

Innately Constrained Learning: Blending Old and New Approaches to Language Acquisition

Elissa L. Newport and Richard N. Aslin
University of Rochester

In S. C. Howell, S. A. Fish, and T. Keith-Lucas (eds.), Proceedings of the 24th Annual Boston University Conference on Language Development. (pp. 1-21).
Somerville, MA: Cascadilla Press, 2000.

1. Introduction

In the last few years we have begun to develop and investigate an approach to language acquisition we have called '*statistical learning*' (a term adopted from a related literature in computational linguistics; see Charniak, 1993). The central notions are a blend of ideas from structural linguistics (cf. Harris, 1951) and nativist perspectives (cf. Chomsky, 1955, 1981), along with more recent proposals using distributional analysis as part of the process of language acquisition (Maratsos & Chalkley, 1980; Morgan, Meier, & Newport, 1987; Rumelhart & McClelland, 1986). As noted by structural linguistics and used by all modern linguists, a great deal of information regarding the basic units and grammatical constructions of a language is provided by the patterns in a corpus from that language: what elements regularly occur together, what elements cannot be interrupted, and the like. At the same time, until recently, it has seemed unlikely that much of this information would be available for the process of language acquisition. First, it has been hard to imagine that a young language learner would be able to conduct the relevant analyses, online and with rapidly changing and evanescent auditory materials, rather than with a tape recorder and pencil. Perhaps more important, as Chomsky has emphasized (1955, 1957, 1981), there is an unlimited number of analyses an open-minded learner could conduct, many of which would find some support in a large corpus. A successful learner must therefore approach the problem with a constrained set of hypotheses about the nature of the regularities languages contain, and limited slavishness to the details of the corpus.

Our own approach has been to consider these approaches and problems anew and in a somewhat different way. By starting with a very simple version of distributional learning, we have been able to show that even young infants can in fact perform remarkably rapid and surprisingly complex computations of the co-occurrences among neighboring elements. More recently we have begun to investigate, step by step, learners' abilities to perform more complex computations, on both speech and other types of patterned materials; and we have begun to compare the abilities of human learners with those of other species. Our results raise the possibility that at least some of the constraints on hypotheses human learners have about language structure may arise from biases and selectivities in the types and complexities of the computations learners are able to perform.

How does a statistical learning approach change the problem of language acquisition? In some ways it does not fundamentally alter the problem at all; but it does importantly alter some aspects of the mechanisms by which the process is implemented. For example, using a corpus, and statistical computations performed over this corpus, as the input to language learning means that a learner is not confined to formulating rules or identifying patterns from single instances, or to being radically misled by individual counterexamples or occasional superficial similarities between two different constructions. As long

as the learner has procedures for distinguishing highly consistent versus infrequent properties from one another (not a trivial claim, it should be noted), a corpus will provide a much richer source of information for identifying grammatical regularities than we have previously thought. At the same time, however, access to the distributional properties of a corpus and the calculations one might perform on it only worsens the theoretical problem of limiting hypotheses: corpora provide yet larger infinities of things to calculate. A crucial part of our approach, then, rests on understanding the limits and selectivities on the computations real human learners are capable of performing, and determining how these selectivities of computation match up with the limitations and recurring patterns of structure across human languages. Insofar as these observations fit together, one might argue that some of the innate components of language acquisition and language structure arise, at least in part, from such computational constraints.

We are currently in the early stages of a long enterprise, asking what kinds of computations different aspects of language require, and then systematically moving from testing very elementary computations to more complex ones. In addition, we are investigating how such computational abilities may be similar or different in various types of learners and on various types of patterned materials. Our expectation is that, while certain aspects of these answers will be the same across learners and domains, in more complex computations there will be differences between human infants and other types of learners, and between humans' abilities to track patterns in speech versus music and vision (otherwise language would not be different in its structure from music or visual patterns). But we are not yet at the point where we can make such claims. Rather, we are just beginning to examine some of the elementary computations language learning requires, offer evidence for some of the computations human learners can perform, and (surprisingly, even at an early stage) suggest some constraints even in some fairly basic computations which appear to be quite limited in humans, in accord with structural properties of natural language.

In the present paper we will review our evidence to date on three aspects of this research program. First, an initial question: Can learners compute a very basic statistical property of input: the probabilities of adjacent sound sequences? We will review our studies of this question for human learners on simple speech materials, on nonlinguistic materials offering patterns in other domains, and also some early evidence on this same question for nonhuman primates. Second, we ask what is learned from statistical computations. Is the product of this type of learning statistical (thus changing rather radically our notions of language representation, as some have claimed)? Or is statistical information being used to acquire representations which are not themselves statistical? Finally, we will overview our initial results on the types of more complex computations we have begun to investigate, the constraints we have begun to discover in the statistical procedures human learners can and cannot perform, and the correspondences between these constraints on learning and the types of patterned regularities which do and do not occur in natural languages.

2. An Elementary Computation: Statistical Regularities among Adjacent Elements

Our first question, then, concerns whether human learners are able to keep track of the sequences of syllables they hear in a stream of speech, and indeed whether they are able to keep track of which sequences are highly consistent, and which are only occasional. This ability might be useful and needed in a number of aspects of language learning, but it is particularly relevant to the problem of *word segmentation*. In a series of studies (Saffran, Newport & Aslin, 1996; Saffran, Aslin & Newport, 1996; Saffran et al., 1997; Aslin, Saffran & Newport, 1998), we have asked: how does the learner determine, from the apparently continuous stream of speech, which sequences of sounds (or, in the case of signed languages, movements and handshapes) form the words of the language? This clearly must involve learning, since languages vary so widely in which sound sequences form words. Part of the answer involves the use of prosodic and rhythmic information, as well as silence at the ends of utterances (Aslin, Woodward, LaMendola, & Bever, 1996; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Jusczyk, Cutler, & Redanz, 1993; Mehler, Dupoux & Segui, 1990; Morgan & Saffran, 1995). However, such cues are not always available for use in initial segmentation (Aslin et al., 1996).

