Bootstrapping language:
Are infant statisticians up to the job?

Elizabeth K. Johnson

1. Introduction

Over the years, spoken language acquisition has attracted the attention of intellects from many disciplines. After much debate, two facts are apparent. On the one hand, it is clear that the ability to learn a language must be at least partially built in to our human psyche. Even chimps, our closest evolutionary cousins, fail to learn spoken language the way human children do. This is true even if they are cared for and spoken to as if they were human children (Terrace, Petitto, Sanders, and Bever 1979; see, however, Savage-Rumbaugh and Fields 2000). On the other hand, it is also clear that human language acquisition crucially depends on experience. Infants exposed to French learn French, infants exposed to Swahili learn Swahili (see also Curtiss 1977). But what must be inherited and what must be learned? And how do infants learn what they need to learn? All current models of language acquisition represent different answers to these very basic questions. Currently, many of the most popular models of early language acquisition are what could generally be described as distributional models. Most of these models place a heavy burden on the learning capabilities of prelingual infants. Some researchers working within the distributional framework strive to show how much children could accomplish in the absence of innate linguistic knowledge (e.g. Elman 1999), whereas others still emphasize the importance of linguistically-motivated constraints or expectations in statistical learning (e.g. Gervain, Nespor, Mazuka, Horie, and Mehler 2008; Mehler, Peña, Nespor, and Bonatti 2006; Yang 2004). The main focus of this chapter will be to examine distributional learning with the goal of better understanding what exactly we do and do not know about the ability of infants to extract linguistic generalizations from the speech signal. Although the discussion presented in this chapter is meant to apply to many levels of spoken language acquisition (e.g. phonology, morphology, syntax), the examples used to illustrate my points will be drawn primarily from the infant word segmentation literature. This is simply a reflection of my own research interests, as well as the fact that much of the recent work on distributional learning has been focused on potential solutions to the word segmentation problem. My primary goal in writing this chapter is to stimulate critical thinking about the key challenges facing different types of distributional models. In the end, I am afraid that this chapter will pose more questions than answers.

       The diversity of languages that a child must be equipped to learn is immense. For example, Finnish has a rich inflectional morphology, highly predictable stress, and vowel harmony whereas English, by contrast, is morphologically impoverished, exhibits difficult to predict stress, and has no vowel harmony. Likewise, German is an intonation language with very long words whereas Vietnamese is a tone language with relatively shorts words. And Italian is a syllable-timed language with a simple syllable structure and clear syllable boundaries whereas Russian is a stress timed

language with complicated syllable structure and less clear syllable boundaries. The examples provided here just begin to scratch the surface in terms of possible sound structure patterns exhibited by the thousands of languages spoken by humans around the world. Remarkably, no matter what their characteristics, all natural languages seem to be acquired equally readily by normally developing children. This poses a potential problem for experience-based accounts of language acquisition. How could a single learning strategy account for how children are capable of mastering (or at least gaining a handle on) all of these different types of linguistic systems with relative ease in the first few years of life?

In the decades prior to the mid 1990s, most researchers would have avoided this dilemma by arguing that the key to language acquisition lies in the linguistic knowledge all humans inherit. Through our innate endowment, we are born knowing all of the possible rules that can exist in a language, therefore reducing language acquisition to the process of simply working out which rules apply to the language currently being learned (Chomsky 1957). In the late 70's and early 80's language researchers who opposed this Chomskian perspective had little ground to stand on, since there was no other fully convincing explanation for how children could acquire the wide varieties of language in the world so quickly. The rules and structure of language were viewed as too complex and unpredictable to be learned, especially by young infants with limited cognitive resources. And early speech perception research initially seemed to support the notion that much of language knowledge was innate, as study after study suggested that neonates were born being able to discriminate virtually all possible phoneme contrasts in much the same way as adults (see Jusczyk 1997, for review).

A brief article published in Science in 1996 marked the beginning of a dramatic shift in the way language researchers viewed early language acquisition (Saffran, Aslin, and Newport 1996). In this study, the authors set out to demonstrate that infants could indeed acquire some complicated aspects of language using experience-dependent learning mechanisms. They chose to investigate the word segmentation problem because the authors viewed word segmentation as a tractable problem faced by all language-learning infants. In this study, eight-month-old infants were exposed to an artificial language containing 4 trisyllabic CVCVCV words. The made-up words were strung together in random order with the stipulation that no single word ever followed itself in immediate succession. The speech stream was produced with a flat intonation and contained no pauses between words. Since the words in this language had no meaning, and the language lacked any prosodic or pause cues to word boundaries, words only existed in the sense that the syllables co-occurred consistently. That is, the words only existed in a statistical sense. Remarkably, after a mere two minutes of exposure, infants listened longer to trisyllabic sequences of syllables that spanned syllable boundaries over the trisyllabic sequences that corresponded to statistical words. This finding revealed that 8-month-olds are capable of segmenting words from an artificial language using statistical cues alone (in the remainder of this chapter, we will often refer to this particular type of statistical segmentation cue as a syllable distribution cue). In the closing words of the authors:

'Our results raise the intriguing possibility that infants possess experience-dependent mechanisms that may be powerful enough to support not only word segmentation but also the acquisition of other aspects of language… the existence of computational abilities that extract structure so rapidly suggests that it is premature to assert a priori how much of the striking knowledge base of human infants is primarily a result of experience-independent mechanisms. In particular, some aspects of early development may turn out to be best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge. If this is the case, then the massive amount of experience gathered by infants during the first postnatal year may play a far greater role in development than has previously been recognized.' (Saffran, Aslin, & Newport, 1996)

