Edinburgh Occasional Papers in Linguistics September 22, 1997

The evolution of incremental learning: language, development and critical periods

Simon Kirby & James R Hurford
simon@ling.ed.ac.uk

1 Introduction

In this article[1], we show how the existence of critical periods follows from the action of natural selection on genomes in which incremental growth can be tuned to chronological age (maturation) or to accumulating input.[2] This article brings together conclusions from two previous papers in *Cognition* (Hurford 1991, Elman 1993) which have been thought to be incompatible.

The key concepts in our discussion are:

1. Incremental growth in cognitive resources, as facilitating language acquisition
2. The nature of the genome's control over such incremental growth – whether growth is a function of maturation or of exposure to data
3. Darwinian natural selection
4. Critical periods in language acquisition

The last of these is our *explanandum*; we will locate these concepts in an explanatory framework, and outline a mechanism, computationally implemented, relating them to each other. Our discussion is concentrated on language acquisition, but our conclusions can be applied to development more generally, so in principle "language acquisition" above can be replaced simply by "development" (*mutatis mutandis*).

In the next section (Sec.2), we briefly review the evidence for critical (or sensitive) periods in language acquisition; in Section 3, we describe the principal structural properties of language acquisition models that rely on the idea of incremental growth, with specific focus on Newport's "Less is More" hypothesis and Elman's "Importance of Starting Small"; in Section 4, we identify two elements crucially missing from such models, **timing** and **evolution**; the remaining sections of the paper set out our own explanatory framework, and report the results of computational simulations.

2 Critical periods

Long (1990) has provided a valuable survey of the evidence for critical periods in language acquisition. He draws five conclusions:

[1]This work was supported by two fellowships at the Collegium Budapest Institute for Advanced Study, and by Economic and Social Research Council research grant R000326551. We also thank Mark Ellison, Jenny Hey and Kevin Gregg for helpful comments: any defects that remain are the responsibility of one of the authors

[2]This is exactly the type of investigation identified by Todd (1996:217) as a fruitful interface between developmental psychology and evolutionary simulation.

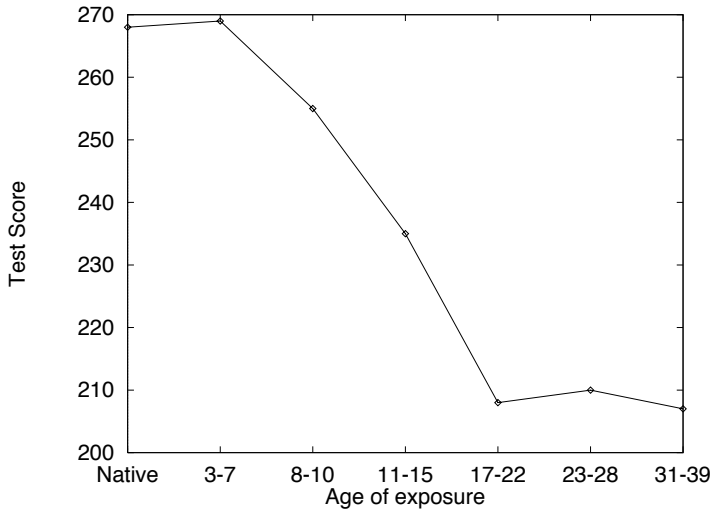

Figure 1: How language competence declines with age of initial exposure to a language (graph taken from Goldowsky & Newport 1993).

> (a) "Both the initial rate of acquisition and the ultimate level of attainment depend in part on the age at which learning begins.
> (b) There are sensitive periods governing language development, first or second, during which the acquisition of different linguistic abilities is successful and after which it is irregular and incomplete.
> (c) The age-related loss in ability is cumulative (not a catastrophic one-time event), affecting first one linguistic domain and then another, and is not limited to phonology.
> (d) The deterioration in some individuals begins as early as age 6 — not at puberty as is often claimed.
> (e) Affective, input, and current cognitive explanations for the reduced ability are inadequate." (251)

Long's article appeared in a journal of second language acquisition, but it deals with both first and second language acquisition, and his conclusions apply equally to both. The evidence for a critical period in first language acquisition has mounted considerably since Lenneberg's original claim (Lenneberg 1967). It includes evidence from feral children and child abuse cases (Curtiss 1980; Curtiss 1977; Goldin-Meadow 1982), and from deaf children's and adults' acquisition of ASL (Woodward 1973; Mayberry *et al.* 1983; Curtiss 1988; Newport 1984; Newport & Supalla 1992). Figure 1 gives a visual impression of the critical period for second language learning.

Johnson & Newport (1989) distinguish between two versions of the critical period hypothesis:

> "*Version One: The exercise hypothesis.* Early in life, humans have a superior capacity for acquiring languages. If the capacity is not exercised during this time, it will disappear or decline with maturation. If the capacity is exercised, however, further language learning abilities will remain intact throughout life.
>
> *Version Two: The maturational state hypothesis.* Early in life, humans have a superior capacity for acquiring languages. This capacity disappears of declines with maturation."

(64)

As both Long and Johnson and Newport point out, these two versions make the same prediction for first language acquisition, but a different prediction for second language acquisition. Long argues that the literature supports a radical form of the maturational state hypothesis. The evolutionary simulations which we report in later sections give rise to a (simulated) developmental program which is consistent with the maturational state hypothesis, but not with the exercise hypothesis.

No significant new data on the critical/sensitive period has emerged since Long's (1990) review[3], and we will proceed on the assumption that his summary above is essentially correct. As he notes, "Here, ... disagreements as to both the facts and their explanation are very pronounced" (252). Some of the sting may be removed from any controversy by reiterating further wise remarks by Long, who draws here on Oyama (1979)[4].

> "Saying something is "biological" or "maturational" does indeed often imply reliably scheduled sequences, changes in anatomical structure and size, and apparent independence from specific environmental contingencies. The terms need not be so narrowly construed, however ...In sum, while maturational constraints are certainly compatible with nativist accounts of learning, they do not entail such views. Oyama suggested that a sensitive period is more usually thought of as the product of a *nature-nurture interaction*, a time of heightened responsiveness to certain kinds of environmental stimuli, bounded on both sides by states of lesser responsiveness." (252-3)

The outcomes of our evolutionary simulations are exactly in keeping with these suggestions of a nature-nurture interaction. Recently, a somewhat polarised antithesis to nativism in language acquisition has been proposed by "constructivists", such as Quartz & Sejnowski (1997). In the constructivist view, growth, rather than being innately pre-programmed, is responsive to input stimuli. This claim applies equally to the growth of physical features (such as neural dendrites) and to the growth of the abstract representations (e.g. of grammar) instantiated in them. But the constructivist's "rather than" poses a false opposition; in a quite real sense, obviously, the developing organism is pre-programmed for growth in response to stimuli.

Any innate pre-programming for input-sensitive growth must have evolved. An evolutionary model necessarily incorporates the key ingredient of variation among individuals; in order for a species to evolve, there must be variation among its members. This allows us a further slight softening of an unduly strict interpretation of critical period claims. We interpret the critical period hypothesis as making a strong statistical prediction, that the vast majority of humans are biologically disposed to lose their language acquisition ability in the early years of life. Evidence that isolated exceptions to the broad trend exist, such as the subjects of White & Genesee (1996) or Birdsong (1992), do not invalidate the significant statistical claim. In a similar vein, the commonplace claim that all normal humans acquire full competence in a language is not invalidated by the relatively rare instances of severe congenital language deficit. (But we believe that making a statistical claim about all humans is less circular than making an absolute claim about all 'normal' (undefined) humans.)

[3]We do not interpret White & Genesee (1996) or Birdsong (1992) as constituting significant attacks on the critical period hypothesis.

[4]Oyama's ideas on nature-nurture interaction are further developed in Oyama (1985).

3 Incremental Learning

We associate the term "incremental learning" with the idea of some learning-related resource starting at a low value, which then gradually increases while (but not necessarily because) the organism matures. Also essential to incremental learning is the proposition that the initial low (immature) value of the resource actually facilitates, or even enables, the early stages of learning. Later stages of learning are in turn facilitated, or enabled, by higher-valued settings of the resource concerned. We shall mention some non-linguistic instances before discussing some specifically linguistic proposals, by Elman and Newport.

Turkewitz & Kenny (1982) pioneered a theoretical position, backed by neurological and developmental evidence across species, that the limitations (or immaturity) of sensory and motor systems may play adaptive roles in ontogeny.

> "In recent years, it has become abundantly clear that James's (1890:499) characterization of the world of the infant as a "blooming buzzing confusion" is simply wrong. There is evidence that the infant's world is structured and that far from being overwhelmed by a barrage of stimulation which only slowly comes to be sorted out, the infant from his earliest days is quite properly characterized as competent and organized. It is our contention that one of the major sources for this organization is the infant's limited sensory capacity." (362)

We cite below some of Turkewitz and Kenny's most telling examples.

> "This limitation includes a fixed [infantile visual, K&H] accommodative system in which objects that are approximately 10 in. away are most clearly in focus (Haynes *et al.* 1965), and an acuity level such that only relatively large objects or large features (Salapatek & Banks 1978) are resolvable. ... the infant's responsiveness to visual stimuli with only low spatial frequencies ensures that large segments of the external world will not be resolved by the infant's visual system, further reducing the amount of visual information available for processing (Salapatek & Banks 1978)." (362)

> "... if vision is effectively restricted to objects within a narrowly circumscribed distance from the viewer, the requirement for size constancy is obviated and an orderly world is attained even in the absence of a level of perceptual organization necessary for the achievement of size constancy. Thus the infant's limited depth of field may make it possible to respond to and learn the relative size of objects even in the absence of size constancy. In that known size is a strong cue for size constancy (Ittelson 1951), the early opportunity to learn relative sizes provided by the infant's limited depth of field may facilitate the development of size constancy. That such is the case is suggested by the recent finding that size constancy, when it appears (at between 4 and 6 months), is initially restricted to very near distances (probably not greater than 70 cm.) (McKenzie *et al.* 1980)." (363)

> "Finally there is some evidence suggesting that during normal development in infants and young children, the availability of multimodal input may disrupt rather than enhance functioning. Thus, Rose *et al.* (1978) report that when 6-month-old infants are allowed to simultaneously see and manipulate objects, there is no evidence that they subsequently recognize the object visually, although such recognition is clearly evidenced when they are given only visual preexposure. ... Renshaw *et al.* (1930) report that the ability of young children to localize a touched spot on their body surface is interfered if they are allowed to use vision during localization. The same procedure results in improved localization in older children and adults." (365)

Bjorklund & Green (1992) survey a range of areas for which there is evidence that

> "...aspects of children's immature thinking are adaptive in their own right. We propose, as have others (e.g. Lenneberg 1967; Oppenheim 1981), that some aspects of the young child's cognitive system are qualitatively different from those of the older child or adult and are well suited to attain important cognitive-social milestones such as attachment or language. In a similar vein, Oppenheim discussed the presence of neurobehavioral characteristics of immature animals that have a specific role in survival during infancy or youth but disappear when they are no longer necessary. These *ontogenetic adaptations* are not simply incomplete versions of adult characteristics but serve specific adaptive functions for the developing animal." (46)

The areas surveyed by Bjorklund and Green include metacognition, plasticity and speed of cognitive processing, egocentricity, and language acquisition. Under the heading 'metacognition', they point out that younger children tend to have optimistically unrealistic views of their own capacities and achievements, as compared to older children and adults. Such limited and immature self-evaluation gives young children the confidence they need, B&G argue, to persevere in engaging in challenging activities. If young children really knew at the outset how difficult and complex life's challenges were to be, they would perhaps never embark on the journey toward ultimate mastery.

