


# Infant Pathways to Language

Methods, Models, and Research Disorders

Edited by  
John Colombo  
Peggy McCardle  
Lisa Freund

 Psychology Press  
Taylor & Francis Group

---

New York London

## 2

---

# *What Statistical Learning Can and Can't Tell Us about Language Acquisition*

RICHARD N. ASLIN AND ELISSA L. NEWPORT

### INTRODUCTION

In fall 1992, Jenny Saffran entered graduate school at the University of Rochester after having spent four years at Brown University working with Jim Morgan and an additional year as a research assistant with Sheila Blumstein. She already had an interest in the interaction between prosody and word segmentation (Morgan & Saffran, 1995) but sought a research topic for her first-year project. Ironically, in two independent consultations with each of us the suggestion was made to Jenny to read a chapter by Hayes and Clark (1970) in which adults were presented with sequences of noises that contained embedded subpatterns. Hayes and Clark reported that after listening to a continuous stream of noises, performance on a two-alternative forced-choice post-test was above chance for recognizing noise sequences that began and ended at subpattern breaks over noise sequences that began and ended within subpatterns. Four years later, after a number of fits and starts, two papers were published on adult (Saffran, Newport, & Aslin, 1996) and infant (Saffran, Aslin, & Newport, 1996) word segmentation from streams of synthetic speech. These two papers used the term *statistical learning* (SL)—from Charniak's (1993) description of algorithms in computational linguistics—to describe the psychological process by which the transitional probabilities from one syllable to another in the continuous speech streams could enable word segmentation and its complement, what Hayes and Clark referred to as *clustering*.

It would appear that Saffran, Newport, & Aslin (1996) and Saffran, Aslin, & Newport (1996) struck a chord in the language acquisition literature.<sup>1</sup> There have been dozens of follow-up experiments in the subsequent decade, and the term statistical learning is now used to describe a subfield of research on impressive feats of rapid learning in a variety of domains within language and visual processing.

adults have access to highly detailed representations of the input distribution. For example, Howes and Solomon (1951) showed that adults' word-recognition threshold in noise is a function of word frequency (estimated from written corpora). Levelt and Wheeldon (1994) showed that adults' picture naming latency is a function of syllable frequency. Vitevich and Luce (1998) showed that adults' lexical decision time is a function of phoneme frequency. And literally hundreds of psycholinguistic experiments have found the need to control stimuli for word and phoneme frequency, implicitly suggesting that subjects would be sensitive to these variables. Thus, there is clear behavioral evidence that mature users of a language have access to distributional information at the word, syllable, and phoneme level. And despite the "in principle" arguments for the implausibility of SL, there is demonstrable evidence that human learners overcome these impediments. How adults, and infants, do so will be discussed after reviewing recent findings on SL that extended the original studies by Saffran, Aslin, & Newport (1996) and Saffran, Newton, & Aslin (1996).

#### Statistical Learning since 1996

In our original studies of SL, we chose to study word segmentation because it is a tractable problem that must be solved by all language learners, and it is illustrative of a distributional learning mechanism that *may* apply more broadly (though we made no claim that it is sufficient). Saffran, Newport, Aslin, Tunick, and Barruecco (1997) showed that overt attention was not necessary for SL in preschoolers or adults, although some general level of attention is required (Toro, Sinnett, & Soto-Faraco, 2005). Saffran, Johnson, Aslin, and Newport (1999) showed that SL operates as well over sequences of nonlinguistic tones—in which the statistical structure mimics that of speech syllables—in infants and adults. Perhaps most importantly, Aslin, Saffran, and Newport (1998) investigated precisely what type of statistic is used in SL on speech streams. They showed that 8-month-olds could solve the word-segmentation problem even in the absence of trisyllable co-occurrence frequency differences. When the frequency of occurrence was equated for the words and part-words that were tested after exposure to speech streams, infants still discriminated words from part-words based on the higher-order statistic of transitional probability.<sup>2</sup> This does not mean that frequency is irrelevant to SL but rather that first-order frequency statistics are not the only class of statistics that infants can compute.