This landmark study was not the first study to discuss how statistical information could be used to extract linguistic generalities from language (e.g. Harris 1955), nor was it the first study to implement the use of an artificial language to study language learning in human adults (e.g. Hayes and Clark 1970; Morgan and Newport 1981; Valian and Coulson 1988) or infants (e.g. Goodsitt, Morgan, and Kuhl 1993). It was not even the first to suggest that syllable distribution cues could help listeners find word boundaries (e.g. Morgan and Saffran 1995). Nonetheless, this remarkable study struck a chord in the field, making two very timely and pivotal contributions that literally reshaped the way researchers thought about (and studied) early language acquisition. First, the Saffran et al. study firmly established the use of artificial languages as a mainstream approach to studying infant language development (see Gomez and Gerken 2000, for review). Second, this study suggested that infants possessed truly powerful statistical learning mechanisms that might enable them to learn far more from the ambient environment than researchers had previously imagined possible. Thus, this one study not only inspired language researchers to develop new hypotheses concerning the development of language skills in children, it also gave infant researchers a very powerful tool for experimentally testing these hypotheses (see Bates and Elman 1997, for discussion).

      In the 15 years since the initial publication of Saffran, Aslin, and Newport (1996), scores upon scores of infant artificial language learning studies have been published (see Aslin and Newport 2008, for a review). Many of these studies have built directly upon Saffran et al.'s initial work, continuing to explore how infants might learn to segment words from speech. Studies have shown that the statistical learning mechanisms that infants use to segment words from an artificial language appear to be domain general since infants apply them to sequences of tones and visual objects just as readily as they apply them to sequences of syllables (e.g. Saffran, Johnson, Newport, and Aslin 1999; Kirkham, Slemmer, and Johnson 2002). And infants have been shown to track not only simple co-occurrence frequencies between syllables (as tested by the original Saffran et al., 1996, study), but also conditional probabilities between syllables (Aslin, Saffran, and Newport 1998). Rats, on the other hand, appear to succeed only at tracking the simpler co-occurrence frequencies (Toro and Trobalon 2005). Some have suggested that rats' inability to track conditional probabilities might in part explain why they do not develop human-like language (Aslin & Newport, 2008). Amazingly, infants have also been shown to readily extract backward transitional probabilities (Pelucchi, Hay, and Saffran 2009b), as well as non-adjacent

relationships between segments in an artificial language (Newport and Aslin 2004). English-learning infants have even been shown to track transitional probabilities between syllables in highly constrained but nonetheless natural Italian speech (Pelucchi, Hay, and Saffran 2009a).

In short, the evidence that infants rely on transitional probabilities between syllables to segment words from speech is quite convincing. If you pick up a recently published undergraduate textbook, chances are you might just read that infants solve the word segmentation problem by tracking transitional probabilities between syllables (e.g. Byrd and Mintz 2010). Specifically, most infant speech perception researchers suppose that infants first segment a limited number of words from speech by tracking transitional probabilities between syllables in the ambient language and positing word boundaries at 'dips' in the these transitions. Then, by analyzing the sound properties of these words, infants could deduce language-specific cues to word boundaries in their language, such as lexical stress placement and phonotactic constraints. Indeed, transitional probabilities between syllables seems an ideal solution to the word segmentation problem because such a strategy relies entirely on bottom-up information, circumventing the need to resort to built-in linguistic constraints to explain how infants manage to segment words from speech.

The impact of the original Saffran et al. (1996) paper goes far beyond the study of word segmentation. Infants' success at using transitional probabilities between syllables to extract words from an artificial language helped increase the popularity of connectionist models of cognitive development, and has inspired dozens of studies examining the possibility that infants use statistical mechanisms to learn virtually every other level of language structure. On the segmental level, tracking the distribution of phonetic contrasts in speech has been argued to help infants work out the phonemic inventory of their language (e.g. Maye, Werker, and Gerken 2002; Peperkamp, Calvez, Nadal, and Dupoux 2006). It may also help infants learn the patterning of speech sounds with respect to word boundaries (e.g. Chambers, Onishi, and Fisher 2003; Seidl and Buckley 2005). On the morphosyntactic level, young infants have been argued to track non-adjacent dependencies between linguistic elements, an ability that could help them learn morphosyntactic patterns such as the is –*ing* dependency in English (e.g. 'is singing' versus 'can singing'; Gómez and Maye 2005). Similar strategies could be used to work out the word class of newly learned words (e.g. e.g. Mintz 2003; Monaghan, Christiansen, and Chater 2007). On the syntactic level, there is some evidence that syntactic grammars are learnable via statistical learning mechanisms (e.g. Saffran 2001; Saffran and Wilson 2003; Thompson and Newport 2007; see however De Vries, Monaghan, Knecht, and Zwitserlood 2008). And on the semantic level, research has suggested that language learners could use simple cross-situational statistics to work out the referent of words in an artificial language (e.g. Smith and Yu 2008).

The sheer amount of information that artificial language learning studies suggest infants might be able to glean from their language input is staggering, and would have been truly unimaginable 20 years ago. In the early days after the initial publication of Saffran, Aslin, and Newport (1996), many researchers saw statistical cues such as transitional probabilities between syllables as just another cue (amongst many cues) that infants might have in their arsenal to use in the task of word

segmentation (e.g. Jusczyk, Houston, and Newsome 1999). And many researchers found it easier to accept that these statistical cues would be used for more 'low-level' problems such as word segmentation rather than the higher-level acquisition challenges such as syntax and word form to referent mapping. But in more recent years, the general zeitgeist of the language acquisition field has truly changed. In the past decade and a half, distributional learning has clearly replaced innate knowledge as the most widely accepted explanation for the better part of children's language learning prowess. Indeed, in some cases, reliance on any sort of innate phonological knowledge has been presented as a weakness in models of developmental speech perception (e.g. see discussion on 'external components' in Monaghan and Christiansen 2010).