Several investigators (Chomsky, 1955; Harris, 1955; Hayes & Clark, 1970; Goodsitt, Morgan, & Kuhl, 1993) have noted that this problem might be solved by keeping track of relative consistency in the sound sequences. This observation can be converted into a statistical form: learners might compute the conditional probabilities between sequential syllables (called *transitional probabilities*; cf. Miller & Selfridge, 1950; Goodsitt, Morgan & Kuhl, 1993; Christophe, Dupoux, Bertoncini, & Mehler, 1994; Saffran, Newport, & Aslin, 1996). Over a speech corpus, those sequences with relatively high conditional probabilities are likely to be inside words, and those with relatively low probabilities are likely to be the accidental juxtapositions of sounds at word boundaries. Following an important study by Hayes & Clark (1970), we asked whether human learners were capable of performing such computations.

Note that the relevant computation involves not merely the tabulation of the *frequency* with which one element follows another; such frequencies must be adjusted by the base rate at which the elements occur individually. That is, what is most relevant to the coherence of sound elements within a word is the *predictiveness* of one sound given another. If 100 instances of sound X are followed by 100 instances of sound Y, and never by any other sound, then predictiveness is high. In contrast, if sound X is followed by sound Y 100 times, but by sound Z 100 times, then the frequency of co-occurrence of X and Y is the same as in the first example, but the predictiveness is much lower. In order for learners to determine which elements belong together consistently, ignoring differences in element frequency (and therefore chance co-occurrence), they must be able to compute a conditionalized statistic. In our own work, we have

discussed a transitional probability metric for this computation, due largely to its history in the field of psycholinguistics, but other metrics (for example, *mutual information* or *conditional entropy*) could also be used and, on the materials we have employed thus far, would give the same results.

2.1 Can human learners perform these computations online?

Our initial study involved presenting adults with an artificial language (Saffran, Newport & Aslin, 1996; see also Hayes & Clark, 1970). The language consisted of trisyllabic ‘words,’ concatenated in random order and spoken by a speech synthesizer with no prosodic or acoustic markers of word boundaries, to create an unbroken 21 minute corpus. Although transitional probabilities varied both within and between words, the transitional probabilities inside words were relatively high, while those spanning a word boundary were relatively low, as is the case for real languages. After exposure to the corpus, subjects were given a series of 2-alternative forced-choice items, each containing a word from the language and either a non-word or a part-word (depending on the experimental condition). Non-words were 3-syllable sequences made of the same syllables used in the language, but in an order which did not occur in the exposure corpus. Part-words were 3-syllable sequences consisting of two syllables in the correct positions and order, and a third syllable which did not occur in that position in the corpus. Subjects were to choose which alternative in each item sounded more familiar. We found in both the non-word and the part-word conditions that subjects performed significantly and substantially above chance, suggesting that adults not only can acquire syllable order, but also can segment a stream of syllables into groups based on the distributional characteristics of the corpus.

In a second study we asked whether 5- to 6-year old children could perform the same task (Saffran, Newport, Aslin, Tunick & Barrueco, 1997). To prevent the children from getting bored during familiarization, we asked them to color on the computer, using a program called KidPix, and we merely played the speech stream in the background, with no instructions to learn or even listen to the sounds. For comparison, adults were given the same exposure. After one or two 21-minute coloring sessions, both the adults and the children performed significantly above chance on a word-nonword forced-choice task (see Figure 1). Thus children can also segment words from fluent speech based solely on statistical information from a continuous corpus. Moreover, this process can apparently proceed implicitly, without subjects’ attention directed at the speech stream or the analytic process.

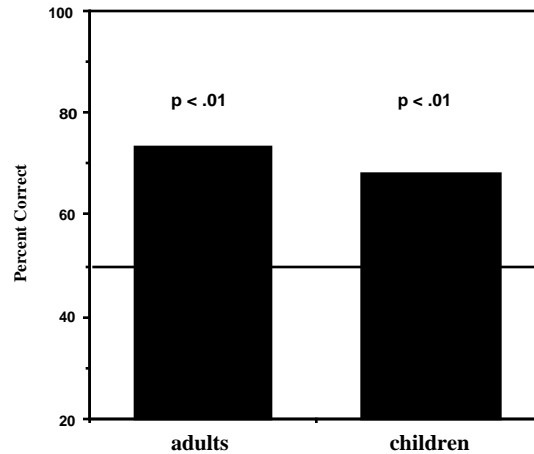


Figure 1. Mean performance of adults and children on a two-alternative forced-choice test of word vs. non-word statistical learning under implicit conditions.

We have also conducted a series of three studies on statistical learning in 8-month old infants (Saffran, Aslin & Newport, 1996; Aslin, Saffran, & Newport, 1998), chosen because this is the age at which word segmentation in natural language acquisition is underway. In our first study, infants were exposed to a simplified corpus of trisyllabic nonsense words (with transitional probabilities inside words of 1.0 and those across word boundaries of .33), presented continuously for only 2 minutes. Then, using the preferential-listening methodology (Jusczyk and Aslin, 1995), we tested each infant with two words from the language and two non-words (made up of familiarization syllables in a novel order). Our results showed that infants listened differentially to words versus non-words, indicating that they could discriminate between them. Because the individual syllables in words and non-words occurred with equal frequency in the familiarization corpus, the results cannot be due to discriminating the frequency of individual syllables, but rather must be due to discriminating syllable order: infants must be noting that the syllables in non-words never occurred in that order in the familiarization corpus. In our second study, we asked whether 8-month olds could perform the more difficult task of discriminating words from part-words. In this study, the familiarization corpus was like that in the first study, but the test items consisted of words and part-words. Part-words in this study were more difficult to discriminate from words than in our previous adult work. Here, part-words consisted of the final syllable of one word and the first two syllables of another word. Thus these part-words had in fact occurred in the familiarization corpus. They differed from words in

having transitional probabilities of .33 and 1.0 (as compared with 1.0 and 1.0 in the words). Infants in this second study also listened differentially to the part-words as compared to the words (see Figure 2). Thus, 8-month-olds do not merely note whether a syllable sequence occurred or not, but apparently can perform an analysis of the statistics of the language corpus.

