Running Head: Simplicity-based approach to language learning

Language learning from positive evidence, reconsidered: A simplicity-based approach

Anne S. Hsu

Department of Cognitive, Perceptual and Brain Sciences

University College London

26 Bedford Way

London, WC1H 0AP, UK

anne.hsu@ ucl.ac.uk

Nick Chater

Behavioural Science Group

Warwick Business School

University of Warwick

Coventry, CV4 7AL, UK

Nick.Chater@wbs.ac.uk

Paul Vitányi

Centrum Wiskunde & Informatica,

Science Park 123, 1098 XG Amsterdam,

The Netherlands

paul.vitanyi@cwi.nl

Keywords: child language acquisition; no negative evidence; probabilistic Bayesian models; minimum description length; simplicity principle; natural language; induction; learning

Abstract

Children appear to be able to learn their native language by exposure to their linguistic and communicative environment, but without requiring that their mistakes are corrected. Such learning from "positive evidence" has been viewed as raising "logical" problems for language acquisition: in particular, without correction, how is the child to recover from conjecturing an over-general grammar, which will be consistent with any sentence that the child hears? There have been many proposals concerning how this "logical problem" can be dissolved. Here we review recent formal results showing that the learner can learn successfully from positive evidenced by favouring the grammar that provides the *simplest* encoding of the linguistic input. Results include the learnability of linguistic prediction, grammaticality judgments, and form-meaning mappings. The simplicity approach can also be "scaled-down" to analyse the learnability of specific linguistic constructions, and is amenable to empirical test as a framework for describing human language acquisition.

Children appear to learn language primarily by exposure to the language of others. But how is this possible? The computational challenges of inferring the structure of language from mere exposure are formidable. In light of this, many theorists have conjectured that language acquisition is only possible because the child possesses cognitive machinery that fits especially closely with the structure of natural language. This could be because the brain has adapted to language (Pinker & Bloom, 1990), or because language has been shaped by the brain Christiansen & Chater, 2007).

A number of informal arguments concerning the challenge of language learning from experience have been influential. Chomsky (1980) argued that the "poverty of the stimulus" available to the child was sufficiently great that the acquisition of language should be viewed as analogous to the growth of an organ, such as the lung or the heart, unfolding along channels pre-specified in the genome. Here, we focus on a specific facet of poverty of the stimulus: that children do not appear to receive or attend to "negative evidence:" explicit feedback that certain utterances are ungrammatical (Bowerman, 1988; Brown & Hanlon, 1970; Marcus, 1993)

The ability to learn the language in the absence of negative evidence is especially puzzling, given that linguistic rules are riddled with apparently capricious restrictions. For example, a child might naturally conclude from experience that there is a general rule that *is* can be contracted, as in *He's taller than she is*. But contractions are not always allowed, as in **He is taller than she's*. The puzzle is that, once the learner has entertained the possibility that the overgeneral rule is correct, it appears to have no way to "recover" from overgeneralization and recognise that restrictions should be added. This is because each contraction that it hears

conforms to the overgeneral rule. If the learner uses the overgeneral rule, then it will from time to time, produce utterances such as **John isn't coming but Mary's*. A listener's startled reaction or look of incomprehension might provide a crucial clue that the rule is overgeneral: however, this feedback is the very negative evidence that appears to be inessential to child language acquisition. Thus, if children do not use such negative evidence, how can they recover from such overgeneralisations? Various scholars argue that they cannot: Restrictions on overgeneral grammatical rules must, instead, be innately specified (e.g., Crain & Lillo-Martin, 1999). Other theorists argue that avoiding overgeneral rules poses a fundamental "logical problem" for language acquisition (Baker & McCarthy, 1981; Dresher & Hornstein, 1976).

One way to defuse the puzzle is to challenge its premise. One possibility is that, despite appearances, children can access and use negative evidence in a subtle form. In this paper, we set aside these contentious issues (e.g., Demetras, Post & Snow, 1986; Marcus, 1993) and argue that, whether or not negative evidence is available to or used by the child, language can successfully be learned without it (following, for example, MacWhinney, 1993; 2004; Rohde & Plaut, 1999; Tomasello, 2004).

One reason for supposing that a learner can eliminate overgeneral grammars from positive evidence alone is provided by recent developments in computing power and machine learning techniques. These developments have led to the development of language engineering systems which can automatically learn non-trivial aspects of phonology, morphology, syntax and even aspects of semantics from positive language input (Goldsmith, 2001; Steyvers, Griffiths & Tenenbaum, 2006; Klein & Manning, 2005). While such systems are still very far from being able to acquire language from mere exposure, the pace of progress suggests that *a priori* barriers to learning may not be insurmountable as once supposed.

In the context of this special issue, it is natural to ask**:** how far can the analysis of formal learning results help us understand, resolve, or at least clarify, the nature of the learning problem a child faces? In particular, is it possible to show formally that, given sufficient positive evidence, language acquisition is possible?

There has been a large literature, dating back at least to Gold (1967) of relevant formal results (e.g., Angluin, 1980, 1988; Clark, & Eyraud, 2007; Jerome Feldman, 1972; Horning, 1969; Jain, Osherson, Royer & Kumar Sharma, 1999; Niyogi, 2006; Wharton, 1974). Rather than reviewing these, and indicating how they may be extended, here we will take a more direct approach. First, we outline our general approach based on a 'simplicity principle' (Section 1) embodied in an "ideal learner" (Section 2). We then outline some recent interconnected formal results on prediction, grammaticality judgments, language production, and mapping between form and meaning (Sections 3-6). We then describe how the simplicity-based approach to learning can be "scaled-down" to provide a practical methodology to analyse the learnability of specific linguistic patterns and briefly outlining how this approach can be linked with experimental data (Section 7). Overall, the contribution of the work reviewed here is to prove that, under certain fairly mild conditions, language acquisition from sufficient amounts of positive evidence is possible; and to indicate how the simplicity-based approach can provide a framework for understanding child language acquisition.