At this point in time, when distributional models of early language acquisition are continuing to climb in popularity, a useful exercise might be to sit back and take stock of both the strengths and weaknesses of these models. As has been summarized in this chapter thus far, the strengths of this type of theoretical approach have been outlined in the literature time and time again and are in large part stunningly (and elegantly) clear. Statistical learning has been argued to provide a universal bottom-up strategy for pulling words out of speech. These models can also explain how infants might begin to learn nearly all other levels of linguistic structure. Since the great majority of descriptions of statistical models in the literature are extremely positive, I would like to take the larger part of the remainder of this chapter to take a more critical stance, and discuss the flip side of the coin. What are the drawbacks to these models? What aspects of these models are under-specified? What behavioral data do these models have difficulty explaining? What assumptions do these models make? Are the assumptions valid? And what does the term 'distributional model' mean? Perhaps the term is so underspecified that it carries little meaning at all. Note that I do not dwell on the problem areas of distributional models because I think these models are weak. On the contrary, the models are quite powerful as evidenced by the impact they have had on the field. And I think that the statistical learning mechanisms that have been revealed in the past 15 years of research are truly exciting, and distributional models are here to stay. Nonetheless, I think it is important to consider some of the challenges still faced by these models because identifying and testing weaknesses in popular models of psychological behavior is an important part of the scientific process. Only by testing and challenging currently popular models can our theories move forward and continue to grow and improve to meet the challenges posed by new data and possible new alternative theories of early language acquisition.

2. Five Challenges Faced by Distributional Models of Language Acquisition:

2.1. Does it scale up?

As mentioned above, language scientists have known for many years that statistical patterns in the input are linked to the linguistic structure of languages. However, these statistical patterns were thought to be too inconsistent and too complex to be acquired by young infants and children. Thus, researchers assumed that children must possess substantial innate knowledge about possible language structures in order to

acquire language as quickly and seemingly effortlessly as they do. But now we have substantial experimental evidence to suggest that infants are far better at picking up statistical patterns in the input than we initially thought. Using the artificial language learning paradigm, researchers have created miniature languages containing patterns that are reflective of linguistically relevant statistical patterns in natural language. Infants' ability to learn many of these patterns within a very short period of time suggests that these same patterns can be learned from natural language input.

However, there is a problem with using these studies as evidence for the ways in which real language learning occurs. The artificial languages used in many infant artificial grammar learning studies are so simplistic that one must wonder whether the ability to learn a pattern in these artificial languages will necessarily scale up to the challenge of learning a pattern in natural language input. Consider, for a moment, the language used in the original Saffran et al. (1996) study on word segmentation. This language contained four trisyllabic words, and each syllable had a simple CV syllable structure. All syllables were clearly enunciated, fully stressed, and of equivalent duration. And no syllable occurred in more than one word. In other words, the language was not very speech-like at all. The authors admit the language is simplistic, but also note that natural language contains many other cues to word boundaries aside from transitional probabilities. Thus, they never claim that the tracking of transitional probabilities alone could solve the whole segmentation problem.

Later publications, however, make stronger claims regarding infants' reliance on transitional probabilities between syllables to segment words from speech. For example, Thiessen and Saffran (2003; 2007) argue that transitional probabilities between syllables are perhaps the first cue used by children to segment words from speech. These cues could then be used to learn other important cues to word boundaries such as the placement of lexical stress. Other studies have argued that phonotactic cues to word boundaries could be learned from transitional probabilities, once again suggesting that transitional probabilities between syllables are the first and most important cue infants use to begin tackling the word segmentation problem (Sahni, Seidenberg and Saffran 2010). The clear assumption in these studies is that infants can track transitional probabilities between syllables in natural speech in much the same way that they can track them in a highly simplified artificial language (i.e. in the absence of any other cues to word boundaries). Analogous arguments have been made for the acquisition of other levels of linguistic knowledge (e.g. Onnis and Christiansen 2008). But is it valid to assume that the ability to track linguistic patterns in a highly simplified language will necessarily scale up to the challenge of natural language?

There are a few ways to test this assumption. In a perfect world, we would know the transitional probabilities between syllables in natural language (or, for example, the statistical likelihood of a word being uttered when a particular object is in sight). Testing the ecological validity of distributional models of word segmentation (or the acquisition of any other level of linguistic structure) would simply involve seeing whether infants were sensitive to this information in the environment. But for numerous obvious reasons, such an experiment is clearly impossible (see Johnson and van Heugten, in press, for discussion). Who is to say precisely how many times a given

child has heard or experienced a particular pattern in the input? And if they did, who is to say the information was attended to?

The next best option for testing the ecological validity of these models might be to present infants with an unfamiliar natural language and see whether they use transitional probabilities to locate likely word boundaries. Such an experiment has in fact been carried out (Pelucchi, Hay, and Saffran 2009a; 2009b). In this cleverly designed study, English-learning 8.5-month-olds were presented with an Italian passage containing repeated tokens of two trochaic bisyllabic words (e.g. fuga and melo). The infants had never heard Italian before, and the syllables in this target word did not occur elsewhere in the passages. Therefore, the researchers could be assured that at least for the English-learning infants they tested, the transitional probability between the two syllables in this word were 100%. The passages contained two additional target words that occurred just as often as the first two target words (e.g. bici and casa). But importantly, the initial syllables of these words also occurred as monosyllabic words in the passage (e.g. ca and bi). For these second set of target words, the transitional probability between the syllables was .33. Amazingly, the English learners not only segmented the Italian target word with a high transitional probability from the passage, they also differentiated between the high transitional probably words and the low transitional probability Italian words. The authors concluded that infants can track transitional probabilities from speech in the face of naturalistic speech sound variation.