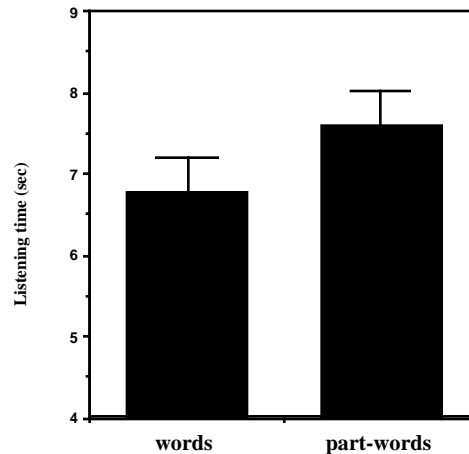


Figure 2. Mean listening times by 8-month-old infants to words and part-words after a 2-minute exposure to a continuous stream of 3-syllable ‘words’.

This second infant study does not, however, demonstrate precisely what statistic the infants are computing, and whether in particular they are capable of computing conditional probabilities among sequential syllables. Because the words of the corpus were each presented with equal frequency, the part-words formed by their junctures were all less frequent than the words. Infants therefore could have been responding to tri-syllabic frequencies, rather than tri-syllabic conditional probabilities. (Either of these would be quite impressive, but conditional probabilities would be more structurally informative in real language learning, as discussed above.) To pursue this issue further, we conducted a third study of 8-month-olds (Aslin, Saffran & Newport, 1998), in which the test words and part-words were matched in frequency of presentation during familiarization, and differed only in conditional probabilities. We achieved this by creating a corpus in which two of the four words presented in the familiarization corpus were more frequent than the other two. This resulted in a corpus in which the two part-words (formed by the juncture between the high-frequency words) occurred with the same frequency as the two low-frequency words. Nonetheless, the transitional probabilities within words were

still higher (1.0 and 1.0) than the transitional probabilities within the part-words (.50 and 1.0). Our results showed that infants continued to discriminate between the words and the part-words, demonstrating clearly that they can compute transitional probabilities (or other equivalent conditionalized statistics) and can use them to segment multisyllabic words from fluent speech. Moreover, this task involves the running computation of 20 different conditional probabilities, each over 45 to 90 occurrences of the component syllables and 9 to 90 occurrences of syllable pairings, during a 3 minute period. Eight of these 20 conditional probabilities are included in our test items. Our study thus more accurately asks not merely whether infants can compute a conditional probability, but rather whether they can compute a large number of such probabilities rapidly and simultaneously.

Taken together, these studies show that adults, children, and infants are all capable of rapidly extracting from fluent speech the statistical patterns among adjacent sound sequences that distinguish words from accidental junctures, even without explicit direction of attention to the materials or the task.

2.2 Comparison with nonlinguistic patterns

This first type of statistical computation -- keeping track of the predictiveness of adjacent syllables in a stream of speech -- is, of course, merely one type of statistical computation that would be needed for learning a language. Higher level aspects of language, which are not limited to patterns among adjacent syllables, could not be acquired through such a computation (Chomsky, 1955, 1957), but instead would require more complex computations, involving the formation of classes, long distance dependencies, and hierarchical structure. Nonetheless, this first elementary type of computation might be extremely useful, not only for the segmentation of words in language, but also for the segmentation of elements in other domains and modalities. In a number of patterned domains, computing the consistency with which elements are in proximity to one another could help to determine which elements form a unit; as in language, elements which are consistently proximal to one another are likely to be part of the same unit, while elements occasionally proximal to one another are likely to be at the edges of two different units (and part of an accidental juncture). One example is that parts of a visual scene which move together or are co-linear are likely to be parts of the same object (Koffka, 1935; Kellman & Spelke, 1983). Another example is that sequences of musical notes or intervals which recur are likely to be a phrase or melody within a larger piece of music. We therefore decided to investigate next whether human learners -- infants as well as adults -- could perform the same sequential probability computations on musical tone sequences that they had performed on speech sequences.

2.2.1 Tone sequences

In our first study of statistical learning in a nonlinguistic domain, we (Saffran, Johnson, Aslin & Newport, 1999) created a tone segmentation task, analogous to our previous word segmentation task, in which the corpus was a continuous sequence of musical notes. Subjects listened to this continuous tone stream, and then were asked to identify which sequences of 3 tones sounded like familiar groups or melodies. The question of interest was whether adults and infants would perform in this task as they did in our word segmentation studies. In our first study, the exposure corpus was a sequence of pure tones whose statistical structure was identical to that used in Saffran, Newport and Aslin (1996). Adults exposed to this tone stream for 21 minutes performed identically to the adults exposed to the speech stream: they judged the 'tone words' (that is, sequences of tones with relatively high transitional probabilities) as more familiar than the 'tone non-words' (sequences of the same tones, but in an order they had never heard), and also as more familiar than 'tone part-words' (sequences of tones with lower transitional probabilities). Their performance on this task was indistinguishable from that on the analogous word segmentation task.