1. Ideal learning using a simplicity principle

Suppose that a learner, whether human or artificial, is faced with a set of positive data, which might, for example, be perceptual or linguistic. Any set of data is consistent with an infinite number of hypotheses: in the case of language, any given set of observed sentences might have

been generated by any of an infinite number of grammars. How can an intelligent agent choose among these infinite possibilities? The simplicity principle provides an attractive answer: it recommends that the learner prefers hypotheses which provide the *simplest* encoding of the data. The simplicity principle has a long history in the philosophy of science and the study of perception (e.g., Mach, 1959), and has been proposed as a general cognitive principle (Chater & Vitányi, 2002).

The simplicity of the code is measured by its length (by convention, in a binary code). Thus, to specify a set of sentences in terms of the grammar requires (i) encoding the grammar, presumably as a list of rules, and, perhaps also a list of probabilities associated with each rule; and (ii) specifying the data in terms of the grammar. If the grammar is probabilistic, this can be done straightforwardly. According to information theory (Shannon & Weaver, 1949), the shortest code length for an outcome with probability $\Pr(x)$ is $\log_2(1/\Pr(x))$; this implies that probable outcomes have short codes, and improbable outcomes have longer codes. The complexity of a hypothesis is therefore the sum of two factors: the complexity of the hypothesis itself, and the degree to which the hypothesis predicts the data. Thus, for example, a wildly overgeneral hypothesis, e.g. a hypothesis that any sentence at all as possible, will be very briefly encoded; but it will fail to predict the data, and the encoding of the data will be extremely complex. Conversely, very complex hypotheses may be ruled out because the first term is too great, even though the hypotheses may precisely predict the data. By choosing the hypothesis with the shortest overall description length for the data, the simplicity principle seeks to find an appropriate middle ground between these two extremes. In Section 7, we shall briefly discuss how this may be done in practice (see Hsu, Chater & Vitányi, 2011).

By searching for the middle ground between over-general and over-specific grammars, the simplicity principle appears to have the potential to learn successfully from positive data. Crucially, an overgeneral grammar will predict the linguistic data poorly---and hence will provide an unnecessarily long code for that linguistic data. The more data available to the learner, the greater the code length "wasted" in comparison with some more restrictive grammar. Hence, a more restrictive grammar will provide a simpler encoding of the linguistic data overall, if sufficient data is available: and hence the overgeneral grammar will be rejected.

This intuition is encouraging but hardly definitive. Knowing that a learner can potentially eliminate a single over-general grammar does not, of course, indicate that it can successfully choose between infinte possible grammars (an infinite number of which will be over-general), and hone in on the "true" grammar, or some approximation to it. We shall see, however, that positive results along these lines are possible.

2 An"ideal" learner

In the results below, we consider what an "ideal" learner, using the simplicity principle, can learn purely from exposure to an (indefinitely long) sequence of linguistic input (i.e., from positive evidence). These results can be obtained based only on the modest assumption that linguistic input is generated by a computable probabilistic process. Informally, we can view this process as embodying a Turing machine (or any other computer) combined with a source of randomness (i.e., a sequence of coin flips). The source of randomness captures the possibility that the process of generating the linguistic input may be non-deterministic (although it need not be); the restriction to computable probability distributions requires that the *structure* in the linguistic input is computable. This restriction is mild because cognitive

science takes computability constraints on mental processes as founding assumptions (Pylyshyn, 1984) and, indeed, specific models of language structure and generation all adhere to this assumption. Finally, for mathematical convenience, and without loss of generality, we assume that the linguistic input is coded in binary form.

Note, in particular, that we allow that there can be any (computable) relationship between different parts of the input---for example, we do not assume that sentences are independently sampled from a specific probability distribution. Our very mild assumption allows sentences to be highly interdependent (this is one generalization with respect to earlier results, e.g., Jerome Feldman, 1972; Wharton, 1974; Angluin, 1980), and allow the possibility that the language may be modified or switched during the input or that sentences from many different languages might be interleaved. Chater and Vitányi (2007) showed that, nonetheless, an ideal learner can learn from positive evidence alone.

Specifically, suppose that the (indefinitely long) linguistic input, coded as a binary sequence, $x\ldots$, is generated by a computable probability distribution, $\mu_C(x\ldots)$. Intuitively, we can view this as meaning that there is a computer program, C, which receives random input, $y\ldots$, with initial segment y, from an indefinitely long stream of coin flips. The *output* of C, given this random input, is a sample from $\mu_C(x\ldots)$. (Strictly, μ is a measure, rather than a probability distribution, as the sequence is infinite; indeed, it is actually a semi-measure. We ignore these technicalities here (see Chater & Vitányi, 2007; and Li & Vitányi, 2008)).

How likely is it that computer program, C, will produce a sequence beginning with $x\ldots$? There may be many sequences of coin flips, $y\ldots$, with initial sequence y, of length $l(y)$, that, when fed in to program C, will produce the same output x. This probability of a given y is equal to $2^{-l(y)}$, which is the probability of generating any specific binary sequence of length $l(y)$ from

unbiased coin flips. The probability of an output with initial segment x, is $\mu_C(x)$; and this is the sum of the probabilities of all the inputs which begin with $y\ldots$ and which generate output sequences that begin with x:

$$\mu_C(x) = \sum_{y:C(y\ldots)=x\ldots} 2^{-l(y)} \qquad (1)$$

In essence, the distribution $\mu_C(x)$ is built on a simplicity principle: outputs which correspond to short programs for the computer, C, are overwhelmingly more probable than outputs for which there are no short programs.