Under the heading of 'plasticity and speed of cognitive processing', Bjorklund and Green write:

> "Slow and inefficient processing through infancy and early childhood may be the factor responsible for the intellectual plasticity observed in humans. Because mental operations are slow, less information is activated and processed automatically. This reduced automaticity makes processing more laborious and ineffective for the young child, but at the same time protects the child from acquiring cognitive patterns early in life that may not be advantageous later on. Because little in the way of cognitive processing can be automatized early, presumably because of children's incomplete myelinization, children are better prepared to adapt cognitively to later environments. ...Cognitive flexibility in the species is maintained by an immature nervous system that gradually permits the automatization of more mental operations ..." (49-50)

Under the heading of 'egocentricity', Bjorklund and Green argue (referring to a study by Mood 1979) that the relative inability of young children to take a perspective other than their own is adaptive in that it facilitates early sentence comprehension. A better way of looking at this is to see young children's better performance on sentences related to their own perspectives as evidence of the children's *incremental* semantic/pragmatic progress. The child's progress into the semantic space of possible sentence-interpretations necessarily starts with practice on interpretations involving concepts familiar from their own infantile experience, or perhaps even innate concepts.

3.1 Less is More

Goldowsky & Newport (1993) describe a computer model of the acquisition of form-meaning pairings. It is assumed that the child can segment out from the stream of speech whatever atomic meaningful forms (morphemes) it contains. (This is no trivial matter, but it seems that the child must achieve it.) It is also assumed that the child can understand enough of the content of what is being said, from contextual clues, to form some kind of semantic structure, containing (linearly unordered) semantic primitives (concepts, or whatever). Adults do

not package their utterances in such a way that the first morpheme necessarily corresponds to the "first" semantic prime in any ordered representation of its meaning. All the child gets (at most) is a sequence of forms, which she knows somehow corresponds to an unordered set of meanings. The acquisition task is to figure out which atomic forms correspond to which atomic meanings. The task is further complicated by the possibility of several morphemes in some utterances, perhaps redundantly, corresponding to a single meaning, and the possibility of some meanings present in the child's interpretation of the situation not being expressed by any of the forms present in an observed utterance.

Goldowsky and Newport devise a simple artificial language, or code, in which atomic meaning-form pairs are pre-defined, and compose multi-form utterances from this code. They then input such utterances, as sequences of forms paired with unordered sets of meanings, to a learning program. This program takes in meaning-form pairs, as extracted in all logically possible ways from the simulated utterances, and builds a table of their correspondences. The resulting table contains many spurious atomic meaning-form pairs, i.e. pairs not envisaged in the artificial code from which the input utterances were generated.

G&N discover an interesting and suggestive way in which the number of such spurious pairings can be reduced, while not losing the genuine pairings. They impose a random filter on the input, such that the input stream is effectively broken into shorter sequences, and the corresponding semantic input is also fragmented. They appropriately call the resulting effect "data loss", and comment:

> "However, the data are not lost evenly: the upper left corner [of their table], containing the one-to-one mappings, retains more data than the rest of the table, since only small pieces of form and meaning can make it through the filter. Thus the model is forced to concentrate on smaller units, much as the child does. This effect we call DATA FOCUS." (131-2)

G&N's general conclusion is:

> "We have shown that a limitation on the ability to perceive or remember the full complexity of linguistic input, as seems to occur in young children, may have unexpected benefits for the learning of morphology. If the child begins acquisition with a very restricted input filter, it will obtain the optimally clean data for the smallest meaningful units in the language. Learning larger units will require a less restrictive filter, but as we mentioned earlier, for *any* structure in the language there is a filter that produces optimal learning of that structure. If you start with very limited capabilities and then mature, you will have each size of filter in turn, and therefore have the chance to learn each structure in the language at the time appropriate for that structure — and you end up learning the entire language optimally." (134)

3.2 The Importance of Starting Small

In a much-cited paper, Elman (1993) describes experiments which show that acquisition of a small but naturalistic context-free language is significantly facilitated by arranging the acquisition device (a recurrent neural net) in such a way that its "working memory" is small at the outset of learning, and grows incrementally during the learning process.

> "The networks are trained to process complex sentences involving relative clauses, number agreement and several types of verb argument structure. Training fails in the case of networks which are fully formed and 'adultlike' in their capacity. Training succeeds only when networks begin with limited working memory and gradually 'mature' to the adult state. This result suggests that rather than being a limitation, developmental restrictions

> on resources may constitute a necessary prerequisite for mastering certain complex domains. Specifically, successful learning may depend on starting small." (71)

That quotation says it all; as Elman's paper is relatively well-known, we will not present a detailed account of the workings of his system, but will restrict ourselves to some points which emerge from a retrospective view taken several years after his paper appeared.

The general relationship between learning long-term temporal dependencies and various degrees of embedded memory in a variety of recurrent neural network architectures is explored in Lin *et al.* (1996). Not surprisingly, architectures with greater 'memory', defined on neural nets as the recycling in various ways of information about previous states of the system, are superior at learning long term dependency tasks than architectures with lower degrees of 'memory'. But the long-term dependency tasks investigated by Lin et al. were much simpler than the learning task that Elman set his system. In Lin et al.'s trials, some instance of long-term dependency was typically the only piece of structural knowledge to be learnt; that is, the training sets contained examples illustrating long-term dependencies and no (or little) other structure. Elman's target knowledge, on the other hand was more naturalistic, and contained elements of structure, such as Noun/Verb categorisation, over and above long-term dependencies, and in terms of which the long-term dependencies were themselves defined.

> "When . . . the network is initially handicapped, . . . the *effective* . . . subset of data . . . contain only three of the four sources of variance (grammatical category, number and verb argument type) and there are no long-distance dependencies. . . . by selectively focusing on the simpler set of facts, the network appears to learn the basic distinction — noun/verb/relative pronoun, singular/plural, etc. — which form the necessary basis for learning the more difficult set of facts which arise with complex sentences." (Elman 1993:84)

In other words, while architectures with constant built-in large memory are clearly better at learning long-distance dependencies, this appears only to be true where the long-distance dependencies are not located in the later stages of an incremental learning scheme. For target knowledge which is incremental in nature, with long-term dependencies located in the 'higher reaches', Elman's conclusion is that constant adult-sized memory is a hindrance, rather than a help.

An important caveat concerns Elman's term 'working memory'. In Elman's system, incrementing the 'size' of 'working memory' involved cutting the link between the context layer of his system and the hidden layer at successively longer intervals. The context layer can be regarded as a register of the internal 'cognitive state' of the machine after processing (parsing) the last several words of input; cutting the link between the context layer and the hidden layer every N words has an effect which one can reasonably surmise is like that of restricting the parsing process to a window or buffer of length N words. Although the term 'working memory' is a suggestive mnemonic for this variable in Elman's model, whose 'size' was incremented during training, it cannot be equated with the working memory which is the subject of extensive theorising in the psychological literature. The principal theorists of working memory are Baddeley and Gathercole (see Baddeley 1986; Baddeley 1990; Baddeley 1992; Baddeley *et al.* 1988; Baddeley *et al.* 1995; Gathercole & Baddeley 1990; Gathercole & Baddeley 1993). The working memory model discussed in psychological literature has several components, including a 'phonological loop', a 'visual sketch-pad', and a 'central executive'. The phonological loop component, as its name suggests, is a kind of buffer in which specifically phonological (or even raw phonetic) information is stored, as an utterance is processed. It is established that the size of this buffer, or something like it, is smaller

in children than in adults, and thus might seem to fit Elman's model. But Elman's model in fact has no phonology, and the part of his system which he labelled 'working memory' is not a buffer containing elements from the input signal, but rather a register of the internal 'cognitive state' of the machine after processing (parsing) the last several words of input. The 'visual sketch-pad' component of working memory models is also obviously not identifiable with Elman's 'working memory'; and there is much less theorising about the nature of the central executive component, so that one cannot make an easy identification with that module, either.

The above discussion is, however, merely terminological. We do not doubt that Elman's system demonstrates a robust result that the gradual incrementation of a particular resource component of a language acquisition device facilitates (or even enables) acquisition of a naturalistic grammatical system. The question of a biologically plausible interpretation of the general 'starting small' idea is taken up next.

The most extensive testing (that we are aware of) of Elman's conclusions is by Joyce (1996). Joyce first replicates Elman's own experiments, as closely as possible given the absence of certain technical details from Elman's account. The replications broadly confirm Elman's own conclusions about the advantages of starting small.

Next, Joyce argues that Elman's method of periodically annulling the impact of the information in the context layer is biologically implausible, and he considers alternative ways of making a recurrent neural net 'start small'. The method with which he experiments in detail is that of applying internal noise to the various channels of information-flow in the net. The equivalent of 'starting small' is now 'starting with high internal noise': this is not a great conceptual leap, as high noise systems will tend to transmit information about local correlations in data relatively faithfully, but are far less likely to transmit accurate information about longer-distance correlations. (Another method of 'starting small', progressively adding nodes to the network during learning, in keeping with Quartz & Sejnowski's (1997) constructivist ideas, is mentioned but not implemented.)

