relative grammar judgments for the 17 constructions analyzed, chosen as controversial cases from the literature. Our experimental results support the possibility that many linguistic constructions that have been argued to be innately acquired may instead be acquired by probabilistic learning.

Our learnability predictions were calculated using a large corpus (COCA) to represent the distributional language input that native English speakers receive. This assumes that the distributional information estimated from this corpus is representative of that which influenced the language acquisition process in our adult participants. It also allows for the possibility that a speaker's certainty about different linguistic rules is updated through adulthood using probabilistic learning. If so, older adults might more certain in their grammar judgments, is a direction for future work.

References

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition, 106,* 87-129.

Bowerman, M. (1988). The 'No Negative Evidence' Problem: How do Children avoid constructing an overly general grammar? In J.Hawkins (Ed.), *Explaining Language Universals* (pp. 73-101). Oxford: Blackwell.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103,* 566-581.

Chater, N. & Hsu, A. (in preparation). Language learning in the limit: theory and practice.

Chater, N. & Vitányi, P.M.B. (2007). `Ideal learning' of natural language: Positive results about learning from positive evidence, *Journal of Mathematical Psychology*, *51*,135-163.

Chomsky, N. (1975/1955). *The Logical Structure of Linguistic Theory*. London: Plenum Press.

Crain, S. (1991). Language Acquisition in the Absence of Experience. *Behavioral and Brain Sciences, 14,* 597-612.

Dowman, M. (2007). Minimum Description Length as a Solution to the Problem of Generalization in Syntactic Theory. *Machine Learning and Language, (in review)*.

Feldman, Jacob (2000). Minimization of boolean complexity in human concept learning. *Nature, 403,* 630-633.

Feldman, J.A., Gips, J., Horning, J. J., & Reder, S. (1969) Grammatical complexity and inference. Technical Report CS 125, Stanford University.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2009). Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One. *Cognitive Science, 33,* 300.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10, 447-474.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In S.Scheler, Wernter, & E. Rilof (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language.* (pp. 203-216). Berlin: Springer Verlag.

Hsu, A. & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science, 2nd revision submitted*.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Perfors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the Stimulus? A rational approach. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society,* 663-668.

Pinker, S. (1989). *Learnability and Cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Regier, T. & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition, 93,* 147-155.

Theakston, A. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development, 19,* 15-34.

Vitányi, P. & Li, M. (2000). Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity. *IEEE Transactions on Information Theory, IT, 46,* 446-464.