The learner's task, then, can be viewed as approximating $\mu_C(x)$. So, for example, if C generated independent samples from a specific stochastic phrase structure grammar, then the learner's aim is to find a probability distribution which matches $\mu_C(x)$ as accurately as possible. To the extent that this is possible, the learner will be able to predict how the corpus will continue; decide which strings are allowed by $\mu_C(x)$ and generate output similar to that generated by $\mu_C(x)$. Framing these points in terms of language acquisition, this means that, by approximating $\mu_C(x)$, the learner can, to some approximation, (i) predict what phoneme, word, or sentence will come next (insofar as this is predictable at all); (ii) learn to judge grammaticality; and (iii) learn to produce language, indistinguishable from that to which it has been exposed. We shall briefly explore these issues in turn in Sections 3-5.

How, then, can the learner approximate $\mu_C(x)$, given that it has exposure to just one (admittedly infinite) corpus $x\ldots$; and no prior knowledge of the specific computational process, C, which has generated this corpus? Our starting point is the assumption that the learner does, at least, have access to a *universal* programming language. From computability theory, it is well known that any programming language with a minimum of sophistication can implement any other. Such universal programming languages include practical artificial languages, from

Fortran, to C++ to JAVA. It follows, of course, that any computational process, *C*, can be implemented in any universal language: if a process has a program in any universal language, then an equivalent program can be written in any other language, for example, by implementing the first language in terms of the second, and then executing the original program. Let us suppose, reasonably, that the brain (and our ideal learner) has the minimal resources required to implement a universal language.

A simplicity-based learner, such as we explore here, will favour simple "explanations" of the corpus x…; that is, it will assign prior probability to each program which generate the corpus *x*, based on its length (for concreteness, where the program is expressed in a standardized binary code). But surely the length of a program depends on the programming language used? What may be easy to write in Matlab may be difficult to write in Prolog; or vice versa. It turns out, though, that the choice of programming language affects program lengths only to a limited degree. An important result, known as the invariance theorem (Li & Vitányi, 2008), states that, for any two universal programming languages, the length of the shortest program for any computable object in each language is bounded by a fixed constant.

The invariance theorem thus allows the code length of computer programs to be specified without committing to a particular computer language. So, by assuming that the coding language that the cognitive system uses is universal, we can avoid having to provide a specific account of the program that the learner uses.

Now suppose the learner assumes only that the corpus, *x*…, is generated by a computable process (and hence makes no assumptions that is generated by specific type of grammar, or indeed, any grammar at all; and makes no assumption that "sentences" are sampled independently, etc.). Then, the probability of each possible *x*… is given by the

probability that this sequence will be generating from the output of a random input, *y*…, of length *l*(*y*) (as before, random coin flips) fed to a universal computer, *U* (technically it is important that this computer is monotone (Chater & Vitányi, 2007), but we ignore this here). Analogous to (1), we can define this "universal monotone distribution" (Solomonoff, 1978) $\lambda(x)$:

$$\lambda(x) = \sum_{y:U(y...)=x} 2^{-l(y)} \qquad (2)$$

where *U(y)* are programs *y* written in the universal programming language. Thus, an ideal learner can use a universal programming language and the simplicity principle to formulate $\lambda(x)$, which serves as the learner's approximation for $\mu_C(x)$.

Note that $\lambda(x)$ is known to be uncomputable (Li & Vitányi, 2008), and hence must be approximated. It remains an open question the degree to which $\lambda(x)$ can be approximated and how this affects learnability results. Promisingly, computable approximations to the universal distribution can be developed into practical methodologies in statistics and machine learning (e.g., Rissanen, 1987; Wallace & Freeman, 1987). Such approximations will be considered briefly below in relation to developing a methodology for assessing the learnability of specific linguistic patterns.

Perhaps surprisingly, it turns out that $\lambda(x)$ serves as a good enough approximation to $\mu_C(x)$ to allow the ideal learner to predict future linguistic input. We show below that this allows the ideal learner to make grammaticality judgments, produce grammatical utterances, and map sound to meaning.

3. Prediction

One indication of understanding, in any domain, is the ability to predict. Thus, if the linguistic input is governed by grammatical or other principles of whatever complexity, any learner that can predict how linguistic material will continue, arbitrarily well, must, in some sense, have learned such regularities. Prediction has been used as a measure of how far the structure of a language has been learned since Shannon (1951); and is widely used as a measure of learning in connectionist models of language processing (Christiansen & Chater, 1994; Christiansen & Chater, 1999; Elman, 1990) .

In the present setting, the task of prediction can be formulated as follows. At each point in a binary sequence *x*, specify the probability that the next symbol is 0 or 1. Given that the data is generated by $\mu_C(x)$ the true probabilities are:

$$\mu_C(0 \mid x) = \frac{\mu_C(x0)}{\mu_C(x)} \; ; \qquad \mu_C(1 \mid x) = \frac{\mu_C(x1)}{\mu_C(x)} \tag{3}$$

where $\mu_C(0 \mid x)$ and $\mu_C(1 \mid x)$ represents the probability that the subsequence x is followed by a 0 and 1 respectively. But the ideal learner does not have access to $\mu_C(x)$, and instead will use $\lambda(x)$ for prediction. Thus, the learner's predictions for the next item of a binary sequence that has started with *x* is:

$$\lambda(0 \mid x) = \frac{\lambda(x0)}{\lambda(x)}; \qquad \lambda(1 \mid x) = \frac{\lambda(x1)}{\lambda(x)} \tag{4}$$