Joyce experimented with the injection of noise at various points in Elman's architecture, namely (a) at the input layer, (b) globally, at the determination of the error to be backpropagated ("appropriate distribution of a noisy error"), (c) in the transmission of the error information during backpropagation ("noisy distribution of the correct error"), (d) at 'synapses', on the weights of connections, and (e) in the (non-)functioning of individual connections. Under all conditions, the level of noise was gradually reduced, as in an annealing regime. The hypothesis tested was whether, in any of these conditions, 'starting noisy' and gradually diminishing the noise internal to the system, would yield results analogous to Elman's results with 'starting small'. The first four methods of injecting noise failed to produce results analogous to Elman's.

> "Four of the five simulations failed to satisfactorily encode the grammar used to generate sentences. Using noise on inputs, synapses, global error and backerror gave results comparable to Elman's first experiment when no modifications were made to the training regime." (58)

In Joyce's fifth condition,

> "...a noise factor was used to determine the probability of a given connection not functioning. The noise factor was pre-set and decayed in the way specified above. ...all synapses which were active in the forward pass are updated in the normal way, whereas inactive synapses remain constant. Reward/blame for the current error is only distributed across the active connections. It can be seen that here we have an example of a changing learner but constant problem. For each pattern there is a new network with the

> same architecture but different connectivity. As learning progresses the average number of connections for the network increases. In this network it is the degree of connectivity which is at first small and then increments." (57-58)

Most interestingly,

> "For the last simulation — noise on transmission, an entirely different pattern of results were found. For all regimes, average error score approached those obtained by Elman in his two successful simulations." (60)

Joyce concludes,

> "Since this simulation attained similar levels of performance on the task used by Elman in his '93 study it both strengthens the hypothesis that limited resources can often be of benefit to a developing learner, but at the same time makes the particular implementation Elman used less suggestive about which particular resource is in fact limited. The use of transmission noise provides a more biologically plausible learning system in terms of both the actual processes incorporated in the learning system and in terms of the behaviour of the system whilst learning (no discrete cut-off for short term memory)." (63)

3.3 Summary

To summarise this section, we will begin by specifying what is common to various models of incremental learning. The rather formal specification below is our own, intended to make precise the actual claims embodied in the idea of incremental learning, and to provide a basis for the computational simulations described later.

- A resource, R, of variable size, up to some maximum m; the available sizes can be expressed as R_1, R_2, $\ldots \mathrm{R}_m$. Models of incremental learning assume that this resource (somehow) increases from the minimal to the maximal value during the learning process.
- A complex, but finite, body of knowledge, K, to be acquired, whose maximum size is calibrated at the same notional number m; incomplete fractions of this body of knowledge can be expressed as $\frac{1}{m}K$, $\frac{2}{m}K$, $\ldots \frac{m-1}{m}K$.
- At least one ordered set, O, of particular fractions of K, ordered in increasing size, each a superset of the previous fraction; this represents the natural order of acquisition of elements of K . The stipulation "at least one" is to recognise that there may be (perhaps trivially) alternative natural orders of acquisition.
- An assumed correlation of the various possible sizes of R with fractions of K, such that possession of exactly R_i and of $\frac{i-1}{m}K$ is a necessary condition for further learning, progressing to $\frac{i}{m}K$.
- A body of data, D, from which K can be learnt, and to which the organism is exposed.

Spurious mathematics needs to be avoided. If terms such as "size" and "fraction" here are to be related to any empirically observable or measurable phenomena, they will almost certainly need to be interpreted in an ordinal, rather than a strictly quantitative sense. That is, for example, R_1 is measurably less than R_2, but not necessarily exactly half of it, as the

numbers 1 and 2 might suggest; and $\frac{2}{m}K$ is verifiably a subset of $\frac{4}{m}K$, but not necessarily exactly half of it, as the numbers 2 and 4 might suggest.

For a model such as Elman's, the resource R is the size of "working memory", that is the inverse of the frequency with which the link between the hidden layer and the context layer is cut – every word, every 2 words, every 3, and so on, up to to some maximum. The body of knowledge K is the set of pairings of initial substrings of grammatical sentences (up to some limit on sentence length) with their appropriate continuations (next word or end-of-sentence). Incomplete possession of K would constitute knowledge of some, but not all, of these appropriate continuations. In Elman's model, there is a natural order of acquisition such that, for example, grammaticality facts due to the frequent juxtaposition of basic syntactic categories are acquired before more sophisticated knowledge, such as long-distance dependencies. At an incomplete stage of learning, only a subset of the appropriate continuations of any particular initial string might be predictable with any confidence. For Elman, the data D was a stream of grammatical sentences generated by his target grammar.

For a model such as Goldowsky and Newport's, R is the size of the holes in the filter from input to intake. The filter may admit an average of only 25% of the input stream of words, or 50%, or 75% up to 100% . K is the set of correct pairings between forms and their meanings. For Goldowsky and Newport, the natural order of acquisition progresses from one-to-one mappings between frequent atomic forms and their meanings to more complex high-level pairings of strings of forms onto complexes of meanings. Irregular form-meaning pairings are also acquired later. The data, D, was a stream of words as defined by the target morphological system.

The specifications above mention subsets and supersets, and it is important to clarify the sense in which we use such terms. What the child acquires is *information* about her language-to-be. Such information can come in various forms, such as parameter settings, or lexical items, for example. The logic of the relation between grammars and languages means that *more* information about what is grammatically possible can result in *fewer* sentences being available to the child. The incrementation during language acquisition that we are concerned with is not any superficial measure of the language itself, as, say, in the number of sentences that the child can produce, but rather the *amount of information* about her language that the child controls.

Say, as a temporary simplifying assumption, that the child's vocabulary throughout the acquisition period stays constant. And say the child starts with no parameter settings. Initially, then, the child's initial grammar, G_0, generates all strings over its vocabulary that are permitted by universal principles. Call the set of such strings L_0. Setting the first parameter results in a more restrictive grammar, G_1, which generates L_1, a subset of L_0. If there are n parameters, grammatical acquisition is complete with the setting of the n^{th} parameter, and the child at that stage will have gone through a succession of grammars $G_0, G_1, G_2, \ldots, G_{n-1}, G_n$ and the corresponding sequence of successively more restricted languages. Acquiring more information (in the shape of parameter settings) progressively narrows down the number of objects (sentences) specified as grammatical in the mind of the acquirer[5]. In this sense, grammar acquisition is a matter of *contraction*, not expansion.

[5]Another assumption is possible about parameter settings, namely the "Subset Principle" Wexler & Manzini (1987), that their default settings generate subsets of the language generated by their other, marked, settings. Here, the detailed relationship between parameter settings and the number of sentences in the language is different from that described above, but still there is no decrease in the amount of information possessed by the child, and an increase in information induced from experience

Obviously, however, a child's language *expands* during acquisition, due to increases in vocabulary and in the length of sentence that can be managed. Vocabulary and principles/parameters are different systems. Adding a vocabulary item increases the set of sentences made available by the child's grammar. We can separate the two systems conceptually, for the purposes of characterising incremental learning, by talking about the acquisition of grammatical parameter settings as if grammars specified strings of "pre-terminal symbols", syntactic categories such as Noun, Verb, Determiner, and the like, rather than strings of actual words or morphemes. Conceivably, the inventory of such pre-terminal categories is universal and is not incremented during language acquisition; new vocabulary items are just attached to various members of this set of categories.

Given the separation of grammatical parameters from vocabulary, we can characterise what would count as an incrementally acquired system in each case. For vocabulary, an incrementally acquired system is one in which there is a natural order to vocabulary acquisition, with the acquisition of items of a certain type being a prerequisite for the acquisition of items of other types. Clearly, much vocabulary acquisition is ordered in this way, with many abstract terms (e.g. *ambition, influence*) only being acquirable once certain other, concrete, terms (e.g. *person, body*) have been acquired. For grammar as specified by parameter settings, an incrementally acquired system is one in which there is a natural order in which (some of) the parameters are set. There is also an interaction between the incremental properties of different subsystems, such as grammar and vocabulary. Probably, no parameter involving some particular syntactic category (or class of categories) can be set until the child has acquired at least one lexical item instantiating that category (or class).

The idea of incremental learning goes beyond Chomsky's (1981:10) requirement of 'epistemological priority'. A theory of language learnability can maintain the epistemological priority of certain foundational components, without abandoning the idealisation of instantaneous acquisition. Incremental learning, on the other hand, entails non-instantaneous acquisition, and involves a scale, or ladder, of epistemological priority; some late-acquired knowledge is epistemologically dependent on earlier-acquired knowledge.

Such an incremental process in syntax acquisition is central to Berwick's (1985) acquisition model. Berwick defines a relation of 'acquisition dependence' between grammar rules, and between sentences; a further notion of 'learning sequence' is derived from his acquisition procedure and the acquisition dependence relation. A general similarity between Berwick's work and Elman's (*mutatis mutandis*) can be read into Berwick's statement, "The acquisition procedure incorporates a filtering function f such that the resulting learning sequences are well-ordered" (175). Thus, for both Elman and Berwick, something internal to the organism imposes an effective ordering on the input data. For Berwick, this ordering derives from the logical structure of the knowledge being acquired; for Elman, the order-imposing component is not inherent in the knowledge being acquired, but independent of it. In this paper, we pursue the question of how such an adaptive order-imposing mechanism could have arisen.

Generalising considerably from Bjorklund & Green's (1992:48) suggestion that egocentricity in children facilitates sentence comprehension, we would argue that the child could not be 'born mature', i.e. already capable of interpreting sentences with quite abstract, impersonal or non-egocentric meanings, although it must be born capable of progressing to an understanding of such meanings. An essential feature of abstract, impersonal or non-egocentric meanings is that they are in some sense projected from more concrete, personal and egocentric meanings. If one were somehow 'born mature', then all concepts, abstract

and concrete, personal and impersonal, ego-oriented and other-oriented, would be on an equal footing, missing the inherently derivative nature of the latter kind of concepts. To take a related example, children are capable of acquiring a 'theory of mind'; that is, normal children end up with an awareness that other people think rather like them. The growth of this awareness must progress first through a stage of *self*-awareness, a theory of one's own mind, or else the final step, the progression to an explanation of others' behaviour in terms of one's own mind, would not be possible.