This study represents the strongest support to date for the notion that infants use transitional probabilities between syllables to extract words from natural speech. Note, however, that although this study does demonstrate that English learners can track transitional probabilities in Italian, it does not demonstrate that infants use transitional probabilities to segment their first words from real world language. One reason this is the case is because the training phase in the Italian study was artificially constructed, so the statistical cues would be particularly salient to infants. To put it more simply, by creating a stream of Italian speech that contained many repetitions of the target words produced by a single speaker in a short period of time and by ensuring that the transitional probabilities between the syllables in the high transitional probability and low transitional probability target words were markedly different (100% and 33%, respectively), this study employed an unnatural sample of a natural language. In English, the dips in transitional probabilities would likely be far less dramatic. And in other languages with many monosyllabic words, such as Thai, it seems likely that the syllable transition cues to word boundaries would be even weaker than in English. Moreover, the choice of language materials used in the Pelucchi et al. study may have impacted the findings, since tracking syllables is probably relatively easy in a syllable timed language with clear syllable boundaries such as Italian compared to a stress timed language with less clear syllable boundaries such as Russian or Hungarian. For this reason, it would be interesting to repeat this study with Russian rather than Italian stimuli (or to test infants learning a language other than English on their ability to track transitional probabilities in English). It would also be interesting to carry out some additional studies to ensure that the infants were truly segmenting the whole words (rather than part of a word) from speech (as is sometimes the case in artificial language studies; see Juscyzk, Houston,

and Newsome 1999; Johnson, 2005; Johnson and Jusczyk 2003b; for a related discussion). Last, but certainly not least, since the researcher's goal was to show that transitional probabilities could be tracked in natural speech, they did not control for intonation and other prosodic grouping cues that likely provided an additional cue to word boundaries in natural language. Thus, for all of these reasons, despite representing a very important and exciting step forward in the study of infant statistical learning abilities, these Italian segmentation studies do not entirely answer our question of interest. They show that exaggerated transitional probabilities can be tracked in natural Italian speech containing many cues to word boundaries, but they do not show that infants can track the naturally occurring transitional probabilities between syllables in natural language (which are presumably much less clear than the 1.0 and .33 ratio used in the Italian study), and use these cues to bootstrap all other language-specific segmentation cues.

A third possible way to test whether infants' ability to track transitional probabilities between syllables can scale up to the challenge of natural language would be to take an artificial language and systematically make it a bit more like a natural language in order to gauge the likelihood that infants can track distributional information in natural language. By doing so, one could begin to address the 'scaling problem' while also maintaining the exquisite experimental control offered by artificial language studies. This is precisely the approach initially taken by Peter Juscyzk and myself. We created a natural speech analogue of the original Saffran et al. (1996) artificial language. We then added a speech cue that either conflicted or aligned with the statistical cue in the artificial language. In both cases, 8-month-olds extracted the words aligned with the speech cues, suggesting that at least at 8 months infants appear to weigh speech cues to word boundaries more heavily than syllable transition cues (see also Thiessen and Saffran 2003; Johnson and Seidl 2009).

In a more recent study, rather than adding an additional speech cue to the speech stream, we instead focused on removing regularities other than the transitional probabilities between syllables that could have been helping infants extract words from simple artificial languages (Johnson and Tyler 2010; see also Johnson and Jusczyk 2003). We had many regularities to choose from, including syllable structure, word length, and uniform phoneme length. There are at least two ways in which these regularities could have helped infants track transitional probabilities in the original artificial language. First, by excluding all other natural language variability, highly simplified artificial languages may highlight syllable distributions as an important pattern in the input (much more so than they are highlighted in natural languages). Second, once infants commence transitional probability tracking in a simplified artificial language, sound structure regularities could serve as additional segmentation cues boosting initial distribution-based hypotheses regarding likely word boundaries (see Sahni, Seidenberg and Saffran 2010, for a related discussion). As a first step to exploring the role of variability in infant's ability to track transitional probabilities, we chose to focus on word length. We created two sets of artificial languages: one with 4 bisyllabic words (the uniform word length language) and one with 2 bisyllabic words and 2 trisyllabic words (the mixed word length language). Despite the fact that both the uniform word length and mixed word length languages contained equally strong statistical cues to word

boundaries, both the 5.5- and 8.5-month-olds we tested only succeeded in segmenting words from the uniform word length language. That is, infants succeeded if they were exposed to the language containing both syllable transition cues to word boundaries and word length regularities, but they failed if they were exposed to the language containing only syllable transition cues to word boundaries. Since all natural languages contain words of varying length, this study gives one reason to fear that infants' ability to track transitional probabilities in a highly simplified artificial language might not scale up to the challenge of natural language.

Note that there are studies that have shown that infants can use transitional probabilities between syllables to segment words from a language containing words of variable length, however, these studies have involved artificial languages that contained both statistical and prosodic cues to word boundaries (Thiessen, Hill, and Saffran 2005). Even adults appear to have great difficulty using syllable distribution cues alone to find word boundaries in artificial languages containing words of varying word lengths (Tyler and Cutler 2009). Given how difficult it appears to track transitional probabilities between syllables (even in a simple artificial language), one is left wondering whether perhaps infants have an easier solution up their sleeve (e.g. see discussion of utterance level prosody in Johnson and Seidl 2008; Johnson and Tyler 2010). Further research in this area is clearly needed.

2.2. Which unit? And once you pick a unit, which statistic?

Artificial language learning studies are not the only line of evidence in support of distributional models of language acquisition. Computational models demonstrating what information statistical learning mechanisms could, in theory, extract from the speech signal also provide another important line of evidence for distributional learning theories. But all computational models have to make some assumptions regarding what sort of information infants can perceive and process, and calculating statistics requires having some unit over which to do the calculations. It is impossible to implement a computational model of early language acquisition without making assumptions about relevant units and types of calculations.