A key result by Solomonoff (1978), which we call the *Prediction Theorem*, shows that, in a specific rigorous sense, the universal monotone distribution λ, described above, is reliable for predicting any computable monotone distribution, μ, with very little expected error. More specifically, the difference in these predictions is measured by the square of difference in the probabilities that μ and λ assign to 0 being the next symbol:

$$\text{Error}(x) = \left(\lambda(0 \mid x) - \mu(0 \mid x)\right)^2 \tag{5}$$

And the *expected* sum-squared error for the *n*th item in the sequence is:

$$s_n = \sum_{x:l(x)=n-1} \mu(x)\text{Error}\,(x) \tag{6}$$

The better λ predicts μ, the smaller s_n will be. Given this, the overall expected predictive success of the method across the entire sequence is obtained by summing the s_n across all *n*:

$$\sum_{n=1}^{\infty} s_n \tag{7}$$

Solomonoff's Prediction Theorem shows that predictions from the ideal learner's λ approximate any computable distribution, μ, so that $\sum_{n=1}^{\infty} s_n$ is bounded by a constant. Thus, as the amount of data increases, the expected prediction error goes to 0. Specifically, the following result holds:

Prediction Theorem (Solomonoff, 1978): Let μ be a computable monotone distribution, predicted by a universal distribution λ (see Li & Vitányi, 2008) for mathematical discussion, and an accessible proof). Then,

$$\sum_{j=n}^{\infty} s_n \leq \frac{\log_e 2}{2} K(\mu) \qquad (8)$$

where $K(\mu)$ is the length of the shortest program on the universal machine that implements μ, known as its Kolmogorov complexity (see Chater and Vitányi, (2007), for further details).

The Prediction Theorem shows that learning by simplicity can, in principle, be expected to converge to the correct conditional probabilities for predicting subsequent linguistic material. This implies that the learner is able to learn the structure of the language---because if not, the learner will not know which sentences are likely to be said, and hence will make prediction errors. This results suggest that, given sufficient positive evidence, linguistic restrictions, such as on the allowed contraction of *is* mentioned above, are learnable from positive evidence. This is because a learner who does not learn these restrictions will continue to predict the ungrammatical form when it is not allowed, and thus accrue an infinite number of prediction errors. Note that while the Prediction Theorem demonstrates that an ideal learner, with sufficient positive evidence, will learn to respect these linguistic restrictions, it does not claim that the learner will necessarily be aware of the specific linguistic theory that underlies the restrictions. Instead, the theorem concerns the predictions of the learner, rather than the specific representational methods that the learner might use.

4. Learning grammatical judgments

The task of prediction naturally extends to that of grammaticality judgments by examining predictions for larger chunks of linguistic material (e.g., words, phrases) and asking how often

the predicted utterance will correspond to a continuation that is a grammatical sentence. One can then ask: How often does the learner overgeneralize what is possible in the language such that its guesses violate the rules of the language (e.g., predict a contraction of *is* where it is not allowed)? Also, how often does the learner undergeneralize what is possible, such that it fails to guess continuations that are acceptable (e.g., never predict a contraction when it is allowed)? Results for overgeneralization and undergeneralization errors are examined in turn.

4.1 Grammaticality errors: overgeneralization

When considering grammaticality, it is convenient to consider language input as a sequence of words, rather than coded as a binary sequence (as was done above). Thus, instead of dealing with distributions, μ, λ, over *binary* sequences, one may consider distributions P_μ and P_λ.over sequences of a finite vocabulary of words. Suppose that the learner has seen a corpus, *x*, of *j-1* words and has a probability $\Delta_j(x)$ of incorrectly guessing a *j*th word which is ungrammatical, i.e., the string cannot be completed as a grammatical sentence. One can write:

$$\Delta_j(x) = \sum_{\substack{k: xk \text{ is ungrammatical,} \\ l(x)=j-1}} P_\lambda(k \mid x) \qquad (9)$$

As before, we focus on the *expected* value $\langle \Delta_j \rangle$:

$$\langle \Delta_j \rangle = \sum_{x: l(x)=j-1} P_\mu(x) \Delta_j(x) \qquad (10)$$

This expected value reflects the expected amount of overgeneralization that the learner makes, starting with different linguistic inputs *x*, multiplied by the probability of occurrence of each *x*. Applying the prediction theorem, and in line with the intuition that overgeneral grammars will

produce poor predictions, it is possible to derive the following 'overgeneralization theorem' (Chater & Vitányi, 2008):

$$\sum_{j=1}^{\infty} \langle \Delta_j \rangle \leq \frac{K(\mu)}{\log_e 2} \qquad (11)$$

That is, the expected amount of probability devoted by the learner to overgeneralizations, in the course of encountering an infinite corpus, sums to a finite quantity. Hence, the expected amount of overgeneralization must tend to zero as more of the corpus has been encountered. Note, too, the expected number of overgeneralization errors depends on the complexity of the language, $K(\mu)$, the length of the shortest universal program that can simulate P_μ. $K(\mu)$, is effectively a measure of the complexity of the underlying computational mechanism generating the language.

The ability to deal with overgeneralization of the grammar from linguistic experience is particularly relevant to previous discussions of the "logical problem" of language learnability, discussed above (Baker & McCarthy, 1981; Hornstein & Lightfoot, 1981; Pinker, 1979; Pinker, 1984). The learner only hears a finite corpus of sentences. Assuming the language is infinite, a successful learner must therefore infer the acceptability of an infinite number of sentences that it has never heard. Thus, *not* hearing a sentence cannot be evidence against its existence. As noted above, this has raised the puzzle of whether it is possible for overly general grammars to be corrected. The overgeneralization theorem shows that an ideal learner using the simplicity principle will eliminate overly general grammars because overly general grammars produce overly long codes for linguistic material.

