Gold's Theorem and cognitive science[*]

Kent Johnson
Department of Logic and Philosophy of Science
University of California, Irvine
johnsonk@uci.edu
http://hypatia.ss.uci.edu/lps/home/fac-staff/faculty/johnson/

[Forthcoming in *Philosophy of Science*]

Abstract. A variety of inaccurate claims about Gold's Theorem have appeared in the cognitive science literature. I begin by characterizing the logic of this theorem and its proof. I then examine several claims about Gold's Theorem, and I show why they are false. Finally, I assess the significance of Gold's Theorem for cognitive science.

1. Introduction

Over the years, negative evidence in language acquisition has received a lot of attention. (Negative evidence is information regarding what possible sentences are *not* sentences of the language to be learned.) It is often thought that children lack negative evidence when they acquire their native language (e.g., Marcus 1993), and that this fact supports a rationalist view of the mind over an empiricist view. One argument from 'No negative evidence' to rationalism centers on the 'Logical Problem of Language Acquisition' (LPLA). Briefly, LPLA goes like this. If a child learning a given language doesn't use negative evidence, then it is logically possible that she would begin to acquire a language whose grammatical sentences included all those of the target language, plus some more. E.g., if a child was learning English, she could begin to acquire a language that included all the grammatical sentences of English, plus sentences of VSO clausal order, like **Kicked Mary the boy* and **Sang Susan*. If the child received no information that these extra sentences aren't part of

[*] I am grateful to Robert Matthews for stimulating my thinking about learnability; his positive influence can be seen throughout this paper. I am also grateful to Brian Skyrms and two anonymous reviewers for much useful feedback. All errors are mine.

English, then she would have no way of knowing that she was not learning English, but a syntactically more expressive language instead. Thus, we would predict that the child would end up acquiring the ‘larger’ language. However, this does not happen: children acquire their target language, and not languages whose sentences contain the sentences of the target language as a proper subset. How children do this is the Logical Problem of Language Acquisition. A common route from LPLA to rationalism goes as follows. If there is no negative information, then there must be some other mechanism that enables the child to learn her language instead of a more expressive language. Such a mechanism would most plausibly be a cognitive ability that somehow prevents the child from entering a situation where negative evidence is needed. Any such cognitive ability would appear to be domain-specific to language and not learned. Thus, the ability must be innate, so rationalists are right about language acquisition and empiricism is false.

LPLA and its use in support of rationalism have received much attention (e.g. Cowie 1999 and citations therein). LPLA is often thought to have a precise mathematical realization in a theorem from E. Mark Gold’s seminal paper ‘Language identification in the limit’ (Gold 1967). Although views differ about what exactly ‘Gold’s Theorem’ shows, a great many cognitive scientists treat it as an attempt to provide a kind of evidence for rationalism: substantial innate knowledge or constraints are needed to facilitate language acquisition. Such an attempt is bold: it would be impressive to solve the millennia-old debate about rationalism vs. empiricism, and to do it with a simple theorem of mathematical logic would be stunning. So it’s unsurprising that Gold’s Theorem has received much criticism, especially from those with empiricist leanings. This criticism has come from many perspectives, including neuroscientific discussions (Deacon 1997; Elman et al. 1996), child language acquisition (Hirsh-Pasek and Golinkoff 1996; Cowie 1997), the nature of concepts (Prinz 2002), and the debate about innate abilities and knowledge (Cowie 1999). Despite it’s impressive impact in cognitive science, Gold’s Theorem is frequently misinterpreted. All of the authors listed above, for instance, have made false – and in some cases wildly inaccurate – claims about the theorem. Indeed, even rationalists, who might welcome support from the theorem, have made incorrect criticisms of the general assumptions that drive it (Chomsky 1986). The widespread confusion about the theorem is especially surprising, since even those who have misunderstood it have claimed that its proof ‘is quite easy to grasp intuitively’ (Cowie 1999, 194).

The aim of the present paper is to clear up these misunderstandings of Gold’s Theorem. I do this in three steps. First (§1), I describe the logic behind Gold’s Theorem and its proof. Next (§2), I examine several claims about Gold’s Theorem, and I show why they are false. Finally, (§3), I assess the significance of Gold’s Theorem for cognitive science. We’ll see that Gold’s Theorem and LPLA are logically very distinct phenomena. While this comparison with LPLA brings out some strengths of Gold’s Theorem, a significant weakness still remains.

Before beginning, a word on the project is in order. Of all learnability results, Gold’s Theorem is the most famous, and was the first one to become well-known in cognitive science. As one psychologist put it, Gold’s theorem had a ‘chilling effect’ on the psychological community when it first became widely known there.[1] The logic of Gold’s Theorem and its proof are quite simple, which makes it an excellent subject of attention for several reasons. First, it’s easy to see precisely what the theorem does and does not say. Second, the theorem’s simplicity is a great source of power. Gold’s Theorem avoids many objections that require assumptions the theorem doesn’t make. Third, the various cognitive interpretations of the theorem’s crucial elements are easy to see. In that sense, Gold’s Theorem presents a useful case study in mathematical modeling for such areas as linguistics, psycholinguistics, and even philosophy of language, which often do not traffic in such things. By studying Gold’s Theorem, we can get an idea of which considerations matter to other, more sophisticated models of learning.

[1] Stephen Jose Hanson, Cognitive Science Proseminar, Rutgers, 1998.

2.

Gold's 1967 paper proves a classification of various groups of languages in terms of their learnability. Most of this work is of interest only from the standpoint of mathematical logic or theoretical computer science. However, one part of the classification looks pertinent to cognitive science. In this section, I characterize the general logic behind this result and its proof.