To put the matter in terms which are even more general (but still contentful — and possibly contentious), we assume an *incremental epistemology*, such as is implicit in a wide range of philosophical, linguistic and psychological ideas. The ideas we have in mind include Quine's 'beginning with ordinary things' (Quine 1960:1–5) and his 'semantic ascent' (270-276), Fodor's (1981) 'triggering hierarchy', and Lakoff's 'internal realism' and 'experiential realism' (Lakoff 1987:§16). The body of knowledge which a human acquires, mainly during childhood, is structured in a way reflecting how knowledge acquired later is built upon prior knowledge. And, further, we claim, the child's devices for knowledge-acquisition are adaptively structured, by such strategies as 'starting small' and 'less is more', to acquire the earlier knowledge first, without distraction by evidence relevant only to the kind of knowledge acquired later.

Incremental learning should be understood against the general background of pronounced human morphological neoteny (see Hofer 1981; Wesson 1993). Morphologically, humans change much less between infancy and adulthood than do chimpanzees. Behaviourally and cognitively, the situation is reversed — the difference between human infant and adult cognition is enormous. Morphological neoteny, along with the longer proportion of life spent maturing, serves to maximise the period during which the organism is plastic and malleable. But this long period of plasticity and malleability is not uniform — it is incrementally structured, with different, or more powerful, parts of the learning organs coming on-line in sequence, corresponding to the incremental, or layered, structure of the acquired adult cognitive state.

4 Gaps in incremental learning accounts

Clearly, the insights that less can be more, and that starting small is important, are valuable for understanding language acquisition. More generally, the paradigm of incremental learning, which we have begun to make precise above, promises valuable insights in a range of domains of development. But two (at least) obvious questions remain unanswered; these are the *WHEN?* and the *WHY?* questions. We address these questions below.

4.1 When? The scheduling problem

Conceivably, a person could either be in a great hurry to learn her language, or take a leisurely lifetime to achieve it. The incremental learning paradigm itself says nothing about this kind of timing. Incremental learning, as we have specified it, leads to the emergence of critical periods. But explanations of critical period effects (e.g. Elman 1993; Bever 1981) typically make no mention of the facts of timing. This is a major omission, as the existence of a critical period is *essentially* a fact about the timing of some period of openness to learning in an individual's life history. According to the evidence summarised by Long (1990),

roughly 15 is "the close of the latest posited sensitive period for language development, that for morphology and syntax" (Long 1990:279). Now, why 15? Why not 30? Why not 55? Indeed, why not 2? A full explanation for a critical period should also explain the particular age range affected.

An earlier work (Hurford 1991) proposed a model which identified puberty as the closing age of the critical period for language acquisition. Long's survey makes it clear that in fact puberty is not very closely identified with the closing age of the critical period. The model to be described below gives results in which the ending of the critical period does not coincide exactly with puberty, but nevertheless is indirectly correlated with it, via a stage in the life-history of a typical individual that we call the "*normal* age of acquisition".

In the experiments of Elman and Goldowsky and Newport, the experimenters manipulated the timing of the increase in the relevant resource in such a way as to enable successful learning. The actual schedules used were never defined in terms of any analogue of chronological age. The trained systems were not living organisms, inexorably maturing or ageing as the seconds tick by. Systems like these could be left for any length of time with no input, and not change their internal states in any way. This explains why Elman, for example, was in no position to make any statements about the timing of the critical period, relative to life history.

4.2 Why? Evolutionary adaptation

In typical incremental learning experiments, discovering the most successful schedule for increase in the resource is the outcome of trial and error. Elman, for example, found that he had to train his system on many times more sentences while its "working memory" was at the initial low value than at later stages in his simulations. But what ensures that the timing of the increase in resource in a real living organism is such as to yield successful learning? Living organisms in real populations are not marionettes whose internal resources are scheduled by experimenters. We propose that the scheduling found in real organisms is also the outcome of a process of trial and error, but blind non-teleological trial and error — Darwinian natural selection. We will work with the assumptions that successful learning of language conveys selective advantage on individuals, that the scheduling of the increase in resource is genetically controlled, and that the details of this genetic control are subject to variation, giving rise to phenotypes which will vary in their success at learning language.

5 Evolution and starting small

The previous sections should make it clear that what is needed is some way of incorporating the less is more and starting small ideas into an evolutionary framework. Hurford's (1991) approach to explaining the critical period suggests that this approach to the "why" question we have posed can, at least in principle, also answer the "when" question. As we shall demonstrate in the remainder of this paper, this synthesis of evolutionary modelling and insights from incremental models of learning is indeed fruitful in explaining the subtle facts surrounding the timing and "shape" of critical periods.

Another reason why the evolutionary approach is interesting relates to the original motivation behind Elman's (1993) paper. His goal was to approach some of the criticisms that neural networks cannot learn context free grammars.

> "These sort of facts (and specifically, the recursive nature of embedded relative clauses) have led many linguists to conclude that natural language cannot be modelled by a finite state grammar (Chomsky 1957), and that statistical inference is not a viable mechanism for learning language (Miller & Chomsky 1963)....So it is reasonable to wonder how connectionist networks (which nonetheless rely heavily, though not exclusively, on statistical inference) possess the requisite computational properties for modeling those aspects of natural language..." (Elman 1993:75)

Those researchers that reject neural networks as a model of human language acquisition use the claim that neural nets cannot learn context free grammars as an argument for there being innate domain-specific knowledge that assists the learning task. To put it simply, the fact that "starting small" improves the network's success at learning long-distance dependencies appears to remove this support from the nativist argument. On the other hand, in order to pose the question of how incremental learning evolved we have to view it as a trait that is coded for in the genome. In other words, the developmental program that underlies Elman's response to the nativist argument is itself innate.

An obvious response to this might be that although the development of the learning mechanism must be innately specified at some level, this does not equate to *domain specific* knowledge. In other words, incremental learning may simply be a very general property of learning in humans which is applied to language purely because it happens to be part of the environment in which we develop. One of the conclusions we draw from the model presented in this paper is that this is unlikely to be the case, and that the timing of the development of the device that learns language in humans *has evolved for that purpose*. In this narrow sense, at least, we are presenting a case for innate, domain specific, knowledge of language being coded in the genome.

6 From the genome to development

If the time course of the development of the resources for learning is innately specified, and we wish to explore how this might have evolved, then we need to understand how it might be coded for in the genome. Logically there are two possibilities that we will explore:

- the development over time is directly specified genetically. In other words, the state of the phenotype at a particular age is determined solely by the genotype.
- the development over time is related to the input received, and this relationship is specified genetically. In this case the state of the phenotype at a particular point in time is determined by the input over the individual's lifetime as well as by the genotype.

The first possibility is perhaps the most obvious one, and is similar to that modelled in Hurford (1991). It seems to be uncontroversial to assume that the timing of the development of the phenotype can be genetically specified in this way. Indeed Nolfi & Parisi (1991) embed an artificial neural network within a genetic algorithm by allowing the (artificial) genome to control the development of the network's connections in this manner.

> "In [earlier simulations] the process that maps the genome of an organism onto a complete network is supposed to be instantaneous. No changes occur in a network during the entire life of an individual. In other words, the individual is born as an adult and there is no ontogenesis or development. A more biologically plausible alternative is temporally distributed mapping. The genome defines and constructs an initial network

> at birth but then it continues for a certain time after birth to determine changes in the network's architecture and weights." (Nolfi & Parisi 1991:8)

The second possibility corresponds to an approach in neuroscience that has had a lot of interest lately: constructivism. Quartz & Sejnowski's (1997) agenda-setting paper advocates a view of the development of the nervous system that highlights the role of the environment in directing neural growth (specifically dendritic arborisation).

> "According to 'neural constructivism', the representational features of cortex are built from the dynamic interaction between neural growth mechanisms and environmentally derived neural activity." (abstract)

Interestingly, Quartz & Sejnowski (1997) take this view as a fundamental challenge to nativism:

> " 'constructivist learning' minimises the need for prespecification in agreement with recent neurobiological evidence indicating that the developing cortex is largely free of domain-specific structure." (again, from the abstract)

Given this comment, it may seem perverse to try and factor constructivist ideas into an *evolutionary* model. The rejection of genotypic prespecification by constructivists must surely be in response to classical learnability theory, which typically treats the learning mechanism as fixed. However, a view of learning as a more dynamic activity, where the computational architecture itself is modified by the environment, says nothing in principle about the degree of innateness involved. As we will show, constructivist-type learning (at least for young learners) specified genetically is an emergent property of our evolutionary simulations.

Given these two ways in which the genome might specify the development of learning resources, which should be used for the simulation? We believe this question is wrongly posed, and the interesting features that the model show arise from this very choice. Instead of artificially making this choice in setting up the model, it is up to evolution to decide how it is to control the development of resources. In other words, the growth of the phenotype is *potentially* responsive both to a purely age-based maturational program, and a more "constructivist" input-sensitive development.

7 Evolutionary simulation

In this section we present the technical aspects of modelling the evolution of incremental learning: a simple model of learning that formalises the central features of "starting small" and "less is more", a model of the genome that allows for the two types of developmental control discussed in the previous section, and finally the genetic algorithm that follows the selective evolution of a population of learners.

7.1 A simple model of learning

Modelling each individual as an Elman net, for example, and then evolving such a population would be prohibitively costly in terms of computer time. In Elman's (1993) experiment, each network was presented with a total of 320000 sentences of varying lengths. In itself this results in a fairly computationally intensive simulation. If we were to multiply this by 100 networks in a population and watch these networks evolve over several thousand generations we are approaching 10^{11} individual sentences. Instead of this, we can take Elman's

results as providing a function that describes a class of *incremental learning problems*:

$$F(L_t, D_t, M_t) = L_{t+1}$$

where L_t is language already learnt at time t, D_t is the available input data at t, and R_t is the degree to which the learning mechanism has developed (i.e. the resources available at time t.) As discussed earlier, we take all three to be simple integer values. So, L can be thought of as a scale of complexity of linguistic ability, D as the richness of the input (or intake) data, and R as the size of resources (working memory in Elman's example, the size of the filter in Goldowsky & Newport 1993). The feature of incremental learning problems that is of interest to us can be captured by:

$$L_{t+1} = \begin{cases} L_t + 1 & \text{if } L_t + 1 = R_t,\ D_t \geq Q. \\ L_t & \text{otherwise.} \end{cases}$$

up to some maximum L, where Q is some threshold for the quality (or richness) of the intake data. Here we are assuming that the various "language stages" enumerated by L each correspond to a "resource stage" enumerated by R.