4.2 Grammaticality errors: overgeneralization

The universal distribution used by the ideal learner was defined as being a combination of all possible (computable) distributions over corpora, and thus all grammatical sentence in the language will always be deemed possible (assigned non-zero probability). However, while an ideal learner will not ever strictly undergeneralize, i.e., deem a grammatical utterance to be ungrammatical, an ideal learner could drastically *underestimate* a sentence's probability of occurence. Thus, one can investigate the extent to which an ideal learner might commit such errors of 'soft' undergeneralization. Formal results have been derived putting an upper bound on the number of soft undergeneralizations an ideal learner will make. Suppose that the learner underestimates, by a factor of at least *f*, the probability that word *k* will occur after linguistic material *x*. That is, $P_\lambda(k|x)\, f \leq P_\mu(k|x)$. One can write the probability $\Lambda_{j,f}(x)$ that the word that is the true continuation will be one of the *k* for which this underestimation occurs:

$$\Lambda_{j,f}(x) = \sum_{k: f.P_\lambda(k|x) \leq P_\mu(k|x)} P_\mu(k \mid x) \qquad (12)$$

The corresponding *expected* probability is:

$$\left\langle \Lambda_{j,f} \right\rangle = \sum_{x: l(x)=j-1} P_\mu(x) \Lambda_j(x) \qquad (13)$$

Then, the following undergeneralization theorem has been derived, which bounds the expected number of undergeneralization errors throughout the corpus, i.e., $\sum_{j=1}^{\infty} \left\langle \Lambda_{j,f} \right\rangle$:

$$\sum_{j=1}^{\infty} \left\langle \Lambda_{j,f} \right\rangle \leq K(\mu) \frac{1}{\log_2 f/e} \qquad (14)$$

so long as $f > e$, where *e* is the mathematical constant 2.71...

Thus the expected number of 'soft' undergeneralizations is bounded, even for an infinitely long sequence of linguistic input and the expected rate at which such errors occur converges to zero. As with overgeneralizations, the upper bound is proportional to $K(\mu)$, the

complexity of the underlying computational mechanism generating the language (including, presumably, the grammar). The more severely one sets the underestimation factor *f* to be, the fewer such undergeneralizations can occur.

In summary, formal results have shown that an ideal learner, using the universal probability distribution, P_λ, derived from the simplicity principle, can learn to make accurate grammaticality judgments that avoid both overgeneralizations and undergeneralizations. In the description above, grammaticality judgments have been framed as the process of guessing which *word* comes next. However, it is important to note that these results extend to all other units of linguistic analysis, e.g., prediction of utterances on the level phonemes, syllables, or sentences.

5. Learning to Produce Language

One method of describing language production is to assume that it is simply a matter of predicting future utterances of arbitrarily long lengths. Thus, a learner, given an entire history of linguistic input, eventually "joins in" and starts *saying* its predictions. Production success can be assessed by how well these productions blend in with the linguistic input --i.e., how well the learner's productions match those that other speakers of the language (i.e., those producing the learner's corpus) might equally well have said. This is, of course, a highly limited linguistic goal, given that a key purpose of language is to express one's own thoughts, which may be diverge from what other's have said before. (We will consider how this limitation can be overcome in the next section.) However, as a first step, one can begin to assess a learner's ability to speak language by assessing whether the learner can blend into the on-going "conversation."

Blending in can be described as the ability to match the actual probability that a new sequence of utterances, *y*, will follow the previous utterances, *x*, which have been heard so far in the conservation. This is the probability $\mu(y|x)$, which reflects the distribution of continued sequences that would be uttered by speakers of the language. As before, the learner's stream of utterances can be defined on any linguistic level, e.g., phonemes, words or sentences. Because the ideal learner generates utterances using the distribution it learned in prediction, λ, the learner will predict continuations according to $\lambda(y|x)$. The learner will blend in, to the extent that $\lambda(y|x)$ is a good approximation to $\mu(y|x)$--i.e., the extent to which the learner has a propensity to produce language that other speakers have a propensity to produce. Note, though, that the objective is now not merely predicting the next binary code, piecemeal; the material to be predicted, *y*, can be an arbitrarily large chunk of linguistic material (e.g., an entire clause or sentence).

It turns out that $\lambda(y|x)$ is a good approximation to $\mu(y|x)$ (Li & Vitányi, 2008): If μ is, as above, a probability distribution associated with a monotone computable process, and λ denotes the universal distribution, then for any finite sequence *y*, as the length of sequence *x* tends to infinity:

$$\frac{\lambda(y \mid x)}{\mu(y \mid x)} \to 1 \qquad (15)$$

with a probability tending to 1, for fixed utterance *y* and growing prior linguistic experience *x*. Thus, viewing (15) in the context of language production, this means that, in the asymptote, the learner will blend in arbitrarily well, so that its language productions are indistinguishable from those of the language community to which it has been exposed.

6. Learning to map linguistic forms to semantic representations

In addition to being able to predict, judge, and produce linguistic regularities, language acquisition also involves associating linguistic forms with *meanings*. Indeed, to the ability to judge grammaticality, or produce language indistinguishable from that of one's speech community, would be pointless unless it were associated with the ability to communicate: to map from utterances to some representation of their interpretations, and back (we remain neutral here about nature of these representations).

A common assumption among researchers (Pinker, 1989) is that the child can infer interpretations from linguistic context. Therefore the problem of learning interpretations from linguistic input can be framed as a problem of induction from pairs of linguistic and semantic representations. One can then show that, given sufficient pairs, the ideal learner is able to learn this mapping, in either direction, in a probabilistic sense. This result means that the mapping between linguistic and semantic representations can be many-to-many. That is, linguistic representations are often ambiguous; and the same meaning can often be expressed linguistically in a number of different ways.