At first blush, choosing a unit to track in the input sounds like a trivial challenge for most models of early language development. The infant speech perception literature provides reason to choose either the syllable or the phoneme as a basic unit to be tracked (see Juscyzk, 1997, for review; see, Vihman and Vihman, in press, for a different perspective). Indeed, models such as PARSER (Perruchet and Vintner 1998) and Swingley's segmentation model (2005) have chosen the syllable as the basic unit, whereas models such as PUDDLE (Monaghan and Christiansen 2010) and BOOTLEX (Batchelder 2002) have chosen the phoneme as their basic unit. Later on, as children begin learning the words of their language (by tracking statistical relationships between phonemes and/or syllables), models typically assume that words (and even bound morphemes) serve as the units over which additional computations are performed.

However, even a very simple assumption like choosing a unit over which a child will perform calculations is not without its controversies. How did infants extract this unit in the first place? Was the knowledge inborn? Or is it simply a

product of the auditory system? Indeed, the classic infant speech perception story is that children are born perceiving syllabic units as well as nearly all segmental contrasts present in the world's languages (Jusczyk 1997; Werker and Tees 1984). However, we now know that this is a bit of an oversimplification. First, most newborn speech perception studies have only been carried out under ideal listening conditions involving the presentation of clearly articulated isolated syllables or words. We still have a lot to learn about how infants perceive speech contrasts in different contexts (e.g. Bortfeld and Morgan 2010) or distracting environments (e.g. Newman 2009). Moreover, not all speech contrasts are initially perceived equally easily (see Juscyzk 1997, for review). We now know that acquiring the phoneme inventory of one's language does not simply involve losing sensitivity to contrasts that do not signal changes in word meanings in the native language, it also involves a fair amount of tuning into sound contrasts that were initially difficult to perceive (Nurayan, Werker, and Beddor 2010; Tsao, Liu, and Kuhl 2006). And some learning environments surely make this more difficult to accomplish than others (Sundara and Scutellaro in press). Moreover, categorizing syllables and segments (and even words) can be complicated by various language-specific coarticulatory and suprasegmental phenomenon (e.g. Curtin, Mintz, and Christiansen 2005). What this means is that regardless of whether you choose a phoneme or a syllable (or even a word) as your basic unit, the information a young infant pulls from the speech signal will obviously be very different from the information an adult (or a highly trained transcriber listening to speech recordings) will pull from the speech signal. This is especially true since many models suggest that infants are tracking these statistics at around 6 months of age or so, before infants have mastered the segmental inventory of their language or even learned whether they are learning a tone language or an intonation language (Mattock, Molnar, Polka, and Burnham 2008).

The situation becomes even more complicated when you consider the differences between broadly transcribed speech and spoken language. Spoken and broadly transcribed language do not just differ in the amount of information they carry, they also differ in the type of patterns they highlight. For example, spoken language contains rich prosodic grouping cues and immense fine-grained acoustic-phonetic variation in the realization of speech sounds. No two realizations of the same word are ever the same. And there is evidence to suggest that infants are sensitive to even more fine-grained information than this from very early on (Johnson 2003; 2008; Johnson and Juscyzk 2001; McMurray and Aslin 2005). Transcriptions, on the other hand, typically represent each realization of a word in an identical fashion. Representing the input in such a categorical fashion assumes that infants have already solved the many-to-one mapping challenges cause by the lack of invariance in the speech signal. It is worrisome that many of the challenges faced by an infant hearing spoken language (such as dealing with connected speech processes like graded assimilations, resyllabification, casual speech reduction, and stress shifts) are typically not dealt with in computational models. Is it fair to assume that infants know what type of acoustic variation is linguistically important, and what type is not? Some studies have suggested that infants are adept at coping with invariant productions of speech sounds (Jusczyk, Pisoni, and Mullennix 1992; Kuhl 1979; van Heugten and Johnson 2010), other studies have suggested that infants have a fair bit of difficulty

recognizing commonalities between acoustically distinct realizations of syllables and words (Bortfeld and Morgan 2010; Houston and Jusczyk 2000; Schmale, Christià, Seidl and Johnson 2010; Singh, Morgan and White 2004). Only further work in this area will be able to clarify the situation.

Some computational modelers have simply dismissed concerns over the fact that fine-grained information was not incorporated in their models, claiming that such information could only help children find information in the speech signal (e.g. Batchelder 2002). In other words, their models are conservative estimates of children's performance because they have not incorporated all of the useful information present in real speech. But statements such as these assume an incredible amount in terms of the speech perception and processing abilities of young infants. Is this the best strategy for researchers who claim to be designing models that reduce the amount of built in knowledge children have to be equipped with? Other modelers have tried to correct for some of the most basic differences between spoken and written language by replacing some dictionary pronunciations with the pronunciations we typically see in specific speech environments, or adding some random variation in the realization of words (Cairns, Shillcock, Chater and Levy 1997; Monaghan and Christiansen 2010; Swingley 2005). Although such an action is a step in the right direction, the gap between spoken and transcribed language is still enormous. And connected speech processes, such as assimilation, are often (if not always) gradient (e.g. Gow 2002). Even if modelers were to use some sort of complex mathematical vector in their models that carried all of the variation (some useful, some likely distracting) contained within the speech signal, this would still not solve the problem of units in computational models. How would infants know to categorize and make use of all of this variation? How would they know to calculate statistics over phonological units without first knowing the units? Very recent work has shown that adding naturalistic variation to a corpus-based model of segmentation seriously hinders the model's performance (Rytting, Brew and Fosler-Lussier 2010), but it seems that a better understanding of how infants perceive and process speech variability is necessary before such conclusions can be extended to human segmentation behavior. It is possible that the same variability that hinders the performance of computational models of word segmentation may in fact help young human listeners (see Johnson 2003; Rost and McMurray 2009, for related discussions).