In fact, although this function encodes the *competence* acquired by an individual with a certain memory and input data, the actual performance is more likely to be given by:

$$L_{t+1} = \begin{cases} L_t + 1 & \text{if } L_t + 1 = R_t,\ D_t \geq Q. \\ R_t & \text{if } R_t < L_t. \\ L_t & \text{otherwise.} \end{cases}$$

As Elman (personal communication) has confirmed, it is likely that if, for some reason, the resource goes *down* during learning, linguistic performance of the networks in his simulation will be affected. Whether this is biologically realistic of course depends on one's interpretation of learning resource. In fact, the inclusion of this factor in the end makes little difference to the simulation results.[6] For the results presented in this paper, this second formulation will be used since it is assumed that any resource that is available for the learning of language is likely to be implicated in some way in the processing of language.

Notice, finally, that language is assumed to be finite. It is not possible for a learner to simply go on learning more and more language throughout his life in our model (unless we decide to allow for this by setting the maximum L to be very high). This maximum L is *not* what is evolving in our simulations. It is assumed that there is a certain "amount" of language out in the community that an individual may or may not be able to learn. Although we acknowledge that language learners *are* able to go beyond the data to some extent, there are constraints (be they cultural/linguistic or physical/biological) that put a limit on quite how far they can go. It is a completely different (although interesting) question that asks how this maximum got to be the way it is. Our interest here, however, is in showing how learning mechanisms adapt to learn a language of a given size.

7.2 The genome

The two possibilities for genetic specification of development outlined in the last section are modelled by having two genome strings for each individual:

[6] The only effect seems to be to increase the variance among runs of the simulation.

T-Genome this is a string of loci corresponding to each life stage of an individual.

L-Genome this is a string of loci corresponding to each possible language stage (i.e. up to maximum L).

Each locus can contain one of three alleles:

Promote promotes the growth of resources – i.e. R may be increased by 1. This is encoded as 1.

Inhibit inhibits the growth of resources – i.e. R may be decreased by 1. This is encoded as –1.

Null leaves the control of development to the other genome. This is encoded as 0.

For an individual at a particular time of life and having a particular language ability, the two types of genome may interact to control the development of resources. If, for example, the T-Genome locus is **Null** and the L-Genome locus is **Promote**, then resources increase. If the L- and T-Genomes disagree, then the result is no change:

$$R_{t+1} = \begin{cases} R_t + 1 & \text{if } G_T(t) + G_L(L_t) \geq 1. \\ R_t - 1 & \text{if } G_T(t) + G_L(L_t) \leq -1. \\ R_t & \text{otherwise.} \end{cases}$$

up to some maximum R and down to some minimum R (zero in fact), where $G_T(n)$ is the value of the allele at the nth locus of the T-Genome, and $G_L(n)$ is the value of the nth L-Genome allele.

Using this general formulation, we can experiment with the effect of either type of developmental control described earlier, and also what might happen given a combination. So, for example, a purely constructivist learner whose learning resources always grew in response to input but never otherwise would have genomes that looked like:

T-Genome $< 0,0,0,0,\cdots,0,0 >$

L-Genome $< 1,1,1,1,\cdots,1,1 >$

Clearly this type of coding allows for an enormous range of possible developmental programs amongst which evolution may select.

7.3 The genetic algorithm

In order to examine how individuals with genomes like those described in the previous section evolve, we use a form of genetic algorithm (see, e.g. Goldberg 1989, Mitchell 1996). We essentially simulate a small population of individuals of various ages growing up, selecting mates, producing offspring, and dying. As they are doing this, they also update their learning resources and language ability as described in the previous sections. There are a number of aspects to this algorithm that should be made explicit:

Population The initial population is a collection of individuals with a rectangular distribution of ages from 0 to the maximum age. They each have genomes that are completely random, with equal probabilities of **Promote**, **Null** and **Inhibit** at each locus. The individuals are androgynous.

Fitness The fitness of each individual is calculated based on that individual's current language ability *and* their resources. The exact fitness function can be varied; in particular it is possible to vary the extent to which non-linguistic fitness (the resource component) affects the calculation of fitness.

Sex and death The simulation in this paper uses *soft selection*. This means that fitness affects the chance of mating rather than the chance of dying. All individuals live to their maximum age, at which time they are removed from the population and are replaced by mating. In this way the size of the population remains constant. Every individual *past puberty* is a potential mate, but the probability of being chosen as a mate is proportional to fitness.

Crossover Finally, the genomes of the new individuals are formed by one point crossover of the two parents' genomes. This means that all the alleles up to a random point are taken from one parent, and the rest from the other. This crossover takes place on the L- and T-genomes separately.

8 Results

The typical initial variables in the simulations presented in this section are shown in the table below:

Variable	Value
Population size	100
Max. life stages	30
Puberty[7]	15
Fitness function	$L + 0.25R$
Mutation rate[8]	0.001
Length of run[9]	10000 life stages
Resource maximum	15
Language stages[10]	7
Intake quality[11]	0.5

The effects of varying some of these initial conditions will be discussed in the following sections, but there are a few results that we will not present, so a few comments are in order.

The fitness function was chosen mainly for its simplicity. Some experiments were carried out with a function that increased the difference between the fittest and least fit in the population, by cubing the sum of language and resources. Interestingly this meant that a higher mutation rate was required to get consistent results across runs. This is probably due to the variation in the population's gene pool being regularly exhausted by overfecundity of a few individuals, hence the need for a higher mutation rate to re-introduce variation. The end results with both types of fitness function are similar, however.

[7] This value is varied in a later section.

[8] This refers to the chance at each locus of a mutation to a random allele.

[9] This varies — in some cases the simulation was left running far longer than was necessary (100000 life stages) to ensure that no late evolutionary changes were being missed.

[10] This value is varied in a later section.

[11] This value is varied in a later section.

The choice of weighting for non-linguistic fitness was also varied. If the weighting was too high (i.e. resources were an important factor) then the individuals never performed well at the linguistic task. If the weighting was very low, then the results were similar to those presented here, although the variation across runs was greater. Again, changing the resource maximum had a similar effect. If it is reduced so that it is close to the minimum that can support language learning, then the effects that we discuss below are still visible, but are less reliable from run to run.

8.1 The effect of one type of development

8.1.1 The L-Genome: a constructivist learner

The first experiment with the simulation is designed to test what emerges if the learners in the population can only adjust their learning resources in response to input. In other words, they only have an L-Genome. Two interesting questions pose themselves: does a constructivist regime evolve? and does a critical period emerge?

Evolution of a constructivist strategy Firstly, we tested the average final language ability of learners with random L-genomes. To do this we ran the simulation 10 times with no evolution and tested all 1000 resulting individuals on their language ability at the end of life. This was then repeated the experiment with evolution "switched on". The results are summarised below:

Condition	Average final L
No evolution	0.339
Evolution	5.956

As can be seen there is a vast improvement with evolution. In fact, the average final language is very close to the maximum possible (the maximum is 6 since there are 7 language stages from 0 to 6).

After evolution, every individuals' L-Genome looks like:

$$< 1, 1, 1, 1, 1, 1, 1 >$$

which means that at each language stage, the learning resources are increased by one, preparing the individual for learning the next stage as soon as the data intake is sufficient. This then is the ideal pure "constructivist" strategy for managing learning resources.

Testing for critical period One of the advantages of using the computational methodology we are employing is that we can perform experiments on the evolved individuals — a kind of "simulated pathology" — that are possible only as rare and tragic natural occurrences in the real world. To test for the critical period we can deliberately deprive the individual learners of language until a particular life stage and look at how their language is acquired over their life time as a consequence. Notice that this degree of explicitness is missing from discussions of the critical period in Elman's (1993) paper.

Figure 2 shows the average final language attained by all the individuals in the 10 runs against the degree of deprivation to which they are exposed. Figure 3 shows the same information as well as the development of language over the lifetime of the individuals for each level of deprivation.

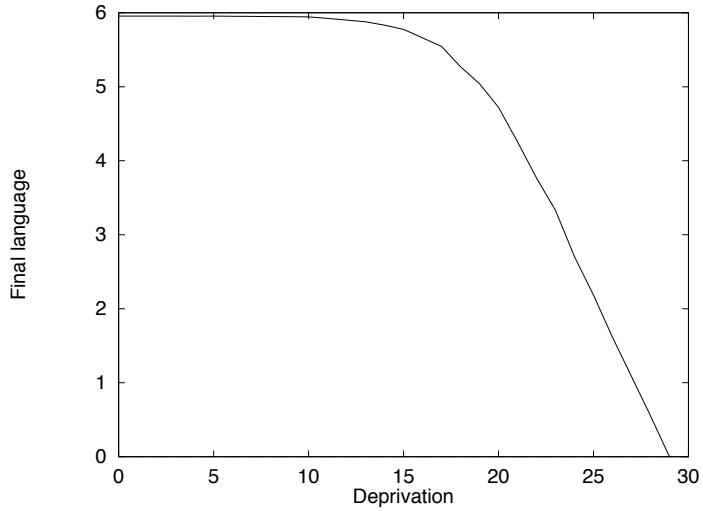

Figure 2: Final language attained against deprivation for evolved learners with only L-genomes. This graph shows that a critical period is not emerging.

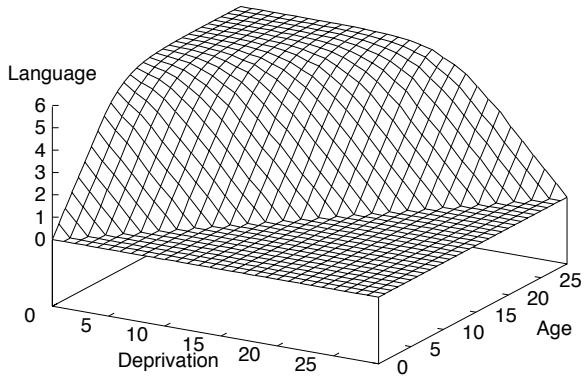

Figure 3: Language over lifetimes against deprivation for evolved learners with only L-genomes. The far edge of this 3d graph corresponds to figure 2; each line perpendicular to this slice shows how an average individual with a particular degree of deprivation progresses through his life.

What these results show is that the critical period does not emerge. Although the final language ability in figure 2 does tail off after a flat start, this is simply because there is less and less time to learn before the individuals reach their maximum age (this can be seen more clearly in the second figure). The ability to learn language is therefore unaffected by deprivation in the case where only an input-based control of development is allowed. This is not at all surprising given the shape of the evolved genome. Since resources increase every time there is enough input, then they will always keep perfectly in track with the requirements of the incremental learning problem, whatever time in life input starts to arrive.