Concretely, we view the learner's problem as learning a relation between linguistic representations (e.g., as the i^{th} string of words), S_i, and a semantic interpretations, I_j, (representing the j^{th} meaning of the string). Suppose that the language consists of a set of ordered pairs $\{<S_i,I_j>\}$, which we sample randomly and independently, according to computable probability distribution $Pr(S_i,I_j)$.

Now we can apply the prediction theorem, as described above, but where the data now consist of pairs of sentences and interpretation, rather than strings of phonemes or words. So, when provided with a stream of sentence-interpretation pairs sampled from $Pr(S_i,I_j)$, the learner

can, to some approximation, infer the joint distribution $Pr(S_i,I_j)$. But, of course, approximating this joint distribution is only possible if the learner can approximate the relationship between sentences S_i and interpretations I_j.

Writing the length of the shortest program that will generate the computable joint distribution, $Pr(S_i,I_j)$, as $K(\Pr(S_i,I_j))$, the prediction theorem above ensures that this joint distribution is learnable from positive data by an ideal learner. Specifically, by (8), this has an expected sum-squared error bound of $\frac{\log_e 2}{2} K(\Pr(S_i, I_j))$. Hence the expected value of error per data sample, will tend to zero because this bound is finite, but the data continues indefinitely.

If ordered pairs of $<S_i,I_j>$ items can be predicted, then the relation between sentences and interpretations can be captured; and this implies that the mapping from sentences to probabilities of interpretations of those sentences, $Pr(I_j| S_i)$, and the mapping from interpretations to probabilities of sentences with those interpretations, $Pr(S_i|I_j)$, are learnable.[1] Thus, we can conclude that the ideal learner is able to learn to map back and forth between sentences and their interpretations, given a sufficiently large supply of sentence-interpretation pairs as data. That is, in this specific setting at least, the relation between form and meaning can be derived from positive data alone.

7. Scaling down simplicity: A practical method for assessing learnability

We have described a range of theoretical results concerning the viability of language learning by simplicity. But how far can the simplicity-based approach be "scaled-down" to inspire

[1] Of course if interpretation I_j where $Pr(I_j)=0$, then the fact that $Pr(I_j, S_i)$ can be approximated arbitrarily well says nothing about $Pr(S_i|I_j)$; similarly for sentences S_i where $Pr(S_i)=0$. But the learner surely needs only learn sentences that express meanings that might actually arise; and interpret sentences that might actually be said, so this restriction is fairly mild.

concrete models of learning? The practical instantiation of the simplicity principle has been embodied using the minimum description length (MDL, Rissanen, 1987) and minimum message length (MML, Wallace & Freeman, 1987) frameworks. And indeed, simplicity has been widely explored as general principle underpinning concrete models in a range of areas of perception and cognition (e.g., Attneave & Frost, 1969; Jacob Feldman, 2000; Hochberg & McAlister, 1953; Leeuwenberg, 1969).

Simplicity-based approaches have also been applied directly to building models of learning various aspects of language (e.g., Brent & Cartwright, 1996; Dowman, 2000; Ellison, 1992; Goldsmith, 2001; Onnis, Roberts & Chater, 2002; Wolff, 1988; Vousden, Ellefson, Solity & Chater, 2011); and closely related Bayesian methods have also been widely employed (Kemp, Perfors & Tenenbaum., 2007; Langley & Stromsten, 2000; Perfors, Regier & Tenenbaum, 2006; Stolcke, 1994).

The simplicity-based approach does not assume any particular linguistic representation: grammar rules can be specified using any type of representation the linguist chooses to use (e.g., phrase structure grammar). A code length can then be assigned to both the rules of the grammar as well as to the corpus under those rules (the corpus might consist of all the utterances that a learner has experienced so far, or a subset of these).

Suppose that we wish to evaluate how much data is required to learn a particular linguistic regularity (we know, from the above results, than any computable regularity will be learned arbitrarily well, given an ideal learner with sufficient data). This can be heuristically assessed by comparing two grammars, which are identical aside from the fact that only one of these captures the regularity. For example, consider how we might assess whether the corpus contains sufficient information to learn the restrictions on cases where *is* can be contracted

(described earlier). A grammar containing this additional regularity requires, of course, greater code-length than one that does not; but, on the other hand, as the resulting model of the language is more accurate, the code length of the corpus, given this more accurate model, will be shorter. Whether the 'balance' favors the more complex but accurate grammar (thus allowing the restrictions on contraction to be learned) depends on the corpus. For a null, or a short, corpus, the advantage of a more accurate language model will not be sufficient; however, once the corpus becomes sufficiently long, the more accurate model will produce a shorter, overall code-length, and the regularity will be learned. The question is: how long does the corpus need to be, for the regularity to learnable?

Inherent in the simplicity principle is the trade-off between simpler vs. more complex grammars: Simpler, over-general grammars are easier to learn, but because they are less accurate descriptions of actual language statistics, they result in inefficient descriptions of language input. More complex grammars, which include linguistic restrictions, are more difficult to learn, but they better describe the language and result in more efficient descriptions of the language. From this viewpoint, language learning can be viewed in analogy to investments in energy-efficient, money-saving appliances. By investing in a more complicated grammar, 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 energy-efficient appliance that saves money with each use. Intuitively, a linguistic restriction is learned when the relevant linguistic context occurs often enough that the accumulated savings makes the more complicated grammar worthwhile.