Fiser and Aslin (2002a) extended these results from the auditory domain to the visual domain by showing that simple visual shapes presented in temporal sequences (albeit at a slower rate than speech) enable adults to extract both first-order (frequency-based) and second-order (conditionalized) statistics. Kirkham, Slemmer, and Johnson (2002) showed that 2-, 5-, and 8-month-olds

are sensitive to first-order statistics in streams of simple shapes. Hunt and Aslin (2001) showed that SL applies to the visual-motor domain in a serial reaction time task with adults. Finally, Fiser and Aslin (2001, 2002b, 2005) moved the SL literature from the temporal to the spatial domain by showing that very similar mechanisms operate in multielement visual scenes, where both adults and 9-month-olds can bind together elements on the basis of first- or second-order statistics to form coherent perceptual "chunks."

Demonstrations of SL in nonhumans, both tamarin monkeys (Hauser, Newport, & Aslin, 2001; Newport, Hauser, Spaepen, & Aslin, 2004) and rats (Toro & Trobalón, 2005) clearly show that at least the simple aspects of SL are not unique to humans. However, only one study of nonhumans has employed the frequency-balanced design of Aslin et al. (1998), which evaluates whether learners show sensitivity to conditional probabilities rather than co-occurrence frequency. In that study the rats failed (Toro & Trobalón, 2005). It is not known whether monkeys are also limited to co-occurrence frequency computations or whether, like human infants, they can compute conditional probability statistics. This failure in rats, however, could be important for explaining why other species do not acquire complex systems like language.

Conditionalized statistics take into account differences in base rates that render first-order statistics less informative for predicting future events. For example, consider a case in which two elements (X and Y) occur frequently but just as often with each other as with other elements, while two other elements (A and B) occur rarely but always with each other. If one simply counted bigram frequency, the number of XY pairs could exceed the number of AB pairs, yet A is more predictive of B than X is predictive of Y. The conditional probability of Y given X and B given A captures this predictiveness better than the frequency of pairs XY and AB does. Interestingly, research on classical conditioning in the rat provides clear evidence of sensitivity to conditional probabilities (Rescorla & Wagner, 1972), but this paradigm places much less demand on memory and computational resources than typical SL paradigms, which require that learners keep track of the conditional probabilities relating many elements at the same time. Again, it is important to note that sensitivity to conditional probabilities does not imply that N-gram frequency of occurrence is unimportant. Highly frequent elements may, for example, serve as anchor points (or filters) that parse the input so that higher-order statistics can be computed over more limited subsets of the data. Several computational models of word segmentation (Swingley, 2005) and form-class learning (Mintz, 2003; Mintz, Newport, & Bever, 2002) use the strategy of operating over only the 200 or 300 most frequent elements.

Another important issue is what use is made of the statistics that are computed from a corpus of input. That is, what decision mechanism operates on those stored statistical values? It seems highly unlikely that statistics computed from

a corpus are retained in memory with sufficient fidelity that microdifferences could be used in making decisions about word boundaries or other properties of the underlying structure (e.g., is a transitional probability difference of 0.43 vs. 0.39 meaningful?). But given a reliable difference in some computed statistic, is the decision rule based on a local minimum or on a hard threshold? Saffran, Newport, & Aslin (1996) referred to “dips” in the transitional probabilities (TPs) at word boundaries and suggested that these dips could serve as a cue to word onsets. However, as noted by Yang (2004), such locally relative dips would not be present for single-syllable words (i.e., there would be low TPs both before and after the single-syllable word). However, a hard threshold (e.g., TPs below some criterion) would serve as a useful segmentation algorithm even for single-syllable words. Given the computational modeling of Swingley (2005), such a hard threshold is a viable mechanism, although to date we have no empirical evidence that such a mechanism actually operates in infants.

A general concern of artificial language (or artificial lexicon) studies is that they may not “scale up” to real corpora. That is, infants (and adults) may successfully use SL mechanisms to solve word segmentation and other problems when the language is extremely small but perhaps not when the problems reflect the size and complexity of real languages. This, of course, is a serious concern, but it may be offset in many cases by the myriad correlated statistical cues to structure that are present in real languages. We certainly recognized this potential problem in Saffran, Aslin, & Newport (1996): “Although experience with speech in the real world is unlikely to be as concentrated as it was in these studies, infants in more natural settings presumably benefit from other types of cues correlated with statistical information (p. 1928).”