8.1.2 The T-Genome: a non-constructivist learner

In the previous section we saw that the critical period does not emerge with an L-genome. We now repeat the experiment with only a T-genome. Again, we compare the random genome result with the evolved populations:

Condition	Average final L
No evolution	1.930
Evolution	3.734

Although evolution improves the final language reached over the random case, it is nowhere near the ability reached by the constructivist strategy. This is understandable since the presentation of language is not deterministic. At each life stage, with the initial settings given above, there is only a 0.5 probability of there being "enough" language intake to potentially progress to the next language stage (in other words, the data intake quality is 0.5). If the development of language resources cannot be sensitive to input and is simply rigidly determined by the age of the learner, then there is always a risk attached to increasing learning resources — it is always possible to overshoot the amount of resource required to learn the next stage of the incremental learning problem. To solve this problem with solely a T-genome, evolution increases resources conservatively as the learner develops. The early stages of a typical evolved T-genome looks like:[12]

$$< 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, \cdots >$$

It appears that this non-constructivist learning strategy is not very successful for the incremental learning problem with input whose rate of presentation/intake is not completely predictable. It is therefore unlikely that this is what is going on in language learning.

Furthermore, we can repeat the deprivation experiments and see that the critical period is not quite what we would expect either. Figures 4 and 5 show a critical period, but it is more catastrophic and earlier than what we would expect. This is understandable since language input only has to be delayed until after the second increment of resources coded by the T-genome in order for it to be completely unlearnable.

[12]The later parts of the evolved T-genome in this condition look rather different: more like those that evolve in the condition described in the next batch of experiments (i.e. all promote). In a similar fashion to those results, the "careful" stage of the T-genome seems to come to an end at about the time the maximum language ability is reached by the population — see later discussion for more details.

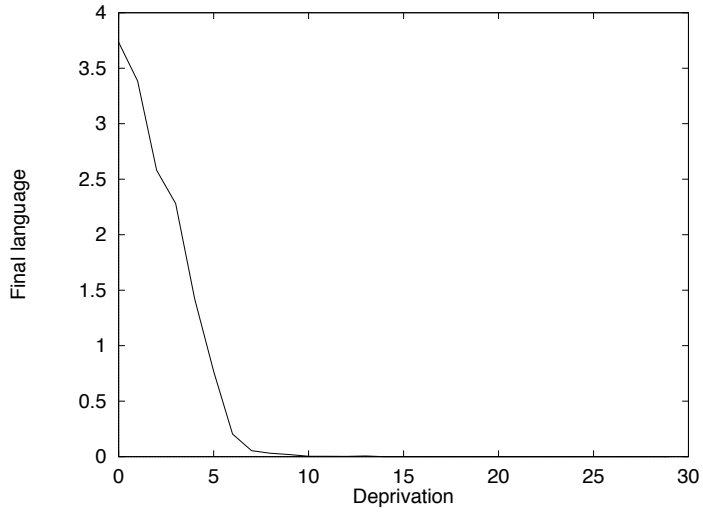

Figure 4: Final language attained against deprivation for evolved learners with only T-genomes. This graph shows an unrealistically early critical period, and an unrealistically low final language ability.

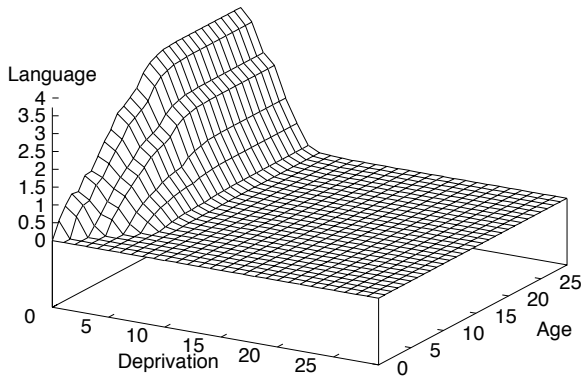

Figure 5: Language over lifetimes against deprivation for evolved learners with only T-genomes.

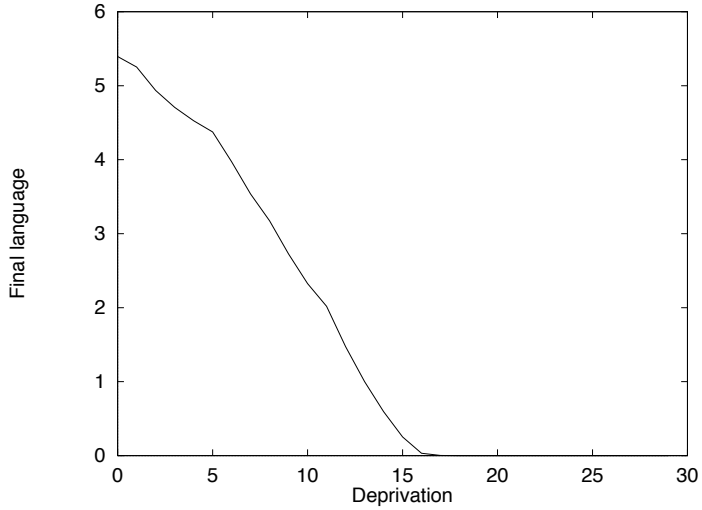

Figure 6: Final language attained against deprivation for evolved learners with both genomes. This graph shows a critical period in many ways similar to that observed in the data in figure 1.

8.2 The effect of combining developmental control

Neither purely input-based nor solely age-based control of development is able to account for the observed behaviour of language learners. A purely constructivist strategy yields no critical period, whilst a completely non-constructivist strategy yields a poor learner, and an unrealistically early critical period.

The next step is to combine both the L- and T-genomes and allow evolution to select which will control the development of learning resources. It is important to realise that the manner in which the two types of developmental control can combine allows for the possibility of control by only one genome, the other, or both at any particular stage of life. This considerably increases the range of variation of possible organisms, and the power of evolution to design developmental programs.

The average final language attained for the random and evolved genomes are shown as before:

Condition	Average final L
No evolution	1.862
Evolution	5.677

Notice first that the final language ability of the evolved population is near the perfect 6, although it is not quite as good as the purely constructivist learner.[13]

The critical period Repeating the deprivation experiments using this new population gives the results shown as figures 6 and 7. There is a clear critical period here, which is not due to running out of time before death as in figure 2. The final language achieved is considerably

[13]This is due to a ceiling effect imposed by the language maximum combined with a greater degree of variance which is a side-effect of having a critical period as we shall see later.

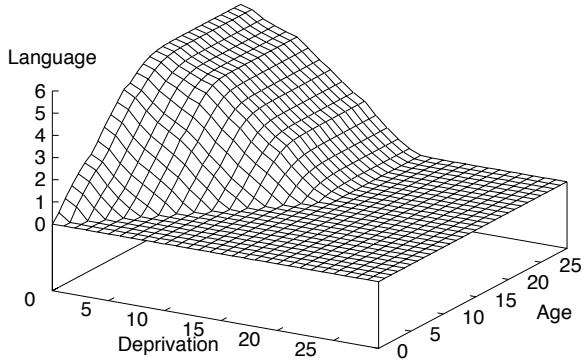

Figure 7: Language over lifetimes against deprivation for evolved learners with both genomes.

better than in figure 4, and the critical period is more lifelike in that it is less catastrophic and later in life.

Down's syndrome Another feature of language learning in exceptional circumstances that is discussed by Lenneberg (1967) is that sufferers of Down's syndrome typically acquire language at a slower rate than the population average *but plateau at a similar age*. We can test if our evolved population behaves in a similar fashion by setting up another simulated pathology. It is assumed for simplicity here that we can simulate the slow learning of Down's children by reducing the intake quality of language during the lifetime of evolved individuals. Importantly, we take a population that evolved under conditions where intake quality was 0.5, and reduced this to 0.25 only for the final experiment. This means — as with the critical period experiments — that the individuals subjected to 0.25 quality possessed genomes which had evolved to cope with different conditions to those to which they are subjected.[14]

The results of this experiment are shown in figure 8. The normal average language grows at a rate of roughly 0.5 per life stage as expected from the quality of 0.5. It also plateaus at 5.677, as discussed earlier. The Down's population's average language grows at roughly 0.25 per life stage, also what is expected given the reduced language quality. However, language ability stops increasing at the same time as in the normal case, consequently having a lower maximum of 2.763. It appears, then that the levelling off of language in the population is not due to reaching a maximum level (or close to a maximum level), but instead to some innate halting of the learning process at a particular age.

[14]This feature of these experimental runs is similar to that argued by the proponents of Evolutionary Psychiatry to be the basis of many human psychiatric problems.

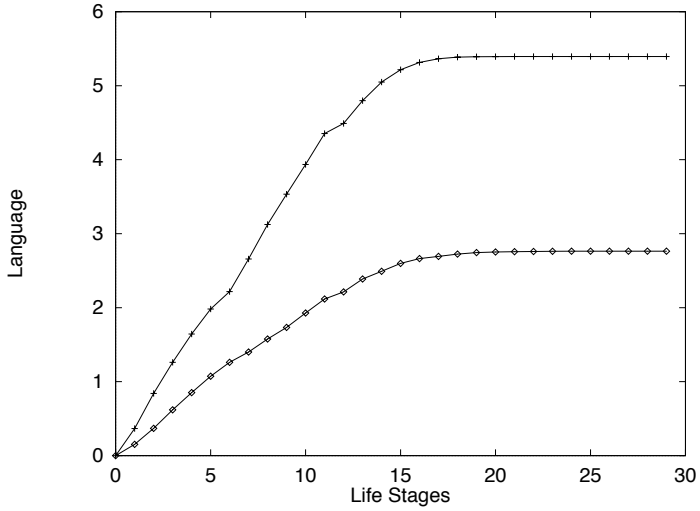

Figure 8: Language against life stages over lifetimes of evolved learners with normal intake quality (upper line), and half intake quality (lower line).

8.2.1 The genomes

In order to understand how this break in the incremental learning process is coded for innately, we can examine the average genome of the individuals in the evolved populations. Figures 9 and 10 show the averaged L- and T-genomes respectively. The L-genome looks very much like the one that evolved in the first experiment with the T-genome disabled. In other words, each locus promotes the growth of learning resources.[15] This means that *as long as the T-genome's alleles are all null* these learners would approximate the constructivists of the first experiment.