We also performed a study of infants (Saffran, Johnson, Aslin & Newport, 1999) identical to the second experiment in Saffran, Aslin, and Newport (1996), but using tones rather than speech. Eight-month-old infants were exposed to 2 minutes of a tone stream identical in structure to the speech stream used in Saffran, Aslin and Newport (1996). After familiarization, each infant was tested with 'tone words' and 'tone part-words,' and they showed differential listening times as in our studies of speech (see Figure 3). Thus the initial statistical learning mechanism we have studied appears to be capable of conducting similar computations on linguistic as well as nonlinguistic auditory materials.

2.2.2 Visual and visuomotor sequences

From the results reported thus far, it is possible that this computational mechanism is specialized for the learning of auditory sequential information. Additional research from our lab, however, suggests that similar sequence-learning abilities are also present for visual and for visuomotor patterns. In a senior honors thesis, Asaad (1998) created a visual analogue of the original Saffran, Newport, and Aslin (1996) study. A 4x3 matrix of blocks was displayed on a computer screen. To create a patterned array, individual blocks were lit one at a time (by turning from white to black for .2 sec), creating a continuous sequence of lit and darkened blocks that rapidly changed position within the matrix. To compare the results of learning patterns in this matrix

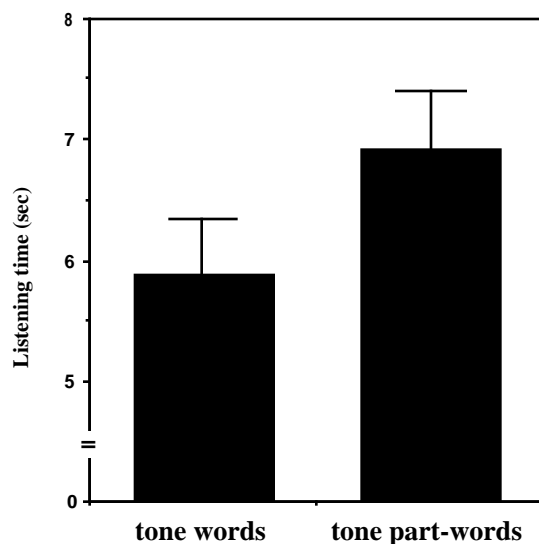


Figure 3. Mean listening times by 8-month-old infants to 3-tone sequences defined by either high (tone words) or low (tone part-words) transitional probabilities.

with our earlier data, each block in the matrix was statistically analogous to a syllable in our original work, and a sequence of three blocks was statistically analogous to a word. Over the presentation, then, some sequences of block patterns occurred with high probability, while others occurred only occasionally. Subjects viewed this continuous sequence of block patterns for 5 minutes. This was then followed by 2-alternative forced-choice test trials in which 3-block sequences analogous to words were paired with 3-block sequences analogous to non-words or part-words, and subjects were asked to choose which formed the better group. As in our speech studies, subjects reliably chose the consistent sequences (analogous to ‘words’) over the less consistent sequences. Fiser and Aslin (1999) have performed a similar study using sequences of visual shapes, and have also found similar results. Together these studies show rather clearly that the sequential statistical learning abilities we have been studying are not limited to the auditory modality.

Finally, Hunt and Aslin (1998) employed a serial reaction time (SRT) paradigm to examine yet another analogue of word segmentation, this time in the visuomotor domain. In this task (e.g., Nissen & Bullemer, 1987), subjects see an array of buttons on a horizontal display; each button is capable of being illuminated. Subjects are told to push each button when it is lit. The paradigm

permits the experimenter to introduce statistical structure into the presented sequence of lights, and to observe learning of this structure through reductions in reaction time at points corresponding to the predictiveness of the light sequence. In Hunt & Aslin's study, the continuous sequence of lights was formed from sets of triplets, analogous to the 3-syllable words in our speech studies. As in our speech studies, then, transitional probabilities between lights from within these triplets were high, while transitional probabilities between lights that spanned a triplet boundary were low. Subjects were tested over the course of 6 days and, as in our speech studies, show clear evidence of learning the statistical structure, as indicated by lower RTs to lights with high transitional probabilities, often with no explicit knowledge of this structure in the learning sequence. Thus, the SRT paradigm suggests that learners are capable of extracting the statistical relations embedded in visuomotor sequences, as they are in auditory sequences.

In short, human learners appear to segment elementary units out of large and apparently unsegmented streams, using the statistical structure embedded within this stream, across a number of modalities and paradigms.

2.3 Comparisons across species

The studies described thus far focus on human learners, who we know (from both laboratory and real-world behavior) are capable of quite extraordinary learning of sequential patterns in domains like language and music. However, precisely where these abilities diverge across species (and particularly compared to other primates) is not well understood. Studies of natural communication among primates (Cheney & Seyfarth, 1990; Hauser, 1996), and especially of primates' ability to learn full or simulated human languages (Gardner & Gardner, 1969; Premack, 1990; Savage-Rumbaugh, 1986), suggest profound and important differences between humans and nonhuman primates, but little research has investigated the underlying computational abilities of different species in order to shed light on where and why these differences in the learning of large domains occur.

In collaboration with Marc Hauser, we have begun to ask whether nonhuman primates can perform the same computations we have studied in human learners. As we have already noted, keeping track of sequential or conditional probabilities among pattern elements could be used in solving a number of segmentation problems. Since human learners can apparently perform this type of computation across a wide range of patterned materials, it is of interest to ask whether nonhuman learners can also perform this computation on at least some of the same materials. Moreover, as we begin to move toward more complex computations in humans -- investigating the formation of classes, long distance dependencies, and the like -- it is of interest to determine where humans and nonhumans begin to differ, as well as what types of materials (speech, tones, visual patterns) learners can perform these computations on. We have begun this line of work by asking whether tamarin monkeys -- a species

whose own natural calls are comprised of a series of call-parts, similar to certain bird songs (Cleveland & Snowdon, 1981; Hauser, unpublished) -- can perform the same task on sequences of speech syllables that we presented to human infants.