#### How Is Statistical Learning Constrained?

The studies conducted since 1996 strongly suggest the existence of a robust SL mechanism in adults, infants, and at least two nonhuman animal species. Given the computational problem of explosive combinatorics (the curse of dimensionality), what enables a SL mechanism to operate without attempting to compute too many statistics (or just the wrong ones, those that mismatch the underlying structure)? Here we consider a set of constraints, some innate and some potentially learned from the input, that allow a powerful SL mechanism to be tractable in finite time.

#### *Preferred Units*

One problem for a SL mechanism concerns which basic units, out of the many available, are the ones on which statistical computations should be performed. Saffran, Aslin, & Newport (1996) presumed that the unit of analysis for initial word segmentation was the syllable. However, recent work with adults (Newport & Aslin,

2004) showed that statistics that reside at the segment (consonant or vowel) level, in the absence of contrastive syllable statistics, are sufficient for word segmentation.<sup>3</sup> Work in progress (Newport, Weiss, Wonnacott, & Aslin, 2004) suggests that adults and infants in fact rely primarily on segment information or on the alignment of segment and syllable information to solve the word-segmentation problem. When the statistics indicating word boundaries were at the syllable level and not the segment level, both adults and infants failed to segment words from streams of speech. By focusing on a subset of the potential types of units for statistical analysis, the learner reduces the combinatorics and simplifies the SL problem.

#### *Gestalt Principles*

A related set of constraints relies on Gestalt principles to bias what is learned. Some elements naturally tend to be linked perceptually, even without any learning process, and statistical relations among these may be most easily learned. For example, Baker, Olson, and Behrmann (2004) conducted a variant of the Fiser and Aslin (2001) studies employing multielement scenes; however, in contrast to Fiser and Aslin, in these studies there was no grid to make clear that each element was an isolated unit prior to learning. In this paradigm, adults are more likely to link together by statistical learning the elements in visual scenes that are connected by thin lines. Another example of a Gestalt constraint in the visual domain comes from Fiser, Scholl, and Aslin (2007). They used dynamic displays in which an object moved behind an occluder and then two objects emerged from the occluder. One object was a perceptually consistent continuation of the preoccluded object's trajectory, while the other was not. Subjects showed a preference for learning the temporal order statistics that conformed to the better continuation—that is, learning the sequence of shapes that appeared as connected.

More relevant to the language domain is a study by Creel, Newport, and Aslin (2004). They presented a sequence of tones as in Saffran et al. (1999), but in one condition the tones alternated between two different octave ranges. This induces a percept called auditory streaming (Bregman, 1990) in which attention is bistable between one or the other of the two octave sequences. Creel et al.'s stimuli had strong statistical relations between nonadjacent tones and weaker cues between adjacent tones. When the tones in both streams came from the same octave, adults learned the weaker statistics among temporally adjacent elements; however, when the streams came from different octaves, adults learned the nonadjacent statistics, favoring a grouping of elements within the same pitch range rather than those that were temporally adjacent. In other words, the perceptual bias of auditory streaming constrained SL. Although the foregoing studies show clear evidence for Gestalt constraints on SL, they are not particularly relevant to word segmentation because perceptual similarity among elements is not a reliable cue to words in natural languages.

### *Social/Attentional Cues*

Language typically occurs in a social context in which there are two or more talkers communicating and in which there is some visual information to complement the auditory information. Baldwin (1993) showed that 14-month-olds are more likely to treat a new label as referring to a novel object that is being looked at by the talker. Based on this work, Yu, Ballard, and Aslin (2005) conducted a study of adults' use of gaze information in a word-segmentation and word-learning task. Learners viewed videotapes of an adult describing the contents of a picture book in Mandarin. In one condition the videotape provided images of the picture book and the pages being turned as the Mandarin speaker described its contents. In the other condition the videotapes also contained information about the talker's eye movements to the picture book as the pictures were being described. The results showed a clear advantage on both word segmentation and word learning for the adults in the gaze condition over the no-gaze condition. These results suggest that, at least in adults, nonlinguistic cues such as eye gaze can aid in learning to segment speech and to attach sounds to referents.