Looking at the T-genome it is obvious that there are now two distinct phases to the life of these learners: an initial phase where control is given over to the L-genome[16] (giving rise to a constructivist strategy), and a final stage where the control of development is insensitive to the presence of input. This, then, is the reason for the critical period and Down's effects. The timing of these effects coincides with the end of the input-sensitive phase of development.

8.3 Normal language maturation age

So far, all the runs of the simulation have used the same values for language maximum and intake quality during evolution. With values of 6 and 0.5 respectively, we would expect a learner to ideally reach maximum language after 12 life stages. Although the learners that evolve do not quite achieve this ideal, the average language ability starts tailing off around this age. Similarly, the Down's syndrome learners also start to flatten out after this age.

[15]The first locus is only promote half of the time. This is because of the way the incremental learning problem is set up; before learning can commence, resources always have to increased to one. It makes sense to do this in the first life stage, but it does not matter whether this is done by the L- or the T-genome.

[16]In this phase, the T-genome does not consist exclusively of null alleles, but these are by far the most common. Furthermore, if there is a promote allele, for example, the effect of this is usually reversed by a following inhibit allele.

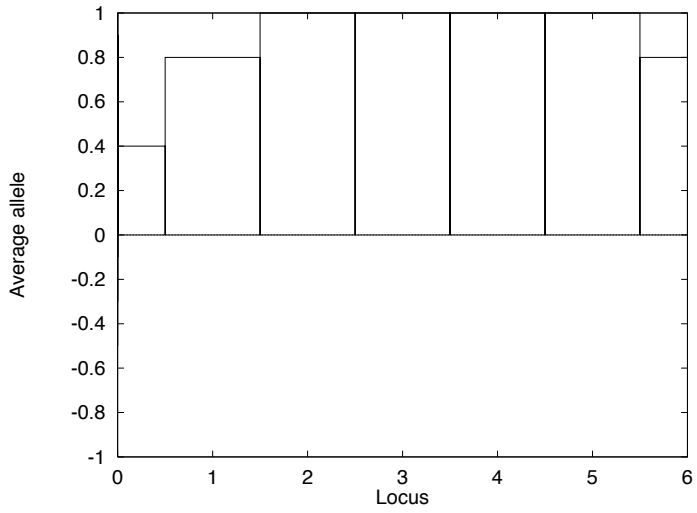

Figure 9: Average allele value against locus for the evolved L-genome.

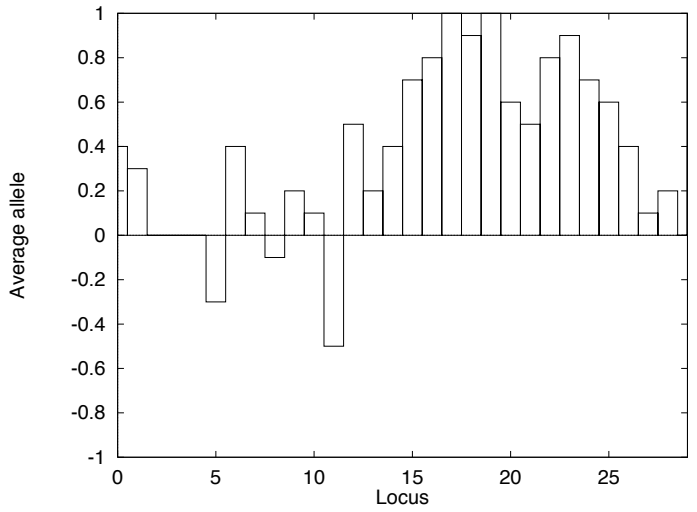

Figure 10: Average allele value against locus for the evolved T-genome. In this graph, the genome appears to be split into two distinct sections.

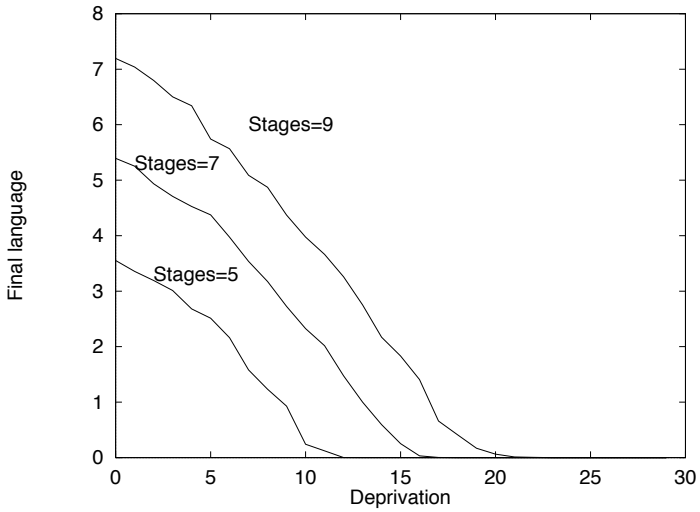

Figure 11: Final language attained against deprivation with quality 0.5 and various numbers of language stages.

Looking at figure 6, then we can see that if evolved learners are deprived language for 12 life stages, they do worse than the non-evolved language maximum (attainment is 1.474, as compared to 1.862 achieved by those with no evolution). This is the first life stage at which deprivation has an effect that is worse than having a random genome and language from birth.

Given these facts, a good working hypothesis might be that there is something about the age at which language is normally acquired in the population that gets coded into the genome which affects the development of individuals causing the critical period and Down's effects. To test this, we can re-run the simulation with different language stages and intake qualities to force the age by which language *can* be learned by the best learner. Figures 11 and 12 show the differing critical periods for these new runs (notice that puberty is the same for each of these runs). Clearly, the timing of the critical period is closely related to expected age at which language will be learnt by the normal population. With the increased number of language stage, the age at which deprivation severely effects learning is later; and with decreased language stages or increased language quality, the reverse is true. It appears that the critical period is indeed timed to occur around the age at which language is normally acquired by the population.

8.4 The effect of puberty

Given the results from the last section, it would appear rather surprisingly, that the timing of the critical period is *not* associated causally with the onset of puberty. Instead it occurs at around the age that language is typically learnt by the normal population. Our final results test this conclusion by varying the onset of puberty in the simulation.

Figure 13 shows the results of varying the age at which organisms can reproduce (and hence the age at which fitness starts to have a role in selection) from 5 life stages through 10,15 and 20. It is clear that the setting of puberty has no effect on the timing of the critical

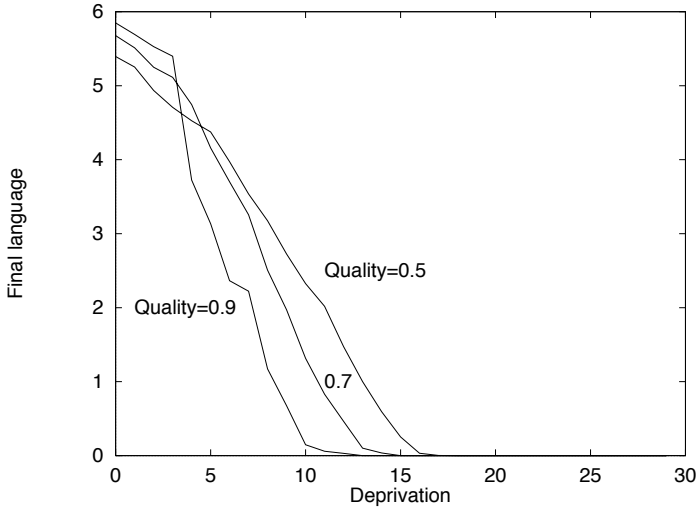

Figure 12: Final language attained against deprivation with 7 language stages and various quality values.

period.

9 Discussion

These results show the emergence of incremental learning with two phases: a first phase where development is sensitive to input (a constructivist phase), and a second phase where development is endogenously controlled. The timing of the switch from one phase to the other varies according to the time when language is typically learnt. This predicts the shape and timing of the critical period and "Down's" effects. Down's sufferers plateau at a similar age to the average population, and the critical period has been argued to become most critical around age 10, arguably the time when the only increase in linguistic knowledge in the population average is vocabulary learning. Crucially, these facts are logically independent. There is no a priori reason why the critical period couldn't be timed at a completely different stage of life from the point at which language is typically learned.

One way of thinking about the simulation results is that evolution is "protecting" the constructivist development of resources for the period of life that it is necessary to do so. It will not allow these resources to be altered in a non-constructivist way since this may upset the incremental learning process. However, after the time when it is safe to assume that language should have been learnt, control is "handed over" to an endogenous developmental process which *maintains* the resources for most of the rest of life (though there may be decay towards the end) even though there is little more learning.

Why puberty? The critical period appears to be timed with the onset of puberty. These results suggest that this is coincidence; that there is no causal connection between puberty and the switch in developmental control. The question, however, should be why language is typically learnt before puberty. A simple evolutionary answer is that there is a pressure to ensure that a skill that has a high social benefit should be completely acquired by the time

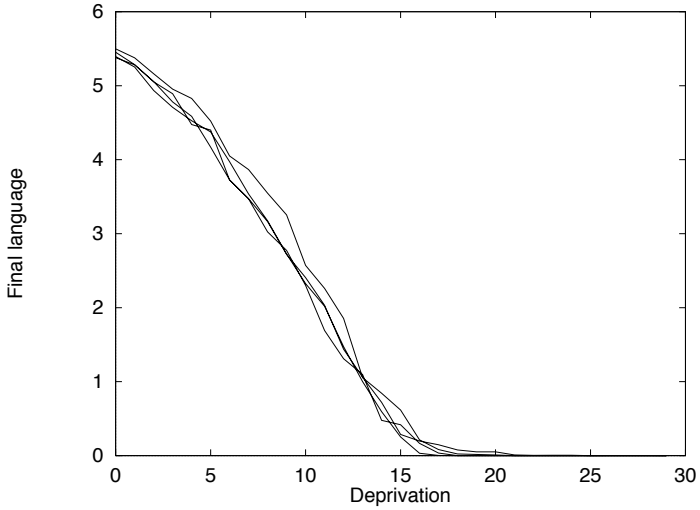

Figure 13: Final language attained against deprivation with 7 language stages, quality 0.5 and various ages for puberty.

mating starts (i.e. puberty). It is likely that language independently evolved *to be the right size to be learnt by puberty* (Hurford & Kirby 1997)[17]. Given this, if the critical period evolves to be timed with the end of normal language learning, then it will appear to be related to puberty. This, however, is an indirect causal connection.