To be precise in our comparisons, we have used the same stimuli and similar testing methods with tamarins as with infants (though with a somewhat longer exposure period). Tamarins are presented in their home cage with a 21-minute exposure to the speech streams synthesized for infants (Saffran et al., 1996), and then are tested individually to see whether they look more frequently to presentations of words versus non-words and part-words. Surprisingly, like human infants, they look more frequently to both non-words and part-words than to words, showing that they too are able to keep track of the order in which syllables occurred in the exposure stream, as well as the probabilities with which syllables followed each other.

We are hoping in the near future to repeat this experiment with another species of nonhuman primate: vervet monkeys. Vervets are more advanced phylogenetically than tamarins, and more recently diverged from the human branch evolutionarily (Fleagle, 1988); in accord with this, they are generally considered to have more advanced cognitive functions than tamarins. On the other hand, their calls appear to be simpler in structure, with individual call types produced one at a time and without strings of call types produced in series (Hauser & Fowler, 1992). It is therefore not clear whether vervets should be expected to perform as well as tamarins on computations of sequential structure in our testing materials. Again, it is of great interest to observe whether the ability to perform such computations on sequential auditory materials will vary across species, and where species and domain differences will appear.

2.4 The next steps

Thus far we have shown that a fundamental computational ability -- the ability to compute the conditional probabilities with which elements in a linguistic or nonlinguistic array co-occur -- is remarkably widespread and general, and may provide an initial basis for segmenting the speech stream in language learning as well as for segmenting units in other patterned domains. We now turn to two important questions that arise in attempting to extend these results: First, what is learned from statistical analysis of this kind? Do learners acquire statistical generalizations, or do they use these statistical outcomes to acquire other (non-statistical) regularities? Second, what other, more complex computations can human learners perform? How far might this approach be extended in understanding how more complex properties of language are acquired, and what are the constraints on human learners' abilities to perform such computations? In the next section we turn to the first of these questions, concerning the products of statistical learning.

3. The Product of Statistical Learning: Statistics or Regularity?

Providing evidence that learners engage in statistical learning -- that they use statistical information computed over a corpus as part of the input to learning -- does not say what they do with this information or how the outcome of learning is represented. In our segmentation work thus far, we have not investigated this further question: We presume that learners do not merely store the transitional probabilities from the input corpus, but instead convert these graded statistics into *words* versus *word boundaries* (though there is no research to date on how such a process might work). However, in another line of work we have investigated a potentially related phenomenon, in which children exposed to probabilistically organized input use these statistical distributions to acquire 'rules.' These data come from a line of research observing children acquiring their native language solely from non-native parents.

3.1 Exposure to inconsistent morphology

Children virtually always acquire their primary language from speakers who are fully fluent in the language. This means that their input is highly regular and systematic: though it may be difficult for a learning theory to explain how the regularities are reconstructed from this input, there *are* rules and patterns underlying the linguistic strings to which the learner is exposed. A statistical approach to this type of learning, then, is an approach which hypothesizes that the rules may be helpfully revealed by computing statistics on the sequences and co-occurrences these rules create in the strings.

To better understand how this process works, we are studying children acquiring their primary language entirely from speakers who are themselves not fluent or native users of the language (Newport, 1999; Ross & Newport, 1996 and in progress; Singleton, 1989; Singleton & Newport, 1994 and in press). The input they provide to their children is thus truly statistical: morphological rules of the language are used only probabilistically, and many inconsistent errors are made. The outcome of acquisition in these circumstances shows that, while children may utilize input statistics to learn parts of their language, they do not merely reproduce these statistics in their own output. Rather, the strongest consistencies are sharpened and systematized: statistics are turned into 'rules.'

Our subjects are congenitally and profoundly deaf children who are acquiring American Sign Language (ASL) as their primary language. All are exposed to some form of ASL from birth or shortly after. However, because their families have imperfect proficiency in ASL (and the children have little or no other sign language exposure, and none which is highly fluent), their input to ASL is very reduced and inconsistent. In one line of work we have observed the children's acquisition of the morphemes of ASL verbs of motion. All of the parents use these morphemes to some degree, but vary in the consistency with which they use morphemes in their required contexts. When they err, they either omit the required morphemes or replace them with ungrammatical forms.

Studies of the children's acquisition of this morphology allow us to see the effects of input inconsistency on the acquisition of these structures.

Our first work on this topic has been a case study of a deaf child, whom we call Simon, acquiring ASL as his native language from his parents (Newport, 1999; Ross & Newport, 1996; Singleton, 1989; Singleton & Newport, 1994 and in press). Simon is the only congenitally deaf son of two deaf parents; both parents were first exposed to ASL in their late teens and now use it as their primary language, with each other and Simon. Simon attends a school where none of the teachers or other students knows ASL; the school uses a form of Signed English which does not contain the morphology or syntax of ASL which we have studied in Simon, and all other students in the school have hearing parents who do not know ASL. Simon's parents' friends are also non-native learners of ASL. In short, Simon's only input to ASL is from his parents. We have filmed this family's signing since Simon was 2 years old, but our first analyses focused on Simon's performance, compared with that of his parents, at a time when he should have completed his acquisition of ASL, at age 7:11.

Simon, his mother, and his father were each tested for their production of the morphemes of ASL verbs of motion. Simon's performance was also compared with that of deaf children of his age who have native signing parents, while his parents' performances were compared with adult native signers and late learners of ASL. In native ASL, verbs of motion involve producing a large number of morphemes in combination, and these verbs are therefore difficult for both late learners and young children to acquire. Each of the morphemes does, however, have a set of obligatory contexts, and is produced by native signers in a highly regular and systematic way.