In order to formally represent a cognitive phenomenon like language learning, we must find mathematical surrogates for various elements of the empirical world. As a bit of background, a *language* can be thought of as merely the set of sentences that are grammatical in that language. I'll say more about this later, but for now we can follow Gold and fix some finite alphabet Σ, and then create Σ*, the set of all the finite sequences of elements of Σ, and then define a language as any subset of Σ*. We also need mathematically precise representations of (i) the learner's environment, (ii) the nature of the learners, including the set of hypotheses that the learner selects from, and (iii) a criterion of successful learning. (Particular representations of (i), (ii), and (iii) can be thought of as a *model* of learning.) To prove Gold's Theorem, the following definitions suffice. We may consider the learner's *environment* to be any infinite sequence $<a_1, a_2, a_3, \ldots >$ of sentences (i.e., elements) from the target language to be learned, with the requirement that every sentence of the language appears at least once in the sequence.[2] The idea is that the sentences are temporally ordered, so that at time t_n the learner receives datum a_n. Thus, the learner's guess as to the target language at time t_n will be based on no more information than $<a_1,\ldots a_n>$, plus any other information that is built into the learner. (The notion of time needn't be taken too literally here, as measuring out equal units. Rather, time merely serves to order the linguistic data. In the life of a child, there may be less than a second between t_{89} and t_{90}, but a whole naptime between t_{92} and t_{93}.) We may represent a *learner* as any function that takes finite initial

[2] Gold didn't consider environments containing sentences that are not from the target language. But cf. e.g. Jain et al. 1999.

sequences of an environment as input, and yields as output a guess as to the target language.[3] A crucial feature of the learner concerns which guesses she can make. If she is utterly unrestricted, then her possible hypotheses form the class of every logically possible language. If she is restricted, her hypothesis space will be smaller. If she is hardwired for just one language, then the hypothesis space contains exactly one language. Given an environment E and a language L, the learner *learns* L *given* E iff there is some time t_n such that at t_n and all times afterward, the learner correctly guesses that L is the target language present in the environment. (Gold himself called this condition 'identification in the limit'.) More generally, if a learner learns L given any environment E (of sentences of L) whatsoever, then she *learns* L. Given a collection C of languages, a learner *learns* C iff she learns every language in C. Finally, if there exists a learner that learns C, then C is *learnable.*

Given these definitions, the logic of Gold's Theorem shows that there are many logically possible classes of languages that no learner can learn. To see this, let C be an infinite collection of languages, which contains as a (possibly proper) subset the languages $\{L_\infty, L_1, L_2, \ldots\}$. Let us also suppose that that L_i is a proper subset of L_{i+1}, yielding the sequence:

(*) $L_1 \subset L_2 \subset L_3 \subset \ldots$

Finally, suppose that a sentence is contained in L_∞ if and only if that sentence is contained in another language L_i.[4] Let's say that any class of languages that meets the conditions we have just imposed on C has the *Gold Property*. (More carefully: a class C of languages has the

[3] The learner could also be allowed to yield a 'no guess' output, indicating that the environmental input so far has not been enough to make it try to guess the target language. Such details don't affect the proof, so I ignore them.

[4] Notice that if a sentence is contained in L_i, then it is contained in all the infinitely many languages L_{i+n}, for any number n.

Gold Property iff C contains (i) a countable infinity of languages L_i such that $L_i \subset L_{i+1}$, for all $i > 0$, and (ii) a further language L_∞ such that for any $i > 0$, x is a sentence of L_i only if x is a sentence of L_∞, and x is a sentence of L_∞ only if x is a sentence of L_j, for some $j > 0$.) Using the definitions given above, we now prove the following theorem:

(GT) Any class of languages with the Gold Property is unlearnable.

To prove (GT), let C be some class of languages with the Gold Property. So C contains infinitely many languages structured as in (*), plus one more language that contains exactly the sentences in the other languages. If a learner could learn C, then for any language L in C, the learner could be given only sentences from L (where these sentences appear in any order, repeats allowed, and every sentence in L appears at least once), and at some point in time, the learner would correctly guess that she was receiving sentences from L, and she would never give up that guess no matter how many more sentences from L she encountered. But this can't happen. To see why, pick your favorite learner ϕ. That is, let ϕ be any function from finite sequences of sentences from a language in C to (names of) languages in C. Given ϕ, we now show that there must be some language L in C and some environment E from L such that ϕ does not learn L given E. Which L and E do the job depends on the exact nature of ϕ. The following strategy for building an environment which will fool ϕ shows that L and E must exist. We start by giving ϕ sentences from L_1, and continue doing so until it 'converges' onto a guess that L_1 is the target language, and will not change its guess as long as it receives only sentences from L_1.[5] (Notice that if ϕ doesn't converge onto a guess that L_1 is the target language, then we could keep extending this initial sequence of sentences from L_1 into

[5] 'Peeking inside ϕ's head' isn't crucial here: we only need to demonstrate that a language and an environment exist. Thus, talk of knowing when ϕ converges is only a metaphorical way of characterizing which environment will fool ϕ.

infinity, which would result in an environment for L_1. Then we would have our L and E.) Once ϕ has converged onto a guess that L_1 is the target language, we then start adding sentences from L_2 to this environment. Since L_2 contains all the sentences in L_1, plus some more, the initial sentences we gave ϕ to get it to converge to a guess of L_1 are also contained in L_2. Now we give ϕ sentences from L_2 until it converges onto a guess that L_2 is the target language (which it must eventually do, unless it is incapable of learning L_2). In general, we get ϕ to converge onto a guess that L_n is the target language, and then we start giving it sentences from L_{n+1}. Since L_n is always a proper subset of L_{n+1}, we are always building a legitimate environment from L_{n+1} (until we switch to L_{n+2}). Moreover, since L_∞ contains all (and only) the sentences from the entire sequence of languages, at no point in this strategy will we give ϕ sentences that are not part of L_∞.

Now ϕ faces an insuperable dilemma. On the one hand, if ϕ changes its guess infinitely many times (e.g., from L_1 to L_2, to L_3, etc.) then by definition, it does not learn the target language. But if ϕ changes its mind only finitely often, then at some point it will have fixed on a particular guess that some language L is correct. But L either is or is not L_∞. If it is L_∞, then ϕ is guessing that it is receiving sentences from the biggest language, even though it has only encountered sentences that belong to a smaller language. Whenever ϕ starts guessing that L_∞ is the correct language, we were at that time trying to get ϕ to guess that some other language L_n was the correct one. Since ϕ won't converge to L_n, our strategy says to keep feeding ϕ sentences from L_n forever, thus creating an environment for L_n. So ϕ fails to learn L_n in this environment. On the other hand, suppose that ϕ converges to the language L_n, for some $n > 0$. But assuming it has followed the only successful strategy so far, once it begins to converge to L_n, our strategy was to begin supplying ϕ with sentences from L_{n+1}, and to do this until ϕ converges to L_{n+1}. Since, by hypothesis, ϕ never does this, we end up creating an environment for L_{n+1} in which ϕ never successfully learns L_{n+1}. So in all cases, there is some language L and environment E such that ϕ does not learn L given E. This completes the proof of (GT).