Setting aside the difficulties of choosing the unit over which infants should calculate their statistics (and the question of how infants overcome the many-to-one mapping problems involved in identifying these units in an adult-like manner), there is another enormous challenge in clearly defining exactly what type of statistical formula children may be working out in their heads. Even in the simple artificial languages specifically designed to address whether infants are capable to tracking a specific type of statistic, there are often multiple ways the patterns in the input could be tracked (e.g. Bonatti, Peña, Nespor and Mehler 2006; Endress and Mehler 2009; Perruchet and Desaulty 2008; Perruchet, Peereman & Tyler 2006). For example, do humans compute primarily forward or backward transitional probabilities? Are they learning rules or patterns? Are they calculating simple statistics or more complex statistics? Are all segments (and syllables) treated equally in these calculations? Are

adjacent relationships easier to learn than non-adjacent relationships? How big does a 'dip' in a transitional probability have to be before a listener hypothesizes that a word boundary has occurred? The list goes on and on. And of course, the type of calculations you assume infants can perform, and the criteria along which you evaluate the success of any given distributional learning model, can have an enormous impact on the conclusions you draw (e.g. Yang 2004).

One unresolved mystery in the literature that I have been interested in revolves around the calculations infants perform over the input they receive from artificial language training. In all studies examining adults' ability to segment words (or tone words) from an artificial language containing no cue to word boundaries other than transitional probabilities between syllables, adults consistently perceive partwords consisting of the last two syllables of one word and the first syllable of another as more 'word-like' than partwords consisting of the last syllable of one word and the first two syllables of another (e.g. all else being equal, if *bupada* and *golatu* are statistical words, then *padago* will sound more word-like then *dagola*; Saffran, Johnson, Aslin and Newport 1999; Saffran, Newport and Aslin 1996). It is not immediately clear to me how attention to transitional probabilities alone can account for this effect because both partwords are equally 'word-like' in terms of their transitional probabilities. The Saffran et al. (1996) infant word segmentation paradigm avoids this interpretational difficulty by only testing infants on the partwords consisting of the last syllable of one word plus the first two of another (the partwords that adults found least 'word-like'). However, one study that has examined infants' sensitivity to transitional probabilities between the first and second syllable of a statistical word versus the second and third syllable has found a pattern very reminiscent of the adult pattern. When exposed to an artificial language like the one used in Saffran et al., infants seem to perceive the last two syllables of a statistical word as more 'word-like' than the first two syllables (Johnson and Juscyzk 2003b). This is despite the fact that the transitional probabilities between the first two and last two syllables in the words were held equal. This behavior of recognizing part of a word as familiar does not neatly match up with the behavior we see in natural language segmentation studies (e.g. in general, infants tend to segment whole words from speech, not parts of words; Houston, Santelmann and Juscyzk 2004; Jusczyk, Houston Newsome 1999). This example serves to illustrate how little we really know about which statistics infants are actually tracking, even in very simplified artificial languages.

Studies on the acquisition of morphosyntactic dependencies provide another example where infant behavior in a natural language study does not necessarily entirely line up with their behavior in a artificial language learning study. Based on artificial language studies, it has been claimed that increased variation in the tokens occurring between two nonadjacent elements should make it easier to learn the relationship between the two non-adjacent elements (e.g. Gómez and Maye 2005). For example, Gomez (2002) found that toddlers exposed to nonsense strings with an A-X-B structure (e.g. pel wadim rud, pel kicey rud, etc) learned the nonadjacent dependency between syllables A and B when 24 different syllables occurred in the X position, but not when only 3 or 12 syllables occurred in the X position. The authors concluded that as the strength between adjacent dependencies diminishes (i.e. as the

number of elements that can occur in the X position increases), infants shift their attention from adjacent dependencies to non-adjacent dependencies. Van Heugten and Johnson (2010) were interested in whether any evidence for this phenomenon could be observed in natural language acquisition. Interestingly, in a study combining a perception study with a corpus analysis, van Heugten and Johnson found no evidence that acquisition of natural language non-adjacent dependencies is impacted by variability in the material intervening the elements forming the dependency. More specifically, they found that Dutch infants appeared to learn the diminutive dependency before the plural dependency despite the fact that their corpus analysis revealed that the plural dependency tends to have far more interceding token variability in the input than the diminutive dependency. Taken at face value, this could be taken as evidence that perhaps the acquisition of non-adjacent dependencies in artificial languages may depend on different computational mechanisms than the acquisition of non-adjacent dependencies in natural languages. Of course, natural language examples cannot be nearly as well controlled as artificial language examples. So it may be the case that other complicating factors were driving the order of acquisition of these dependencies. Clearly, one exception does not necessarily break the generality. Nonetheless, it does certainly motivate one to want to investigate this issue further.

To summarize, a broadly transcribed orthographic representation of speech (even with some corrections for typical pronunciation variants in particular contexts) is a completely different sort of animal than the speech actually produced in everyday interactions (e.g. Johnson 2004; Shockey 2003). Even something as simple as a syllable count can become complicated when normal everyday reduction is taken into account. Consider that the word *probably* can be produced with one [prai], two [prabli], or three [prabəbli] syllables in different speech conditions. And connected speech processes, such as assimilation and coarticulation, are not easily captured in most current computational models because they are graded rather than binary all-or-none phenomenon. Adults are not thrown off by the lack of invariance in the speech signal and appear to be highly sensitive to the acoustic-phonetic detail of utterances. Recent research has shown that adults use this information to work out the intended meaning behind others utterances. We are only just beginning to understand how infants deal with variation and acoustic-phonetic detail in speech. We know very little about how infants perceive speech in context. And even if we assume that infants can extract the same information that adults can extract, we still have very little understanding of which of the many possible statistical analyses infants are performing over the input. When you combine these concerns with those expressed in the previous section of this paper ('Does it scale up'), you are faced with the real life possibility that the solutions natural language learners use to master natural languages may not just be quantitatively different from those used by artificial language learners, they may in fact be qualitatively different. We are not the first to worry about these types of issues. In the words of Soderstrom, Conwell, Feldman and Morgan (2009)

'… *extant models have been hand-crafted for particular problems, selecting relevant properties of input and learning mechanisms to arrive at predetermined output structures.*

Although these models may provide proofs-in-principle on possible ways to solve language learning problems, they do not illuminate how the learner knows how to select and organize the input appropriately for any particular task, what the most appropriate out- put representations might be, or how the learner chooses specific statistical analyses to pursue.'