Simon's parents sign like other late learners: they use virtually all of the obligatory ASL morphemes, but only at middling levels of consistency. On relatively simple morphemes (the movement morphemes of ASL), they average 65-75% correct usage. In contrast, Simon uses these morphemes much more consistently (almost 90% correct), fully equal to children whose parents are native ASL signers. Thus, when input is quite inconsistent, Simon is nonetheless able to regularize the language and surpass his input models. On more difficult morphemes (the handshape classifiers of ASL), where his parents were extremely inconsistent (about 45% correct), Simon did not perform at native levels by age 7; but even here he did surpass his parents.

To examine a greater range of inconsistency in linguistic input, Ross & Newport (in progress) are studying deaf children acquiring their sign language from *hearing* parents. These parents have learned to sign only slightly before their child, and their fluency in the language is often extremely limited. In one analysis we have compared these children and their parents to native signing families, on the same morphology as was studied in Simon. The subjects we have studied thus range from native input (for control subjects) through moderately consistent input (for subjects with deaf late-learning parents, like Simon, and also some with hearing late-learning parents) to extremely inconsistent input. A summary of the data for two of these children and their

parents on movement morphemes is shown in Figure 4. As can be seen, while the parents show quite variable levels of performance on these morphemes (note especially Sarah's mother), their children perform at native or near-native levels. On the more difficult handshape classifiers (not shown in Figure 4), the children do not attain fully native proficiency, but their consistency substantially exceeds that of their input.

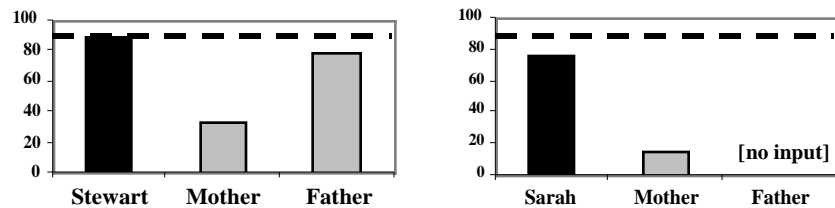


Figure 4. Percent correct production of ASL movement morphemes, for two children as compared with their hearing parents. The dotted line indicates performance of children receiving native input.

What is the mechanism by which children overcome such high degrees of inconsistency in their input? Two hypotheses seem possible. One hypothesis is that children know, innately, that natural language morphology is deterministic, not probabilistic, and acquire this morphology in accord with this knowledge. However, an alternative hypothesis is related to the statistical distribution of mappings between form and meaning in the input data, and, we propose, the tendency of children to sharpen and regularize these distributions as they learn. In the children's input, there is a particular pattern of consistency and inconsistency which we believe is relevant for learning: In each semantic context, one form (typically the correct form in native ASL) is used with some moderate degree of consistency, while errors are highly inconsistent. Thus, for example, 65% of the verbs referring to 'falling events' use the FALL morpheme. The remaining 35% do not use a single alternative form; instead this 35% is comprised of a scattering of other forms, with no one of these used with substantial frequency. This distribution -- one moderately consistent form, and others highly scattered and extremely inconsistent -- appears to be one in which learners will acquire only the major form, and thereby regularize the language as they learn it. In ongoing experimental studies we are investigating the precise patterns of input which produce this kind of 'creolization' effect, and also examining whether children, with particularly limited abilities to learn complex data, are also particularly apt to produce this type of outcome (Hudson & Newport, in progress).

This line of work, though different in method and focus than our statistical learning research, provides evidence that statistical distributions may play a

variety of important roles in acquisition. At the same time, the results suggest that learners do not merely learn and reproduce such statistical distributions, but instead may use these distributions to acquire rules and regularities. Indeed, the tendency of learners to use statistical data to form such rules may underlie certain phenomena of creolization and language change.

4. Non-adjacent Regularities: Constraints and Selectivities of Learning

The final issue we address in the present paper concerns how to extend the work we have done thus far into the more complex computations needed for learning the more complex structures of natural languages. We began our investigations of statistical learning by examining learners' ability to perform a first, elementary computation: what sequences of sounds (or tones, etc.) occur immediately after one another? To expand from here, we have begun, in a systematic way, to examine learners' abilities to compute more complex regularities, involving long distance dependencies, the formation of classes, and hierarchical structure; and also to examine how these more complex computational abilities might begin to differ across domains and species of learner. Our expectation is that, as we examine increasingly complex structures and computations, learners' abilities will begin to diverge across domains and species, with the most complex (or language-specific) of these computations reserved for human language learners. A question of great interest as we take these steps concerns where these divergences will appear, and whether we can understand some of the differences -- between adult and infant language learners, between humans and nonhumans, and for humans across different types of patterned materials -- as arising from basic differences among these learners in the types of computations they are able to perform over differing types of elements.

Our next step in this enterprise has been to examine learners' ability to compute non-adjacent regularities -- and indeed, their ability to compute precisely the same type of regularities we have already been examining, but between elements one unit apart, rather than immediately adjacent. To our surprise, we have quickly begun to see limits on learners' abilities to take this step.

4.1 Non-adjacent syllables versus non-adjacent segments

Many types of patterned regularities, in languages and in other domains, involve relationships between units which are immediately adjacent to one another in time or space. Indeed, one might expect that learners are better at performing computations between nearby units than between those which are quite distant. However, humans are clearly not restricted to learning such local regularities, and human languages contain many types of regular patterns among elements not immediately next to each other (Chomsky, 1955, 1957). In a new line of work we have therefore begun to investigate the types of non-adjacent

statistical regularities learners can compute (Newport, Aslin, & Calandra, in progress).