Gold's Theorem is found in Theorem I.8 of Gold 1967. (GT)'s proof characterizes the general logic behind Theorem I.8's proof. However, Theorem I.8 contains some technical details that are not relevant here, and are not included in (GT). For thoroughness, I have included an appendix containing a more detailed discussion of Gold's central results. The details in question are due to Theorem I.8's role in classifying learnable classes of languages. Nothing major hinges on the differences between (GT) and Theorem I.8. (Indeed, Gold 1967, 461 characterizes Theorem I.8 along the lines of (GT).) So I'll use the term 'Gold's Theorem' to refer to the general logical fact given in both (GT) and Theorem I.8, distinguishing the latter two only when relevant.

Gold's Theorem (and the results in Gold 1967, 1965 more generally) spawned a sizeable industry in computer science and mathematical logic in Formal Learning Theory. Some of this research follows Gold in studying learnability from a very abstract perspective (e.g. Feldman 1972; Valiant 1984; Osherson et al. 1986; Jain et al. 1999; Kelly 1996), while other research studies language acquisition in particular, using empirically motivated constraints (e.g. Wexler and Culicover 1980; Gibson and Wexler 1994; Niyogi and Berwick 1996; Bertolo 2001; Nowak, Komarova, and Niyogi 2001). In the next section, I examine some of the attention Gold's Theorem has received from the cognitive science community.

3.

Despite its simplicity, Gold's Theorem has been frequently misunderstood. A particularly striking example comes from Deacon (1997), who writes:

> [Gold] provided a logical proof which concluded that, without explicit error correction, the rules of a logical system with the structural complexity of a natural language grammar could not be inductively discovered, even in theory. What makes them unlearnable, according to this argument, is not just their

> complexity but the fact that the rules are not directly mapped to the surface forms of sentences. The result is that sentences exhibit hierarchic syntactic structures, in which layers of transformations become buried and implicit in the final product, and in which structural relationships between different levels can often produce word-sequence relationships that violate relationships that are appropriate within levels. From the point of view of someone trying to analyze sentence structure (such as a linguist or a young language learner), this has the effect of geometrically multiplying the possible hypothetical rules that must be tested before discovering the 'correct' ones for the language. (Deacon 1997, 127-128)

There are many problems with Deacon's criticism. For one thing, it mislocates the difficulty for learning *within* individual languages, and not in the *relation between* various possible languages. Deacon suggests that Gold's Theorem depends on the 'complexity' of the various languages. But as we've seen, the internal structure of the languages is irrelevant. To prove (GT), no constraints whatsoever were placed on the nature of the individual languages. Individual languages may be arbitrarily complex; all that matters for (GT) is *how* the various languages in the class are *related* to one another. In particular, it is crucial to (GT) that the class of possible languages have the Gold Property (cf. (*)), but it's irrelevant whether or not a particular language's sentences 'exhibit hierarchic syntactic structures'. (For similar reasons, it's incorrect that 'the space of possible grammars is too massive to select from without information about which grammars are wrong' (Prinz 2002, 210). After all, an infinite class of languages no two of which share a single sentence in common is trivially learnable.)

The distinction between complexity within languages and complexity between languages is important. The idea that complexity within a language could be relevant to learnability suggests that a single language could be unlearnable. Such an idea is not only suggested by Deacon, it is explicit in Elman, Bates, et al 1996. They write 'Interestingly, sentences with relative clauses possess exactly the sort of structural features which may make (according to Gold 1967) a language unlearnable' (343).[6] Several things are incorrect about an individual language being unlearnable in Gold's sense. First, the notion is not defined: learnability only applies to classes of languages. Second, any class containing exactly one language would be trivially learnable, because there exists a constant function that always guesses that language. If the notion of learnability were extended in the natural way to apply to individual languages in non-singleton classes of languages, a similar argument would also show that all such languages are learnable in Gold's sense. (E.g., we might say that L is learnable in the class C iff there is a learner which, given any environment from L, learns L. But every such L is learnable: for each L, there is a constant learner who always guesses L.)

Other researchers have mistaken Gold's Theorem to be about particular learners, not collections of learners. Hirsh-Pasek and Golinkoff (1996), for instance, downplay the significance of Gold's work by claiming that 'Gold's learner was an unbiased learner' (Hirsh-Pasek and Golinkoff 1996, 2, 5). Pullum and Sholz speak of 'the procedure the learner is assumed to use' in the proof of Gold's Theorem (2003, 130). Similarly, Cowie writes that 'children are not Gold-style learners: they do not test every logically possible grammar against the data', and that Gold requires the learners to have an unbounded memory capacity and an ability to 'test every logically possible grammar against the data.' (1999, 195; cf. also 1997, 30). But (GT) proves that *every* learner fails to learn the specified class of languages, regardless of memory abilities. Many of these learners operate in ways that can be accurately described as employing biases, strategies, or memory limitations. Although most of these biases, etc. are pointless, some of them are psychologically interesting, such as: 'assume the language is generated according to rules R_1, …, R_n, unless there is a strong amount of counterevidence', and 'hypothesize the less restrictive language L instead of L' if no sentence that is in L' but not L occurs within 1,000 sentences after hypothesizing L''. The

[6] Elman, Bates et al. go on to report the difficulty the first author had with training a connectionist system to process simplified sentences with relative clauses.

proof of (GT) shows that regardless of what biases, strategies, or memory limitations children might have, they do not aid in learning certain classes of languages. So unless we know that the class of natural languages lacks the Gold Property, it's incorrect to say that

> general-purpose constraints such as 'Prefer more general hypotheses,' or 'Make a tentative universal generalization that all Fs are Gs if you've encountered *n* instances of Fs that are Gs,' would do the trick to ensure the learnability of natural languages'. (Cowie 1999, 195)

Regardless of how 'Prefer more general hypotheses' is spelled out formally, any learner that employs it falls within the scope of Gold's Theorem.[7] The fact that the theorem applies to all learners, including those with built-in biases, etc., is an important one, and I return to it below in §3.