Recently, a practical framework for assessing learnability of a wide variety of linguistic constructions under simplicity has been proposed by Hsu and Chater (2010). Using natural-

language corpora to simulate the language input available to the learner, this framework quantifies learnability (in estimated number of years) for any given construction with restriction rules, such as those governing the contraction of *is* mentioned in the Introduction. When using this framework there are two main assumptions that are left to the discretion of the user. The first assumption is the description of the grammatical rule to be learned, i.e., a description of an original, incorrect (over-general) grammar and the new, correct grammar, which contains the restriction rule. This description can be made under any grammar encoding scheme the linguist chooses. The second assumption is the corpus approximating the learner's input. Given these two assumptions, the framework provides a method for quantifying an upper bound on learnability from language statistics alone. The framework allows for comparison of results which arise from varying these two main assumptions, providing a common forum for quantifying and discussing language learnability. This framework assumes an ideal statistical learner and thus provides an upper bound on learnability based on language statistics. However, measures of learnabity should give an indication for how relatively statistically learnable constructions are in reality.

While the details of implementing this framework are described elsewhere (Hsu & Chater, 2010; Hsu, Chater & Vitányi, 2011), an intuitive description of how this framework works is as follows: Under this framework, the learnability is affected by three factors: (1) Complexity of the rule to be learned. More complexity increases grammar cost and decreases learnability. (2) How often the missing "restricted form" would be expected to appear based on other similar constructions. For example, consider the restriction that *is* cannot be contracted at the end of a sentence. This restriction will be more easily learned if other contractions do often appear at the end of a sentence—which would make the absence of *is*-contractions more

suspicious. On the other hand, if other contractions also rarely appear at the end of a sentence, the absence of *is*-contractions at the end of the sentence will be difficult to notice. (3) How often the construction being learned appears in the language input. (1) and (2) determine how many occurrences are needed for learning and (3) (estimated from corpora serving as input) then will determine how many years it will take to accrue the number of occurrences needed.

Hsu and Chater (2010) used this general framework to quantify learnability using the assumption that grammars can be described using probabilistic context free grammars and that a learner's input can be approximated using corpora of adult speech and writing, such as the Corpus of Contemporary American English (COCA). They found that the predicted years required to learn a construction varied widely from being immediately learnable to taking over a lifetime. It was hypothesized that support for the learnability predictions could be found by examining the difference in grammatical acceptability between the correct and incorrect form of the construction, as judged by adult native language speakers. This was supported in an experiment on adult native language speakers in Hsu, Chater and Vitányi, (2011). Figure 1a shows the range in predictions for 15 constructions from Hsu and Chater (2010), sorted in descending learnability. Figure 2a shows how often the constructions occur per year, estimated from COCA. Note that the occurrence rate does not monotonically decrease with the years required to learn the construction because of the other factors that affect learnability (1) and 2) listed above). Results found that the prediction of MDL learnability analysis were supported: as learnability (i.e., log(1/predicted years needed) increased, the difference in the grammatical acceptability of the grammatical vs. ungrammatical form of the construction also increased.

8. Conclusion

In this paper, we have reviewed some recent results concerning learning language from experience by employing the simplicity principle: that is, favouring models of the language to the degree that they provide a short encoding of the linguistic input. We have shown theoretical results that indicate that an "ideal learner" implementing the simplicity principle can learn to predict language from experience; to determine which sentences of a language are grammatical to an arbitrarily good approximation (assuming, somewhat unrealistically, that the corpus of linguistic experience is noise-free, i.e., containing only grammatical sentences); produce language and to map between sentences and their interpretations. This "ideal" learning approach is valuable for determining what information is contained in a corpus. Yet it cannot be implemented computationally, as the relevant calculations are known to be uncomputable (Li Vitányi, 2008). Nonetheless, we have also shown how a local approximation to such calculations can be used to choose between different grammars which do or do not contain specific regularities (especially those concerned with exceptions) that have been viewed as posing particular problems for theories of language acquisition. Overall, these results form part of a wider tradition of analytic and computational results on language learning which suggest that general *a priori* arguments about whether language acquisition requires language-specific innate constraints can be replaced by a more precise formal and empirical analysis.

Figure 1:

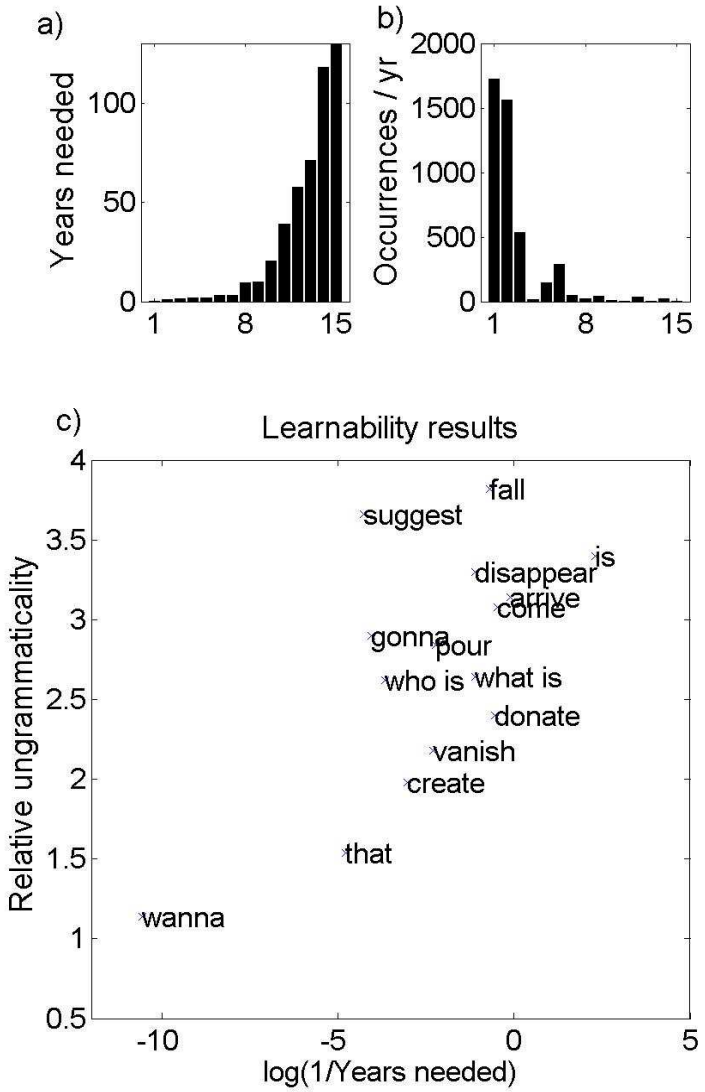

Figure 1: Simplicity framework's learnability predictions and experimental evidence: These results are re-plotted from Hsu, Chater & Vitányi (2011). (a) Estimated years required to learn construction. (b) Number of occurrences per year (estimated from COCA). (c) Relative grammaticality vs. learnability for Sentence Set 1 ($r = 0.67; p = 0.006$). Relative grammaticality judgements were elicited from 200 native English speakers in an online study. Learnability is log of the inverse of the number of estimated years needed to learn the construction.

Reference List

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

Angluin, D. (1988). Identifying languages from stochastic examples. Technical Report. Department of Computer Science, Yale University.

Attneave, E, & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception & Psychophysics, 6*, 391-396.

Baker, C. L. & McCarthy, J. J. (1981). *The logical problem of language acquisition.* Cambridge, Mass: MIT Press.

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.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93-126.

Brown, R. & Hanlon, C. (1970). *Derivational complexity and order of acquisition in child speech*. (J.R. Hayes ed.) New York: Wiley.

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

Chater, N. & Vitányi, P. (2002). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7,* 19–22.

Chomsky, N. (1980). *Rules and representations.* Cambridge, MA: MIT Press.

Christiansen, M. H. & Chater, N. (1994). Generalization and connectionist language learning. *Mind & Language, 9,* 273-287.

Christiansen, M. H. & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23,* 417-437.

Christiansen, M. H. & Chater, N. (2007). Generalization and connectionist language learning. *Mind & Language, 9,* 273-287.

Alexander Clark and Rémi Eyraud. (2007). Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, 8, 1725-1745.

Crain, S. & Lillo-Martin, D. (1999). *Linguistic theory and language acquisition.* Oxford: Blackwell.

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.). *Proceedings of the Twenty Second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Dresher, B. & Hornstein, N. (1976). On Some Supposed Contributions of Artificial Intelligence to the Scientific Study of Language. *Cognition, 4,* 321-398.

Ellison, M. (1992). The machine learning of phonological structure. PhD Thesis, University of Western Australia.

Elman, J. (1990). Finding Structure in Time. *Cognitive Science, 14,* 179-211.

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

Feldman, Jerome (1972). Some decidability results on grammatical inference and complexity. *Information and Control, 20,* 244–262.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2007). Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Gold, E. M. (1967). Language identification in the limit. *Information and control, 16,* 447-474.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27,* 153-198.

Hochberg, J. & McAlister, E. (1953). A quantitative approach to figure "goodness." *Journal of Experimental Psychology, 46,* 361-364.

Horning, J. J. (1969). A study of grammatical inference. Technical Report CS 139, Computer Science Department, Stanford University.

Hornstein, N. & Lightfoot, D. (1981). *London: Longman, 1981*. London: Longman.

Hsu, A. & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science, 34,* 972-1016.

Hsu, A., Chater, N., & Vitányi, P. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition, 120,* 380-390.

Jain, S., Osherson, D. N., Royer, J. S., & Kumar Sharma, A. (1999). *Systems that learn (2nd edition)*. Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypothesis with hierarchical Bayesian models. *Developmental Science, 10,* 307-321.

Klein, D. & Manning, C. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition, 38,* 1407-1409.

Langley, P. & Stromsten, S. (2000). Learning context-free grammars with a simplicity bias. *Proceedings of the Eleventh European Conference on Machine Learning,* 220-228.

Leeuwenberg, E. L. J. (1969). Quantitative specification of information in sequential patterns. *Psychological Review, 76,* 216-220.

Li, M. & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications* (3rd Edition). Springer.

Mach, E. (1959). *The analysis of sensations and the relation of the physical to the psychical*. New York: Dover Publications (Original work published 1886).

MacWhinney, B. (1993). The (il)logical problem of language acquisition. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 61–70). Mahwah, NJ: Erlbaum.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language, 31*, 883–914.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition, 46*, 53-85.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition? *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 720-725.

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. (1979). Formal models of language learning. *Cognition, 7*, 217-283.

Pinker, S. (1984). *Language learnability and language development*. (7 ed.) Harvard Univ Pr.

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

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*, 707-784.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: Bradford Books/MIT Press.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B, 49*, 223–239.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72*, 68–109.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell Systems Technical Journal, 31*, 64.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory, IT, 24*, 422-432.

Steyvers, M., Griffiths, T., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*, 309-318.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models.* Department of Electrical Engineering and Computer Science, University of California Berkeley.

Tomasello, M. (2004). Syntax or semantics? Response to Lidz et al. *Cognition*, *93*, 139–140.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B, 49*, 240–251.

Wharton, R. M., (1974). Approximate language identification, *Information and Control*, 26, 236-255

Vousden, J.I., Ellefson, M.R., Solity, J.E., & Chater, N. (2011). Simplifying reading: Applying the simplicity principle to reading. *Cognitive Science*, *35*, 34-78.

Wolff, J. G. (1988). Learning syntax and meanings through optimisation and distributional analysis. In Y. Levy, I. M. Schlesinger & M. D. S. Braine (Eds.), *Categories and processes in language acquisition*, (pp. 179-215). Hillsdale, NJ: LEA.