In our first study on this topic, we constructed an artificial language, identical in many ways to those we have already studied, but in which the regular patterns which define words (and which should permit word segmentation) are contingencies between syllables not next to each other. For example, consider the following words:

diykiytae	powkiygaa	keykiybuw	WORD STRUCTURE:
diyguwtae	powguwgaa	keyguwbuw	diy--tae, pow--gaa, key--buw
			kiy/guw

In this vocabulary the words are formed by a perfect 1.0 predictability between syllables 1 and 3; but syllable 2, which intervenes, varies over the same range (kiy/guw) for all words. By allowing words to follow one another at random, we can create a speech stream in which the transitional probabilities between every adjacent syllable are .33 to .5, both within and between words; but the transitional probabilities of 1.0 between non-adjacent syllables 1 and 3 of our words should still define the difference between within-word and between-word transitions. In a series of studies we asked whether subjects could learn to segment a stream made up of these words, using the same methods as in our previous studies. To our surprise, our results suggest that subjects cannot learn this regularity. Subjects run for 2 sessions show chance performance on 2-alternative forced-choice tests of words vs. non-words (items formed from the same syllables in random order). Indeed, one subject run for 10 consecutive days scored only minimally above chance on a word vs. non-word task, and no better than chance on a more difficult word vs. part-word task.

This failure to learn at first surprised us: the statistical regularities are precisely the same as, or even easier than, those which our subjects in earlier experiments learn with no difficulty, except that they apply to non-adjacent syllables. However, it happens that no natural language forms its word structure in exactly this way.¹ While many natural languages have word formation principles involving non-adjacent units, they work somewhat differently than the language we first built. Consider the following words:

dowkiybae	powgiytae	WORD STRUCTURE:
daakuwbey	paaguwtey	d-k-b, p-g-t
dowkuwbae	powguwtae	ow/aa, iy/uw, ae/ey
daakiybey	paagiytey	

In this vocabulary the words are formed by a 1.0 predictability between non-adjacent *segments*, rather than between non-adjacent *syllables*: d-k-b and p-g-t. This is much like the structure of words in Hebrew and Arabic, where the consonants form the stem of the word, and the vowels change to mark inflectional contrasts (singular versus plural, present versus past). By limiting

which words can follow one another, we can create a speech stream in which the transitional probabilities between adjacent syllables and adjacent segments is the same within and between words; but the transitional probabilities of 1.0 between non-adjacent segments 1, 3, and 5 (the consonants) should define the groupings into words. In dramatic contrast to our study of non-adjacent syllables, above, subjects exposed to this language with non-adjacent consonant segments show excellent learning.

Why should there be this difference in subjects' ability to learn different types of non-adjacent statistical regularities (corresponding neatly to whether the patterns occur or do not occur in natural languages), and what does it tell us about human learning? There are two ways of conceptualizing this result. One possibility is that learners innately know the particular principles by which natural languages are structured, and use this knowledge to learn linguistic (and language-like) materials (Chomsky, 1965, 1981). When faced with language-like materials which do not conform to the patterns of natural languages, they may not be able to learn them, particularly if the patterns are quite complex. Another quite different possibility is that there are more general constraints on learners' ability to acquire non-local patterns -- constraints which apply to a variety of materials, including language, sequential auditory materials, and perhaps visual and visuo-motor materials as well -- and that these constraints shape the patterns which languages evolve.

Our hypothesis about these particular results, which we are testing in ongoing work, is that the difficulty of learning non-adjacent regularities is ameliorated when the regularities apply to units of like kinds (and when the intervening unit is perceptually different). On this hypothesis, the organization of non-adjacent segments is learnable because the related units are all consonants, with vowels as the intervening items. In a new study we reversed this pattern, making the related units vowels and the intervening variable units the consonants (as in languages with vowel harmony, such as Turkish). In a final study in this series we will return to the case of non-adjacent syllables, this time making the related syllables 1 and 3 phonetically quite similar to one another and the intervening syllable 2 quite different, to see if this results in learnability. If such a language is as learnable as the non-adjacent segment languages, it would suggest that our hypothesized principles about distance and unit similarity might be the correct explanation of both our obtained learnability differences and the differences between the types of word formation principles which do and do not occur in natural languages.

4.2 Non-adjacent regularities in other domains and learners

We are currently in the process of asking whether infants and nonhuman primates can perform either of these types of computations, and also whether human learners show the same abilities and selectivities when the patterns are musical or visual. Given the results on human adults, our clear expectation is that human infants will, at some age, show the same pattern of performance as

adults on these language materials (we know by age 18 they perform as described). However, it is not clear whether very young infants will demonstrate either the ability to compute non-adjacent regularities or the prerequisite ability (for the non-adjacent segments task) to parse a syllable into consonant versus vowel segments. An important question we are studying, then, concerns the age at which each of these abilities appears.

What do we expect to find with nonhuman primates? These paradigms begin to involve complexities, in both distance (non-adjacency) and units of representation (segments rather than syllables), that may well exceed the abilities of any learners other than humans. Indeed, our initial runs on the non-adjacent segment paradigm with tamarin monkeys show chance performance; we are still in the midst of verifying this result with additional exposure, and then (if confirmed) attempting to delineate which of the complexities exceeds primates' processing abilities. Whatever the immediate outcome, our ultimate aim is to use these paradigms to determine the point at which these types of learners diverge; because we know tamarins do not acquire human languages, a failure at some point in our studies is the expected and interesting result.

5. Conclusions

As we have noted, we are in the early stages of a long enterprise, examining further how learners might be computing the various kinds of regularities of natural languages -- the formation of classes (cf. Maratsos & Chalkley, 1980; Mintz, 1996; Mintz, Newport, & Bever, 1995), long distance dependencies, and hierarchical structure -- as well as how various kinds of learners will compare in this regard, across domains and species. We are hopeful that, as we gain further insight into where learners and domains diverge, and where human learners in particular show selectivities in their computational abilities, we may begin to better understand certain aspects of how language regularities are acquired, as well as why those regularities may have some of their universal patterns and constraints.