There is a possible reason why Hirsch-Pasek and Golinkoff and Cowie have misinterpreted Gold's results. At various points, Gold employs a particular learning strategy called 'identification by enumeration' (458-462). A learner embodying this strategy is able to somehow enumerate the guesses it can hypothesize (i.e., it can figure out what the *n*th language is, for any *n*). At each point in time, the learner guesses the first language in this enumeration that is consistent with the data so far. Identification by enumeration is used in several theorems to establish the learnability of a certain classes of languages. But these results are all positive: they simply show that there exists a learner that learns the language class in question. Gold never explores whether an algorithm that behaves like a child could learn the same language classes. Given that Gold used identification by enumeration to show the learnability of certain classes of grammars, it is trivially false that this learner will 'test

[7] Thus, we can take a slightly stronger position than Matthews 2001, 225, who locates the problem with the unclarity of such a constraint. The real problem is that every precise implementation of the constraint fails to produce a successful learner.

every logically possible grammar against the data.' All the classes Gold considered were infinite, so for a learner to test every language in any such class would guarantee that the learner does not learn that class of languages. Thus, it may be that Hirsch-Pasek and Golinkoff and Cowie have misunderstood Gold's Theorem to apply only to a particular learner that was used in various results other than Gold's Theorem. (Evidence that this is the case for Cowie comes from her (incorrect) description of the proof of Gold's Theorem in Cowie 1999, 194 fn. 19, and her remarks about 'Gold-style learners' and their constraints on 195.)

A third type of misinterpretation of Gold's Theorem concerns the restrictions on the environments that learners experience. Hirsh-Pasek and Golinkoff claim that

> Gold's learner received syntactic information in isolation from other forms of input (e.g., input from the environmental context, prosody, or social interaction). That is, Gold's learner heard a series of sentence strings and had to induce the units and rules of language as if in a vacuum. (Hirsh-Pasek and Golinkoff 1996, 2)

Such an interpretation is certainly natural: Gold's paper focuses on determining which strings are part of a language, and this seems almost by definition to be about only the syntax of the language. Indeed, Gold himself thought that the languages he was investigating were 'too simple to do anything with' (Gold 1967, 448). But whether we interpret Gold's Theorem as being about only syntax depends on what we take to be the nature of the elements of the language. Gold's only constraint here is that all the elements of a language be finite sequences drawn from some fixed finite set, and as we saw in §1, even that restriction is not needed to prove (GT). (GT) remains valid even if every language is countably infinite in size, and each element contains an infinitude of information (e.g., each sentence could be an infinite string, or something even more complex). In any case, even if languages are sets of strings of elements from a fixed finite alphabet, this is scarcely a restriction to syntax alone:

the fixed finite alphabet could contain the few hundred phonemes that are present in natural languages, plus elements to mark out syntactic structure, along with elements that enable us to describe any relevant semantic and pragmatic features. Thus, a ‘sentence’ could be e.g., an ordered sequence $<a_{11},\ldots, a_{1n}, a_{21},\ldots, a_{2m},\ldots, a_{jk}>$, where $< a_{11},\ldots, a_{1n}>$ describes the syntax of the sentence, $< a_{21},\ldots, a_{2m}>$ describes its meaning, and the other elements describe the relevant features of the ‘environmental context, prosody, or social interaction’. Indeed, sentences may be regarded as finely individuated sentence ‘tokens’, with a description of their context contained within. (That is, one might regard a sentence as containing not only phonological and syntactic information, but also information about appropriate contexts of use, typical age of acquisition of the various constructions involved, relative difficulty of processing, etc.~~.~~) In short, Gold’s notion of a ‘sentence’ hardly restricts what information sentences may contain. Since unlearnability results like (GT) quantify universally over learners, they hold even for very sophisticated learners, including those that make very subtle use of each sentence. For instance, the theorem applies to those learners that recognize that if $<a,\ldots,a_m,\ldots,a_n>$ is a sentence, where a_m is some bit of ‘environmental context’, then so is $<a_1,\ldots,a'_m,\ldots,a_n>$, for any environmental context a'_m that differs from a_m in some specified fashion. So despite Gold’s own remarks, Gold’s Theorem applies to highly complex and informative languages, and individual sentences can carry at least as much information as is presented by individual tokenings of natural language sentences in contexts.

The final possible misinterpretation I will mention comes from a criticism by Chomsky of some related work developed from Gold’s paper. Osherson, Stob, and Weinstein (1984) used Gold’s model of learning as a basis to prove a number of results in Formal Learning Theory. However, Chomsky warns, ‘one must be cautious in relating their results to our concerns here. [Osherson et al.] are considering E-language, not I-language, and are restricting attention to weak rather than strong generative capacity of grammars’ (Chomsky 1986, 149-150, fn. 89). A couple definitional points will make Chomsky’s charges clear. For present purposes, an E-language can be considered to be the set of sentences that are grammatical in a ‘public’ language, whereas an I-language is a state of a speaker’s mind or brain in virtue of which she speaks the language she does (cf. Chomsky 1986, 19 – 23). The strong generative capacity of a grammar can be thought of as the set of structural descriptions of sentences as they are generated by a grammar, whereas the grammar's weak generative capacity concerns only the sentences without any structural description. (To see the difference, let grammar G_1 be defined by the rule $X \rightarrow Xaa$, and let G_2 be defined by the rule $X \rightarrow aXa$. Both grammars have the same weak generative capacity: they both produce the set of even numbered *n*-tuples of *a*'s. However, their strong generative capacity is different: to produce *aaaa*, G_1 first applies the rule to ∅, which puts two *a*'s to the left, and then applies the rule to that string, which puts two more *a*'s to the left, yielding *[[aa]aa]*. But G_2 first applies its rule to ∅, which puts an a on either side, and then applies the rule again, yielding *[a[aa]a]*.) Thus, Chomsky's two charges against learning models like Osherson et al.'s – and by extension Gold's – are that they are only about sets of sentences, not minds, and that they ignore the constituent structure of sentences. Both charges are serious. If either one is right, the psychological interest of Gold's Theorem is severely compromised. I take each charge in turn. (I focus on the case for Gold, although my arguments apply to Osherson et al. too.)