2.3. Can distributional models predict children's difficulties?

A good test of any model is its ability to predict errors or difficulties in performance. One very nice aspect of distributional models of word segmentation is that they make strong testable predictions about where infants should make segmentation errors. Below we summarize some behavioral tests of these predictions. More research is clearly needed in this area, but at present, it seems that the infant data does not entirely line up as neatly as one would like with the predictions made by the strongest (or perhaps simplest) proposed models of distributional learning. I will illustrate this issue with examples from the segmentation literature.

If a language learner were to rely heavily on transitional probabilities between syllables to extract an initial cohort of words from speech, there are two types of errors that we would expect to see. First, children should segment frequently co-occurring words such as 'see it' or idiomatic phrases such as 'piece of cake' as one word. Likewise, children should over-segment words containing highly frequent spurious words or morphemes embedded within them such as the 'be' in 'behave', 'a' in 'parachute', or 'ing' in 'singing'. To some degree, corpus studies have supported the notion that such errors should be seen. For example, in a study designed to demonstrate that infants can segment words from speech by tracking transitional probabilities between syllables, Swingley (2005) reports the following patterns in his results:

'Examination of the Dutch false alarms suggests two factors that conspired to reduce the accuracy of the Dutch bisyllable analyses. One was the number of fixed expressions consisting of pairs of monosyllabic words. For example, the Dutch false alarm hou vast ("hold on") contains two words that hardly ever occurred in other contexts. As noted previously, several bisyllabic false alarms were conventional expressions, particularly in the Dutch analyses. More importantly, Dutch infant-directed speech contains more trisyllabic words than similar English speech; on occasion these words were not detected as trisyllables, but did trigger postulation as bisyllables. Examples include the first two syllables of boterham ("sand- wich"), mannetje ("little man"), mopperen ("being grumpy"), vreetzakje ("little eating-too-much person"), and zonnetje ("little sun"), and the last two syllables of olifant ("elephant") and eventjes ("just" or "for a little while"). Some of these words are morphologically complex, consisting of a bisyllabic word and the diminutive suffix -je or -tje. The Dutch diminutive is productive and frequent, making full trisyllables containing the diminutive suffix difficult to extract. Thus, to some degree the relatively low accuracy of the Dutch analyses can be traced to structural properties of the language.'

Behavioral studies, however, suggest that infants perform far better than a distributional model relying solely on transitional probabilities would predict. Research has shown that infants (like adults) do not, as some models might predict, simply pull out recurrent patterns in the input (Mattys and Jusczyk 2001). Moreover, infants do not over-segment multisyllabic words containing spurious function words embedded within them. For example, 7.5-month-olds familiarized with passages containing trisyllabic words such as 'parachute' subsequently recognize the word 'parachute', but not the words 'pair' or 'chute' (Houston, Santelmann and Jusczyk 2004). This is despite the fact that the syllable 'a' is a highly frequent word in English that might be expected to trigger over-segmentation of longer words containing the syllable 'a'. At the same time, 7.5-month-olds familiarized with three word phrases such as 'pair of mugs' pull out 'pair' and 'mugs' as units, but not 'pair of mugs' (Johnson, van der Weijer and Jusczyk 2001). These studies suggest that young infants may have some strategy besides simple syllable distribution analyses to help them detect the intended word boundaries in speech. Other studies point to similar conclusions. Eight-month-olds familiarized with a passage containing the word 'catalogue' segment out the word 'catalogue', but not the word 'cat'. At the same time, 8-month-olds familiarized with a passage containing repetitions of the phrase 'cat a log' segment out the word 'cat' but do not segment out the word 'catalogue' (Johnson 2003). In a related study that controlled for both intonation boundaries and the occurrence of spuriously embedded function morphemes, 7.5- and 12-month olds recognized the word 'toga' only when they were familiarized with passages containing repetitions of the phrase 'toga lore'. Infants did not recognize the word 'toga' when they were familiarized with a passage containing repetitions of the phrase 'toe galore' (Johnson 2003; 2008a). And finally, in a study examining Dutch-learning 11-month-olds' segmentation of polysyllabic words ending in the highly frequent diminutive suffix '–je', it was found that for the most part, infants tended to perceive the suffix as part of the polysyllabic word (i.e. the suffix was not blindly 'stripped off' as might be predicted by some distributional models; Johnson 2008b).

Taken together, these studies suggest that 1) infants do not always over-segment polysyllabic words containing frequently occurring spurious function words embedded within or attached to them, and 2) repeatedly re-occurring strings of words are not necessarily under-segmented (see, however, Johnson, 2003, for evidence that infants perceive idiomatic renditions of word strings as more word-like than their literal word string counterparts).

At the same time, there are some findings in the literature that appear to show that infants really should (at least sometimes) make the errors predicted by simple distributional models of word segmentation. For example, studies have shown that infants segment the nonsense word 'breek from the utterance 'thebreek' (which contains a real function word within it) but infants do not segment the word 'breek' from the utterance 'kuhbreek' (which does not contain a real function word within it; Shi, Cutler, Werker and Cruickshank 2006). The authors interpreted this study as evidence that the infants 'stripped off' the frequent word 'the' to discover the new word 'breek' (see Christophe, Guasti, Nespor, Dupoux and Van Ooyen 1997, for discussion). Can we reconcile this evidence for function word stripping with the fact

that the occurrence of 'a' in catalogue and parachute does not cause 8-month-olds to over-segment these utterances into three word phrases?