One obvious caveat should be emphasized in closing: The types of statistical information we are investigating are clearly not the only kinds of information used in learning languages, and the computations we observe in learners are clearly not the only processes involved in language acquisition. This research is thus not intended to dispute the existence of other, quite different types of representations, processes of learning, or innate constraints on acquisition. At this early stage in our investigations, we believe an interest in these computational capacities is quite compatible, for example, with a view that they serve as the handmaiden of more substantive modules of knowledge and learning, or alternatively, with a view that they themselves direct the acquisition of structure. We hope that our research will provide some empirical answers to these questions.

Endnotes

* This paper was originally delivered by Elissa Newport as the Keynote Address at the conference. However, because most of the research was conducted collaboratively with Richard Aslin, we have produced the written version together. We are grateful to our collaborators on the work we have described: Jenny Saffran, Jenny Singleton, Danielle Ross, and Marc Hauser, and also to long-term collaborators Jim Morgan and Lila Gleitman, for their important contributions to our research and our thinking. This research was supported in part by NIH grant DC00167, NIH grant HD37082, and NSF grant SBR98-73477.

1. While some natural languages have somewhat similar structures (for example, the English *be-ing*), the cases we have found differ from our constructed cases in ways that might well be important for learnability. We are grateful to Katherine Demuth for her insightful comments on this topic.

References

- Asaad, P. (1998). *Statistical learning of sequential visual patterns*. Honors thesis, Department of Brain & Cognitive Sciences, University of Rochester.
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9, 321-324.
- Aslin, R. N., Woodward, J. Z., LaMendola, N., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan and K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum.
- Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93-120.
- Charniak, E. (1993) *Statistical Language Learning*. Cambridge, MA: MIT Press.
- Cheney, D., & Seyfarth, R. (1990). *How Monkeys See the World*. Chicago: University of Chicago Press.
- Chomsky, N. (1955/1975) *The Logical Structure of Linguistic Theory*. New York: Plenum Press.
- Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton.
- Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris.
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes* 13, 221-268.
- Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical approach to the bootstrapping problem for lexical acquisition. *Journal of the Acoustical Society of America* 95, 1570-1580.

- Cleveland, J., & Snowdon, C. T. (1981). The complex vocal repertoire of the adult cotton-top tamarin, *Saguinus oedipus oedipus*. *Zeitschrift für Tierpsychologie*, 58, 231-270.
- Fiser, J. & Aslin, R. N. (1999) Learning higher order statistics of simple shape-sequences. European Conference on Visual Perception, Trieste, Italy.
- Gardner, R. A., & Gardner, B. T. (1969). Teaching sign language to a chimpanzee. *Science* 165, 664-672.
- Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language* 20, 229-252.
- Harris, Z. S. (1951) *Methods in Structural Linguistics*. Chicago: University of Chicago Press.
- Harris, Z. S. (1955). From phoneme to morpheme. *Language* 31, 190-222.
- Hauser, M. D. (1996). *The Evolution of Communication*. Cambridge, MA: Bradford Books/MIT Press.
- Hauser, M. D., & Fowler, C. (1992). Declination in fundamental frequency is not unique to human speech: Evidence from nonhuman primates. *Journal of the Acoustical Society of America*, 91, 363-369.
- Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the Development of Language*. New York: Wiley.
- Hunt, R. H., & Aslin, R. N. (1998). Statistical learning of visuomotor sequences: Implicit acquisition of sub-patterns. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* 29, 1-23.
- Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress pattern of English words. *Child Development* 64, 675-687.
- Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology* 15, 483-524.
- Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt Brace.
- Maratsos, M., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's Language, Vol. 2*. New York: Gardner Press.
- Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G.T. M. Altmann (Ed.), *Cognitive models of speech processing*. Hillsdale, NJ: Lawrence Erlbaum.
- Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *American Journal of Psychology* 63, 176-185.
- Mintz, T. H. (1996). *The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories*. Unpublished doctoral dissertation. Department of Brain & Cognitive Sciences, University of Rochester.

- Mintz, T. H., Newport, E. L., & Bever, T. G. (1995). Distributional regularities of form class in speech to young children. In J. N. Beckman (Ed.), *Proceedings of NELS 25* (Vol. 2). Amherst, MA: Graduate Linguistic Student Association.
- Morgan, J.L., Meier, R.P., & Newport, E.L. (1987) Structural packaging in the input to language learning: Contributions of intonational and morphological marking of phrases to the acquisition of language. *Cognitive Psychology* 19, 498-550.
- Morgan, J. L. & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development* 66, 911-936.
- Newport, E. L. (1999). Reduced input in the acquisition of signed languages: Contributions to the study of creolization. In M. DeGraff (Ed.), *Language Creation and Language Change: Creolization, Diachrony, and Development*. Cambridge, MA: MIT Press.
- Nissen, M. J. & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology* 19, 1-32.
- Premack, D. (1990). Words: What are they, and do animals have them? *Cognition* 37, 197-212.
- Ross, D.S., & Newport, E.L. (1996) The development of language from non-native linguistic input. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (eds.), *Proceedings of the 20th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition (Vols. 1 and 2)*. Cambridge, MA: MIT Press.
- Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996) Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35, 606-621.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996) Statistical learning by 8-month old infants. *Science* 274, 1926-1928.
- Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A., & Barrueco, S. (1997) Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science* 8, 101-105.
- Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999) Statistical learning of tonal sequences by human infants and adults. *Cognition* 70, 27-52.
- Savage-Rumbaugh, E. S. (1986). *Ape Language: From Conditioned Response to Symbol*. New York: Columbia University Press.
- Singleton, J. (1989). *Restructuring of Language from Impoverished Input: Evidence for Linguistic Compensation*. Doctoral dissertation, University of Illinois.
- Singleton, J.L., & Newport, E.L. (in press) When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*.