Pace Chomsky, learning models such as Gold's can certainly be interpreted as being about the mind. The fundamental mathematical relation learners enter into is only that of using an initial finite segment of the environment to arrive at a guess as to the target language:

(**) $\phi(<s_1,\ldots,s_n>) = X.$

There is nothing in Gold's Theorem (or in his learning model more generally) that forces any particular interpretation of (**). For instance, although the learner is represented as a mathematical function ϕ, the theorem obviously doesn't assume that learners are nothing more than mathematical functions. Rather, it only assumes that the relevant structure of a learner's behavior can be so represented. Similarly, although languages are *represented* as sets of sentences, there is no need to say that they really are (just) sets of sentences. They could be states of the mind associated with those sets. Equation (**) simply describes the result of applying ϕ to a sequence. A natural psychological interpretation of (**) is that when a learner whose learning strategy is correctly modeled by ϕ experiences the environment represented by $<s_1,\ldots,s_n>$, (ceteris paribus, of course) her brain will be configured so as to instantiate language X. Interpreting Gold's model this way makes it clearly about I-languages. (Gold's discussion of the abstract model of identification (Gold 1967, 456 – 458) provides a useful clarification of the required connection between the information from the environment and the correct answer.)

The formal requirements Gold imposes on languages are also minimal enough that the theorem can be interpreted as being about strong generative capacity. In the first place, a result about unlearnability quantifies over all learners, including those whose guesses process sentences in terms of constituent structure. Clearly, a class C of languages is learnable when the languages are individuated by their strong generative capacity only if C is learnable when they are individuated by their weak generative capacity. The first task requires a learner to correctly determine the language that both agrees with the unstructured strings contained in the environment, and which agrees with the structure that those strings have. The second task only demands that the learner determine a language that satisfies the first of these requirements. But Gold's Theorem shows that if C has the Gold Property, and the languages in C are individuated by their weak generative capacity, then C is unlearnable. And if C is unlearnable, and $C \subseteq C^*$, then C* is unlearnable. Hence, C* is unlearnable if for each language in C, there are several languages in C* individuated by their strong generative capacity. Through its quantification over all learners, Gold's Theorem eliminates the possibility that strong generative capacity could be of help here. But in the second place, Gold's Theorem can also be interpreted as being 'directly' about languages individuated by their strong generative capacity. The sentences a learner receives can contain their constituent structure or derivational history, just as they can contain other information. Not all learners will use this information, but some will. But by including such structural information in the sentences, there is a very real difference between e.g., the grammars G_1 and G_2. G_1 generates sentences like *[[aa]aa]*, and G_2 generates *[a[aa]a]*; neither grammar generates *aaaa*. Thus, a learner that converged to G_1 instead of G_2 when given sentences from G_2 would simply be wrong.

In this section, we've seen several misinterpretations of Gold's Theorem.[8] Many of these errors have been made multiple times. There are, however, several useful discussions of the psychological interpretive scope of Gold's Theorem; e.g., Matthews 1984 and Demopoulos 1989. I suspect that if these two papers had been read more carefully, the errors discussed above would have been avoided. Although Gold's Theorem is immune to many objections, there still remains the question of whether it successfully provides a challenge to empiricism. That is the topic of the next section.

4

Although Gold's Theorem is frequently misinterpreted, it's nonetheless significant that such a wide range of authors – including linguists, philosophers, psychologists and neuroscientists – have taken the time to address the theorem. As with LPLA, much of the interest in Gold's Theorem is fueled by some deep beliefs about the nature of the mind, in particular rationalism vs. empiricism. Not every author explains how Gold's Theorem is supposed to support rationalism (although cf. Demopoulos 1989), but the following seems to capture the main line of argument. According to standard empiricist views, children learn language by a more or less straightforward inductive process. The child is placed in the company of speakers of the language (e.g., English), and gradually, the child 'catches on'. The empiricist holds that the process of 'catching on' is driven only by some very general learning abilities in the child. In particular, there is no reason to suppose that the cognitive mechanisms responsible for language acquisition are somehow particular to the domain of language

[8] And there are more; e.g. Howe 1993, 27 makes multiple incorrect claims about Gold's Theorem.

acquisition. Instead, these learning mechanisms are the same kind that enable us to learn that e.g. dogs like meat. If this is right, then there should be no substantial limitations on the possible languages that the child might hypothesize as correct. But now Gold's Theorem seems to raise a problem. If there are no substantial limitations on the child's possible hypotheses, then this class of hypotheses will have the Gold Property. So by Gold's Theorem, this class of languages should be unlearnable. But this result is unacceptable – we do learn our languages! So there must be constraints on which languages the child can hypothesize. So the child approaches the task of language learning with specific information about natural languages, where this information realizes the needed constraints. This information is not part of our general learning apparatus. Furthermore, because it is needed early on in the child's life, and appears to be very complicated, it is unlikely to have been learned. So the information must be innate. So rationalism about language acquisition is on the right track and empiricism is false.

The argument just given has the following form:

(1) If there are no constraints on language acquisition, then either children have access to negative data or natural languages are unlearnable.[9]

(2) If they exist, the constraints in question must be innate.

(3) Children don't have access to negative data.

(4) Natural languages are learnable.

(5) ∴ There are innate constraints on language acquisition.

One common form of empiricist response to this argument involves challenging (3) (e.g., Prinz 2002; Cowie 1999, 1997). Although it is common to hold that children do not use

[9] Strictly speaking, the results of Gold 1967 allow for a third possibility: the environment is Primitive Recursive and the learner's hypothesis is an effective enumeration of the language. However, there doesn't appear to be any interesting psychological interpretation of this fact.

negative evidence in language acquisition (e.g. Marcus 1993), Prinz and Cowie argue that 'indirect' negative evidence is available to the learner. E.g., Prinz argues that simply failing to encounter a certain (type of) sentence could be sufficient to avoid the learning problem Gold's Theorem presents. Prinz writes:

> Suppose that children, like some recurrent connectionist networks, make predictions about what sentences or words they will hear while listening to adults. A failed prediction could be used as evidence that the rule underlying the prediction was wrong. If learners make predictions of this kind, they have a rich source of negative data without ever being corrected or responsive to correction. (Prinz 2002, 210)

But Gold's Theorem establishes that *all* learners fail to learn classes of languages with the Gold Property, including learners who use indirect negative evidence in the way Prinz suggests. Similar remarks apply to other biases, such as Cowie's 'Prefer simpler hypotheses', etc. quoted above. So attempts to defeat the argument in (1) – (5) by appealing to indirect negative evidence fail.