### What Are the Limits of SL?

Like other examples of implicit (unsupervised) learning, SL was initially thought to involve minimal overt attention. This was based on the Saffran et al. (1997) study of preschoolers (and adults) who were not instructed to listen to the streams of speech but nevertheless learned to segment them. Of course, one can never know if learners are occasionally directing their overt attention to the speech streams even though, on average, they are not attending to them. Perhaps an occasional "monitoring" of a speech stream is sufficient to extract the underlying statistics. Toro et al. (2005) recently showed that a dual task reduced the performance of adults on a SL task. A more direct test of the role of overt attention was conducted in the visual domain by Turk-Browne, Jungé, and Scholl (2005). They had adults watch streams of simple visual shapes as in Fiser and Aslin (2002a), but they required their subjects to attend only to shapes in one color by having them look for a rare shape repetition. The shapes in the unattended color required no monitoring. The results showed that only statistics residing in the attended color stream were learned, even if there were intervening shapes in the unattended color.

Another limitation on SL is temporal adjacency, though (as described already) this interacts with other Gestalt principles. Newport and Aslin (2004) created streams of syllables in which the statistics forming word groupings resided in nonadjacent syllables. In these stimuli, the TP from syllable 1 to syllable 3 was 1.0, whereas the TP from syllable 1 to syllable 2, from syllable 2 to syllable 3, and from syllable 3 to syllable 1 of the next word was 0.5. Across a large series

of experiments, adults failed to learn the words in these nonadjacent syllable languages. In contrast to these negative results, Peña, Bonatti, Nespor, and Mehler (2002) reported successful learning of nonadjacent syllables under similar conditions, but using a different speech synthesizer, a different set of phonetic elements, and Italian- rather than English-speaking adults. It remains unclear why this discrepancy exists, but repeated attempts in our lab to observe learning of these nonadjacent syllable languages by 8- to 24-month-olds have failed over the past three years. In contrast, word groupings formed from the statistics among nonadjacent phonemic segments (consonants or vowels) are readily learned by adults. In this case, the perceptual similarity of the nonadjacent elements (being all consonants or all vowels) may overcome the preference for adjacency.

A final impediment to SL is the familiarity of the elements themselves. Many studies have demonstrated that statistical relations among speech sounds and tones are easily learned (at least in streams that contain adjacent statistics). However, Gebhart, Newport, and Aslin (2004) reported that statistical groupings among unfamiliar nonspeech sounds were very difficult to learn. It took adults more than 45 minutes of exposure across three sessions to extract the same simple triplet-based statistical structure among adjacent elements that was learned in our initial studies in 2 minutes with speech stimuli. This result is similar to findings on early word learning (Fennell & Werker, 2003; Stager & Werker, 1997; Swingley & Aslin, 2007) in which unfamiliar auditory word forms are much more difficult to associate with a picture of an object than familiar word forms.

The foregoing limits on SL—attention, adjacency, and familiarity—all concern the extraction of information from surface forms. But is SL limited only to surface forms? One of the important aspects of language that is not captured by the SL mechanism we have described so far is transfer from one set of surface tokens to novel tokens of the same underlying type (or category). This has been termed *rule learning* (RL) by Marcus, Vijayan, Bandi Rao, and Vishton (1999), who showed that 9-month-olds could learn a pattern such as ABB instantiated in 16 different exemplars and then generalize that ABB pattern to novel exemplars. Since SL operates on surface forms, Marcus et al. argued that RL and SL are qualitatively different mechanisms and suggested that RL necessarily involves encoding variables and relations, not the statistical properties of specific elements.

Additional work in the RL paradigm (Gómez & Gerken, 1999; Saffran & Wilson, 2003) provides strong support for the kind of learning with generalization reported by Marcus et al. (1999) and shows that it is not based on perceptual or phonetic similarity of the surface forms. However, evidence of RL does not eliminate SL as a contributor to the extraction and formation of rules. For