10 Conclusions

The main points raised by these results are:

1. The critical period is inevitable with problems that require incremental learning.
2. The timing of the critical period is predicted to coincide with when these problems are typically learnt.
3. If a child has impoverished uptake, then she will plateau at around the time full language would be typically learnt by the population.
4. These results are only possible with independently evolved input related and age-related control of the incrementation of the learning resource.

Finally, we believe that this work has some interesting implications for conceptions of innateness. In contradiction to the position of Quartz & Sejnowski (1997) domain specificity is not incompatible with constructivism, or indeed any model that highlights the importance of statistical learning. In our view the language acquisition device is the result of the gradual

[17]If we assume that learners cannot regularly go very far beyond the data presented to them, then the limit on language size may have evolved historically through a process of cultural evolution. Whatever amount of language is available in the community in which the learner grows up will then affect the *biological* evolution of the learning strategies to efficiently acquire that language. This hypothesised cultural/biological co-evolution raises many issues that go beyond this paper (see Kirby 1997; Kirby & Hurford 1997 for some discussion).

evolution of general purpose learning mechanisms put to the task of learning a culturally evolving system: language. The ways in which the particular task of learning language is nativised may not be immediately obvious. In the case discussed in this paper nativism shows up as a complex developmental program which is tailored to the task of learning language. It is expected that as we explore other aspects of language learning and evolution with the kinds of methodology developed here, we will have to rethink what it means to have a domain specific Language Acquisition Device.

References

BADDELEY, ALAN D. 1986. *Working Memory*. Number 11 in Oxford psychology series. Oxford: Clarendon Press.

—— 1990. *Human memory : theory and practice*. Hove, Sussex: Lawrence Erlbaum.

—— 1992. Working memory: the interface between memory and cognition. *Journal of Cognitive Neuroscience* 4.281–288.

——, C. PAPAGNO, & G. VALLAR. 1988. When long-term learning depends on short-term storage. *Journal of Memory and Language* 27.586–595.

——, BARBARA A. WILSON, & FRASER N. WATTS. 1995. *Handbook of memory disorders*. Chichester: Wiley.

BERWICK, ROBERT C. 1985. *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

BEVER, THOMAS G. 1981. Normal acquisition processes explain the critical period for language learning. In *Individual Differences and Universals in Language Learning Aptitude*, ed. by K.C. Diller, 176–198. MA: Newbury House, Rowley.

BIRDSONG, D. 1992. Ultimate attainment in second language acquisition. *Language* 68.706–755.

BJORKLUND, DAVID F., & BRANDI L. GREEN. 1992. The adaptive nature of cognitive immaturity. *American Psychologist* 47.46–54.

CHOMSKY, NOAM. 1957. *Syntactic Structures*. The Hague: Mouton.

——. 1981. *Lectures on Government and Binding*. Foris.

CURTISS, SUSAN R. 1977. *Genie: A Linguistic Study of a Modern day "Wild Child"*. New York: Academic Press.

—— 1980. The critical period and feral children. *UCLA Working Papers in Cognitive Linguistics* 2.21–36.

—— 1988. Abnormal language acquisition and the modularity of language. In *Linguistics: The Cambridge Survey: Vol.2. Linguistic Theory: Extensions and Implications*, ed. by F. Newmeyer, 96–116. Cambridge: Cambridge University Press.

ELMAN, JEFFREY L. 1993. Learning and development in neural networks: the importance of starting small. *Cognition* 48.71–99.

FODOR, JERRY A. 1981. The present status of the innateness controversy. In *Representations: Philosophical Essays on the Foundations of Cognitive Science*, ed. by J.A. Fodor, 257–316. Brighton, Sussex: The Harvester Press.

GATHERCOLE, SUSAN, & ALAN D. BADDELEY. 1990. Phonological memory deficits in language-disordered children: Is there a causal connection? *Journal of Memory and Language* 29.336–360.

——, & ——. 1993. *Working Memory and Language*. Hove, Sussex: Lawrence Erlbaum.

GOLDBERG, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

GOLDIN-MEADOW, S. 1982. The resilience of recursion: A study of a communication system developed without a conventional language model. In *Language Acquisition: State of the Art*, ed. by E. Wanner & L.R. Gleitman, 51–77. Cambridge: Cambridge University Press.

GOLDOWSKY, B. N., & ELISSA J. NEWPORT. 1993. Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In *The proceedings of the 24th Annual Child language Research Forum*, ed. by E. Clark, 124–138. Stanford, CA: Centre for the Study of Language and Information.

HAYNES, H., B.L. WHITE, & R. HELD. 1965. Visual accommodation in human infants. *Science* 148.528–530.

HOFER, MYRON. 1981. *The Roots of Human Behaviour: an introduction to the psychobiology of early development*. Oxford: W.H. Freeman.

HURFORD, JAMES. 1991. The evolution of the critical period for language acquisition. *Cognition* 40.159–201.

——, & SIMON KIRBY. 1997. Co-evolution of language-size and the critical period. In *New Perspectives on the Critical Period Hypothesis and Second Language Acquisition*, ed. by David Birdsong. Lawrence Erlbaum. In press.

ITTELSON, W.H. 1951. Size as a cue to distance. *American Journal of Psychology* 64.54–67.

JAMES, WILLIAM. 1890. *The Principles of Psychology, Volume 1*. New York: Holt.

JOHNSON, JACQUELINE S., & ELISSA J. NEWPORT. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology* 21.60–99.

JOYCE, JONATHAN, 1996. When/why/of what is less more? Master's thesis, Centre for Cognitive Science, University of Edinburgh.

KIRBY, SIMON. 1997. Fitness and the selective adaptation of language. In *Evolution of Language: Social and cognitive bases for the emergence of phonology and syntax*, ed. by J. Hurford, C. Knight, & M. Studdert-Kennedy. To appear.

——, & JAMES HURFORD, 1997. Learning, culture and evolution in the origin of linguistic constraints. Submitted to ECAL97.

LAKOFF, GEORGE. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

LENNEBERG, ERIC H. 1967. *Biological Foundations of Language*. New York: Wiley.

LIN, TSUNGNAN, BILL G. HORNE, & LEE C. GILES. 1996. How embedded memory in recurrent neural net architectures helps learning long-term temporal dependencies. Technical Report Computer Science Technical Report CS-TR-3626 and UMIACS-TR-96-28, University of Maryland, College Park, MD 20742.

LONG, MICHAEL H. 1990. Maturational constraints on language development. *Studies in Second Language Acquisition* 12.251–285.

MAYBERRY, S., S. FISCHER, & N. HATFIELD. 1983. Sentence repetition in American Sign Language. In *Language in Sign: International Perspectives on Sign Language*, ed. by J. Kyle & B. Woll, 83–97. London: Croom Helm.

MCKENZIE, N.R., H.E. TOOTELL, & R.H. DAY. 1980. Development of visual size constancy during the first year of human infancy. *Developmental Psychology* 16.163–174.

MILLER, GEORGE, & NOAM CHOMSKY. 1963. Finitary models of language users. In *Handbook of Mathematical Psychology II*, ed. by R. Luce, R. Bush, & E. Galanter. Wiley.

MITCHELL, MELANIE. 1996. *An Introduction to Genetic Algorithms*. Cambridge MA: MIT Press.

MOOD, D.W. 1979. Sentence comprehension in preschool children: Testing an adaptive egocentrism hypothesis. *Child Development* 50.257–250.

NEWPORT, ELISSA. 1984. Constraints on learning: Studies in the acquisition of american sign language. *Paper and Reports on Child Language Development* 23.1–22.

——, & TED SUPALLA, 1992. Critical period effects in the acquisition of a primary language: I. the influence of maturational state on the acquisition of complex morphology in american sign language. University of Rochester.

NOLFI, STEFANO, & DOMENICO PARISI. 1991. Growing neural networks. Technical Report PCIA–91–15, Institute of Psychology C.N.R., Rome.

OPPENHEIM, R.W. 1981. Ontogenetic adaptation and retrogressive processes in the development of the nervous system and behaviour. In *Maturation and Development: Biological and Psychological Perspectives*, ed. by K.J. Connolly & H.F.R. Prechtl, 73–108. Philadelphia: International Medical Publications.

OYAMA, SUSAN C. 1979. The concept of the sensitive period in developmental studies. *Merrill-Palmer Quarterly* 25.83–103.

—— 1985. *The Ontogeny of Information: developmental systems and evolution*. Cambridge: Cambridge University Press.

QUARTZ, STEVEN R., & TERRENCE J. SEJNOWSKI. 1997. The neural basis of cognitive development: a constructivist manifesto. *Behavioral and Brain Sciences* . To appear.

QUINE, WILLARD VAN ORMAN. 1960. *Word and Object*. Cambridge MA: MIT Press.

RENSHAW, S., R.J. WHERRY, & J.C. NEWLIN. 1930. Cutaneous localization in congenitally blind versus seeing children and adults. *Journal of General Psychology* 38.223–238.

ROSE, S.A., A.W. GOTTFRIED, & W.H. BRIDGER. 1978. Cross-modal transfer in infants: Relationship to prematurity and socioeconomic background. *Developmental Psychology* 14.643–652.

SALAPATEK, P., & M.S. BANKS. 1978. Infant sensory assessment: Vision. In *Early Behavioral Assessment of the Communicative and Cognitive Abilities of the Developmentally Disabled*, ed. by F. Minifie & L. Lloyd.

TODD, PETER. 1996. The causes and effects of evolutionary simulation in the behavioural sciences. In *Adaptive Individuals in Evolving Populations: Models and Algorithms*, ed. by Richard Belew & Melanie Mitchell. Addison-Wesley.

TURKEWITZ, GERALD, & PATRICIA KENNY. 1982. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Developmental Psychology* 15.357–368.

WESSON, ROBERT. 1993. *Beyond Natural Selection*. MIT Press.

WEXLER, KENNETH, & RITA MANZINI. 1987. Parameters and learnability in binding theory. In *Parameter Setting*, ed. by Thomas Roeper & Edwin Williams. Dordrecht: Reidel.

WHITE, LYDIA, & FRED GENESEE. 1996. How native is near-native? the issue of ultimate attainment in adult second language acquisition. *Second Language Research* 12.

WOODWARD, J.C. 1973. Some characteristics of pidgin sign english. *Sign Language Studies* 3.39–59.