The previous discussion also establishes an important difference between Gold's Theorem and LPLA. LPLA can be avoided if learners utilize indirect negative evidence along the lines Prinz and Cowie suggest. If a learner hypothesizes a rich language, but does not encounter any sentences of a certain type, she may alter her hypothesis to a simpler language. However, indirect negative evidence is irrelevant to Gold's Theorem. To learn a class of languages with the Gold Property, a learner needs more than indirect negative evidence, such as explicit negative data. A datum of this form can be represented as a pair <*s*, 0>, where *s* is an element of Σ* that is not part of the target language. (Positive data – <*s'*, 1>, where *s'* is part of the target language – is also needed.) Since explicit negative data could also solve LPLA, LPLA is in this respect logically weaker than Gold's Theorem.

In addition to attacking (3), several empiricists have also challenged (2), the claim that the constraints in question must be innate (e.g. Prinz 2002; Cowie 1999; cf. Matthews 2001 for critical discussion). The notion of innateness is nowhere modeled in Gold's Theorem, and Gold's results are silent about whether any such constraints must be innate or not. Since innateness is not part of Gold's Theorem, I won't address this issue.

Despite all the mistaken criticisms of Gold's Theorem, a serious problem remains for the theorem in terms of its psychological utility. The problem concerns the notion of learnability in (1) and (4). So far, we've used the term "learnability" as Gold used it, so that a class of languages C is learnable iff there exists a function ϕ such that for any environment E for any language L in C, ϕ permanently converges onto the hypothesis of L as the target language after some finite time.[10] But there is another option. We could also take learnability to have a psychologically more natural meaning, perhaps along the lines of: a class C of natural languages is learnable iff given almost any normal human child and almost any normal linguistic environment for any language L in C, the child will acquire L (or something sufficiently similar to L) as a native language between the ages of one and five years. (This isn't a theory or conceptual analysis of learnability in psycholinguistics; it's only a ballpark characterization of how the term is used there.). I'll use *acquirable* for the latter psychological notion, and *identifiable (in the limit)* for Gold's notion. Acquirability and identifiability are two very different criteria of learnability, even when children are identified with learning functions and natural languages are identified with sets of sentences. In fact, a primary source of (undue) concern with Gold's Theorem is due to conflating identifiability with acquirability.

A major difference between identifiability and acquirability is in the placement of two restricted quantifiers. Given a target language L in a class C, identifiability requires that for *every* environment, the learner converges to L after *some* finite amount of time. That time can vary wildly from environment to environment. Indeed, there needn't be a finite upper

[10] I assume here that each language has exactly one name.

bound on the time to convergence. It's easy to construct a learnable class of languages such that for every successful learner there is an infinite sequence E_1, E_2, … of environments for L such that the learner first guesses language L on environment E_i no earlier than time t_i. On the other hand, acquirability entails that there exists *some* time after which, given *any* normal environment, a normal child learner will have converged to the correct environment. That is, children always acquire their language within a certain amount of time. Morgan has estimated that a child acquires her language after encountering about 4,280,000 sentences (Morgan 1989, 352). In general, if the relevant notion of learning includes some finite upper bound *n* on the time to convergence, then very few classes of languages will be identifiable in the limit. To see this, let C be a class containing two languages L and L' that contain some elements in common. Now construct a text such that the first *n* sentences are contained in both L and L'. If the learner has converged to L, then complete the text by continuing on with elements of L'; otherwise continue on with elements of L. In either case, the learner fails to identify the target language by the *n*th sentence. Thus, as a model of human language acquisition, identifiability is very crude, and so is not a 'plausible idealization of the learning situation' (Demopoulos 1989, 79) unless the class of acceptable environments is severely restricted.[11] In addition to its psychological crudity, identifiability is hard to compare with acquirability. While identifiability quantifies existentially over all learners, acquirability quantifies (almost) universally over only normal children. So with other things held constant, it is much easier for a class to be identifiable than acquirable. But on the other hand, identifiability quantifies universally over all environments, however odd or repetitive, whereas acquirability quantifies (almost) universally only over the *normal* environments. Thus, acquirability deals with fewer environments than identifiability does, so there is less opportunity for a collection of problematic texts to show up and render a class unacquirable.

[11] Although identifiability is not psychologically natural, it is logically natural. Just as a Turing machine computes a function by using any finite amounts of time and space to arrive at the answer, so too a learner learns by using any finite amount of time to settle into a steady state.

Moreover, acquirability allows the learner to converge not onto L itself, but onto a sufficiently similar language L'. In these respects, it is easier for a class to be acquirable than identifiable (contra the assertions of Pullum and Sholz (2003, 130)). So neither acquirability nor identifiability entails the other.

We can now see the dilemma for interpreting (1) – (5): should learnability be interpreted as identifiability or aquirability? If we use acquirability, then (4) looks true, but no argument whatsoever has been offered for (1). (1) is suggested by Gold's Theorem, but Gold's Theorem is about identifiability, which we've seen is strikingly different from acquirability. If we interpret (1) and (4) in terms of identifiability, then it's unclear why (4) should be true. Just because a class C of (natural) languages is acquirable by children doesn't mean that there couldn't exist a collection of logically possible but highly abnormal environments that would fool all learning functions. Clearly there are logically possible environments that no child could successfully use. In general, the relation of Gold's Theorem to normal child language acquisition is analogous to the relation between Gödel's first incompleteness theorem and the production of calculators. Gödel's theorem show that no accurate calculator can compute every arithmetic truth. But actual calculators don't experience difficulties from this fact, since the unprovable statements are far enough away from normal operations that they don't appear in real life situations. Similarly, child language acquisition may be restricted by Gold's Theorem, but this restriction only applies to cases that don't occur in any normal environment, and thus have no practical significance.

In sum, Gold's Theorem appears interesting to cognitive science when identifiability and acquirability are confused. When we distinguish these notions, we undermine the argument that Gold's Theorem is supposed to support. Early on, we observed three crucial components of a model of learning: (i) the learner's environment, (ii) the nature of the learner, including the set of hypotheses that the learner selects from, and (iii) a criterion of successful learning. Gold's decisions for (i) and (iii) are too liberal to be psychologically interesting. Since these aspects of the model are vitiated, we cannot use it to draw conclusions about the necessity of negative data in language acquisition, (contra Prinz 2002, 210; Pinker 1989, 10; Nowak et al. 2001, 114; cf. Gold 1967, 453 – 454). Nor can we draw the weaker conclusion that 'the space of human languages would have to have some very special properties if they were to be learned only from positive instances of the language' (Williams 1987, ix). In fact, as long as the notion of identifiability in the limit from any environment has no obvious psychological interpretation, there is little of psychological interest to be concluded from Gold's Theorem.

5

Despite its simplicity, many authors have taken Gold's Theorem to threaten some fundamental views about the mind, and they have responded with various criticisms. However, many of these attacks are misguided, for largely formal reasons. But a look at the details shows that Gold's Theorem is still of questionable direct relevance to cognitive science. However, the theorem is still of considerable historical importance. It helped make the psychological community aware of the possibilities for mathematically modeling psychologically relevant aspects of learning. Moreover, it showed that these models can, at least in principle, establish psychologically interesting limitations on possible hypotheses about cognitive activities like language acquisition. In this sense, then, Gold's Theorem provides a useful cautionary tale about the difficulties of precisely articulating a theory of learning – empiricist or otherwise. Thus, despite its limitations, Gold's Theorem belongs on a short list of great results in mathematical modeling in cognitive science.

Appendix

In this section, I characterize the main result of Gold 1967. The interested reader will benefit from a basic grasp of the fundamental aspects of recursion theory (e.g., Shoenfield 1993).

Gold 1967 determined precise boundaries for various learning models. Gold's yardstick for measuring which models could learn which classes of languages was the following sequence of mathematically natural classes of languages, which are ordered by the subset relation:

(A1) *Gold's classes of languages:* Finite $\subset$ Superfinite $\subset$ Regular $\subset$ Context-free $\subset$ Context-sensitive $\subset$ Primitive Recursive $\subset$ Recursive $\subset$ Recursively Enumerable

This sequence begins with the class of all finite languages (i.e., finite subsets of Σ^*, as explained above). Next is a 'superfinite' class of languages, which contains all finite languages plus one infinite language. Then come the classes of regular, context-free, and context sensitive languages. These classes can be identified by the type of rules they allow. Regular languages can be characterized using only rules of the forms $A \rightarrow B$ and $A \rightarrow aB$, where lower-case letters are terminal expressions and uppercase letters are nonterminal expressions; context-free languages need only rules of the form $A \rightarrow \gamma$, where Greek letters stand for either terminal or nonterminal expressions; context-sensitive languages need only rules of the form $\alpha A\beta \rightarrow \alpha\gamma\beta$. The class of natural languages is not contained in the regular languages, and is sometimes thought to be context-free (although Higginbotham (1984) disproves this). Finally, we have the Primitive Recursive, Recursive, and Recursively Enumerable languages. These three classes of languages contain all computable languages. In fact, the class of Recursively Enumerable languages contains languages that are so complex that the best computer program possible can only list the sentences that are in the language, remaining silent about some of those that are not. The subset relations in (A1) are useful because of:

(A2) If $C \subseteq C'$, then if C' is learnable, so is C. Similarly, if C is not learnable, neither is C'.

Ceteris paribus, by shrinking the hypothesis space of possible languages from which the target language could be drawn, the learner is more likely to hone in on the target. But by increasing the hypothesis space, the learner is more likely to miss the target.

Given the typology of language-classes in (A1), Gold explored which classes could be identified in the limit under various circumstances. Gold's general model of learning differed from the model used in the present paper in two ways. First, the learners were required to be computable functions, instead of just any old function whatsoever. Moreover, Gold considered two sorts of learners. The first sort make their guesses about the target language by producing a Turing Machine that computes the characteristic function of the language (a characteristic function for a language L is a function χ such that for any sentence s, $\chi(s) = 1$ if $s \in$ L, and $\chi(s) = 0$ if $s \notin$ L). In Gold's terminology, such a learner uses the *tester naming relation.* The second sort of learner hypothesizes about the target language by producing Turing Machines which, when started, write out a list of all the sentences in the language; these learners use the *generator naming relation*. Due to a well-known result in recursion theory, generators can be constructed from testers. Thus,

(A3) If a learner learns a class of languages with the tester naming relation, then there is a learner who learns it with the generator naming relation.

Gold also considered six different types of learning environments. If a learner only received sentences from the target language, then the environment was called a *text*. (Texts were also constrained to present each sentence of the target language at least once.) Gold considered three ways of presenting a text. The text might be *arbitrary*, and thus be any sequence of sentences from the target language, where each sentence occurs at least once. Or the text might be *Recursive*, and thus meet the additional condition that the text be produced by some Recursive function. Finally, the text might be *Primitive Recursive*, and thus meet the even more stringent condition that it be produced by a Primitive Recursive function. This fact is useful when coupled with:

(A4) A learner learns a language using a class E of environments only if it learns it using any subset of E.

The remaining three types of environment are *informants*, which supply the learner with negative information. At each point in time the informant gives the learner an element of Σ* along with the information whether or not that element is in the target language. Gold considered three different types of informant. Interestingly, they all turn out to be equivalent in the sense that a class of languages is learnable using one form of informant iff it's learnable using another sort of informant. Since they are equivalent, we may think of an informant as providing the learner at each point in time with any sentence at all, along with the information whether it is part of the target language (with the proviso that each sentence in the target language be provided at least once).

The central result of Gold 1967 can be given as follows:

(A5) *Learning with informant:* Using any form of informant and either form of learner (i.e. ones that guess with testers or generators), the class of Primitive Recursive languages is identifiable in the limit, but the class of Recursive languages is not.

(A6) *Learning with text:* Using any form of text and either form of learner *except* the combination of Primitive Recursive text and the generator naming relation, the class of finite languages is identifiable in the limit, but the superfinite class of languages is not.

(A7) *An anomaly:* Using Primitive Recursive text and the generator naming relation, the class of Recursively Enumerable languages is identifiable in the limit.

This classification is organized into a handful of theorems. (GT) has its counterpart in the second clause of (A6). This was proved as Theorem I.8:

> (A8) *Theorem I.8:* Using information presentation by Recursive text and the generator-naming relation, any class of languages which contains all finite languages and at least one infinite language L is not identifiable in the limit. (Gold 1967, 470)

Gold's proof of (A8) is somewhat more complicated than the one given above, because he uses some ideas developed in earlier theorems. By (A4), (A8) establishes non-learnability for any class of texts that contains the Recursive ones. (The proof of (GT) does likewise, if the learner is a Recursively Enumerable function.) Moreover, by (A3), a learner will also fail to learn a superfinite class of languages if it tries to produce a tester instead of a generator for the language. Finally, by (A2), it follows that only the class of finite languages is learnable using Recursive or arbitrary text. (The learnability of the class of finite languages is secured by the following algorithm: at each time, guess that the target language is the smallest language consistent with all the sentences encountered so far. Since a text must present each sentence from the target language at least once, there will be some finite time at which all the target language's sentences will have been presented, and at that time, the learner will correctly guess the target language, and will continue to do so forever more.[12]) It is worth noting, however, that the logic of Gold's Theorem is purely combinatorial: as (GT) shows, there is nothing special about the class of languages being mostly finite or about the names of the languages being Turing Machines. Any sort of names for the language will do, and the class of languages need only have the Gold Property.

The basic strategy for proving Gold's Theorem is again used in Theorem I.9, which shows that a learner receiving Primitive Recursive text and using the tester naming relation

[12] Cf. Osherson, et al. 1986 and Jain et al. 1999 for more study of this issue.

cannot learn a superfinite class of languages. However, the proof is complicated by a clever strategy for generating a Primitive Recursive text that will fool the given learner.

Finally, Angluin proved the following characterization of learnability.

(A9) A class C of Recursively Enumerable languages is identifiable using arbitrary text iff every language L in C has a finite subset T such that for all L’ $\in$ C, if T $\subseteq$ L’ then L’ $\not\subset$ L. (Angluin 1980, 121 – 122)

REFERENCES

Angluin, Dana (1980), “Inductive Inference of Formal Languages from Positive Data”, *Information and Control* 45: 117 – 135.

Bertolo, Stefano (ed.) (2001), *Language Acquisition and Learnability*. Cambridge: CUP.

Chomsky, Noam (1986), *Knowledge of Language*. Westport, Conn.: Praeger.

Cowie, Fiona (1997), “The Logical Problem of Language Acquisition”, *Synthese* 111: 17 – 51.

— 1999, *What’s Within?* Oxford: OUP.

Deacon, Terrence W. (1997), *The Symbolic Species*. New York: W. W. Norton.

Demopoulos, William (1989), “On Applying Learnability Theory to the Rationalism-Empiricism Controversy”, in Robert Matthews and William Demopoulos (eds.) *Learnability and Linguistic Theory*. Dordrecht: Kluwer, 77 – 88.

Elman, Jeffrey, Elizabeth Bates, Mark Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett (1996), *Rethinking Innateness*. Cambridge, MA: MIT Press.

Feldman, Jerome (1972), “Some Decidability Results on Grammatical Inference and Complexity”, *Information and Control* 20: 244 – 262.

Gibson, Edward, and Kenneth Wexler (1994), “Triggers”, *Linguistic Inquiry*, 25: 407 - 454.

Gold, E. Mark (1965), “Limiting Recursion”, *The Journal of Symbolic Logic* 30: 28 – 48.

— (1967), “Language Identification in the Limit”, *Information and Control* 10: 447 – 474.

Higginbotham, James (1984), “English is Not a Context-Free Language”, *Linguistic Inquiry* 15: 225 – 234.

Hirsh-Pasek, Kathy, and Roberta Michnick Golinkoff (1996), *The Origins of Grammar: Evidence from Early Language Comprehension*. Cambridge, MA: MIT Press.

Howe, Christine J (1993), *Language Learning: a Special Case for Developmental Psychology?* Hillsdale, NJ: Lawrence Erlbaum.

Jain, Sanjay, Daniel Osherson, James S. Royer and Arun Kumar Sharma (1999), *Systems That Learn (2^{nd} ed).* Cambridge, MA: MIT Press.

Kelly, Kevin (1996), *The Logic of Reliable Inquiry*. Oxford: OUP.

Marcus, Gary (1993), “Negative Evidence in Language Acquisition”, *Cognition* 46: 53 – 85.

Matthews, Robert J. (1984), “The Plausibility of Rationalism”, *The Journal of Philosophy* 81: 492 – 515.

— (2001), “Cowie”s Anti-Nativism”, *Mind and Language* 16: 215 – 230.

Morgan, J. L. (1989), “Learnability Considerations and the Nature of Trigger Experiences in Language Acquisition”, *Behavioral and Brain Sciences*. 12: 352 – 353.

Niyogi, Partha, and Robert C. Berwick (1996), “A Language Learning Model for Finite Parameter Spaces”, *Cognition* 61: 161 – 193.

Nowak, Martin A, and Natalia Komarova, and Partha Niyogi (2001), “Evolution of Universal Grammar”, *Science* 291: 114 – 118.

Osherson, Daniel, Michael Stob, and Scott Weinstein (1984), “Learning Theory and Natural Language”, *Cognition* 17: 1 – 28.

— (1986), *Systems That Learn.* Cambridge, MA: MIT Press.

Pinker, Steven (1989), *Learnability and Cognition*. Cambridge, MA: MIT Press.

Prinz, Jesse (2002), *Furnishing the Mind*. Cambridge, MA: MIT Press.

Pullum, Geoffrey K. and Barbara C. Scholz (2003), “Linguistic Models”, in Marie T. Banich and Molly Mack (eds.) *Mind, Brain, and Language*. Mahwah, NJ: Lawrence Erlbaum Associates, 113 – 141.

Shoenfield, Joseph (1993), *Recursion Theory*. New York: Springer-Verlag.

Valiant, L. G. (1984), “A Theory of the Learnable”, *Communications of the ACM* 27: 1134 – 1142.

Wexler, Kenneth and Peter W. Culicover (1980), *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

Williams, Edwin (1987), “Introduction”, in T. Roeper and E. Williams (eds.) *Parameter Setting.* Dordrecht: Kluwer, vii – xix.