There are at least two good ways to reconcile these findings. It is very possible that the 8-month-olds tested in the catalogue segmentation study had not yet learned the function words embedded within the trisyllabic words. Perhaps the results would have been different had we tested slightly older infants. Another possibility is that infants are sensitive to the acoustic-phonetic differences between spurious and intended renditions of words. In other words, infants may be able to do something not predicted by most distributional models of word segmentation: they may be somehow sensitive to speakers intended productions and differentiate between real and intended function words (see Conwell and Morgan 2007; Johnson 2003; for a related discussion). Using the fine-grained acoustic-phonetic structure of natural utterances to work out the parse of a speakers' intended message may actually be necessary to explain word segmentation, given the prevalence of spurious embedded words in languages such as English (McQueen, Cutler, Briscoe and Norris 1995). This notion fits nicely with a growing body of literature demonstrating that attention to the fine-grained acoustic phonetic structure of speech is also important for adult speech perception (e.g. McMurray, Tanenhaus and Aslin 2002; Shatzman and McQueen 2006; Spinelli, McQueen and Cutler 2003). Note, however, that if this is true, then this suggests a very large gap between what infants are attending to in the speech signal and what information corpus models based on orthographic transcriptions are tracking. It may be the case that models of developmental speech perception could predict infants' segmentation errors much more accurately if they took some of the factors discussed in this section into account. Clearly additional research is needed in this area.

2.4. How much innate knowledge is too much innate knowledge?

By the end of their first year of study in Psychology, undergraduates are already citing the mantra that no behavior is completely learned or innate. Rather, all human abilities depend on a combination of experience-independent and experience-dependent factors. Of course they are right. Given what we know today about the complex interaction between our biological endowment and our environment, it would be ludicrous to believe anything else. Indeed, from the mid nineties forward, nearly all (if not all) proponents of distributional models of language development would have heartily agreed with this statement. And even prior to the popularization of distributional models, even the staunchest supporters of the nativist perspective had to allow a substantial learning component into their models because all languages are incredibly unique (in other words, even if the parameters along which languages can vary are innate, they still have to be set). Indeed, one could argue that infants' statistical learning abilities simply justify the assumptions underlying many parameter setting models.

But given that language abilities must be based on a combination of learned and innate factors, how much innate knowledge is too much (or too little) innate knowledge? Some have invoked Occam's Razor to answer this question. According to Occam's Razor, if you have two hypotheses that explain an observation equally well,

then the simplest most parsimonious hypothesis is best. Using this sort of logic, some have argued that a model allowing for any built-in (or innate) constraints on processing is less acceptable than one that can do without. But is this a justifiable use of Occam's Razor? Who is to say whether it is more parsimonious to propose that 1) infants have calculator-like brains that can expertly track and extract a specific generalization from a particular multi-level complex pattern in the input, or 2) infants are born with predispositions or innate constraints (be they perceptual or cognitive) that narrow the possible range of solutions infants consider when faced with working out a particular pattern in their language input? Perhaps we need to allow for both possibilities, and avoid ruling out the latter possibility a priori simply because an experience-based calculation (no matter how complex) can in theory do the work (especially given the concerns expressed above in the section 'Which unit? And once you pick a unit, which statistic?').

To make this discussion more concrete, I will illustrate my point with an example from the domain of word segmentation. Brent and Cartwright (1996) improved the performance of their distributional model for word segmentation by requiring that all possible parses contain a vowel. Similar constraints have been built into other adult models of word segmentation (Frank, Goldwater, Mansinghka, Griffiths and Tenenbaum 2007) and online word recognition (Norris, McQueen, Cutler and Butterfield 1997). Others have proposed a linguistically motivated and very useful constraint that no word should contain more than one syllable carrying primary stress (Yang 2004). Still others have suggested that statistics are tracked differently across different types of speech segments (Mehler, Peña, Nespor and Bonatti 2006). Note that what all of these suggestions have in common is that they are not suggesting ad hoc constraints on early speech perception, they are suggesting linguistically motivated constraints. However, some members of the modeling community seem to be suggesting that if possible we should reject the need for these constraints because the use of such a constraint assumes innate knowledge and is therefore not parsimonious (e.g. Monaghan and Christiansen 2010). But is this justified?

It seems to be that one needs to be careful over confusing what is the most parsimonious way to design a computational model versus what is the most parsimonious way to explain early language acquisition. Rather than blindly invoking Occam's Razor, a better way to address this issue might be to at least give linguistically motivated constraints a healthy consideration. For example, why not run perceptual studies designed to test whether infants actually possess a constraint against considering segmental strings lacking vowels as possible words? Indeed, there is some evidence in the literature that such a constraint might exist. For example, English-learning 12-month-olds behave like adults in that they segment 'win' from 'window' but not 'wind' (Johnson, Juscyzk, Norris and Cutler 2003). Additional behavioral data with younger infants would be useful to help decide whether it is parsimonious to include such a constraint in computational models of word segmentation. Note, however, that if such constraint was included in computational models, it would have to be flexible enough to deal with language-specific exceptions such as Slovakian prepositions (Hanulikova, McQueen and Mitterer 2010).

In short, it is clear that infants have excellent statistical learning abilities, and the distribution of linguistic patterns surely provide infants statisticians with useful information. However, does this mean that we should allow no innate biases or learning constraints beyond a tendency to look for statistical patterns in the input? It seems to me that built-in constraints to models of early language learning can be implemented in a parsimonious fashion, and in the end I suspect they will be very necessary to explain several aspects of early language acquisition.

2.5. Looks can be deceiving: the potential dangers of rich interpretations

Nearly all of the infant behavioral studies testing distributional models of language acquisition that have been reported thus far in this chapter have been based on one very simple dependent measure: length of look. Indeed, most infant testing paradigms use length or speed of look because we cannot ask infants to verbally indicate whether they, for example, recognize a word or grammatical construction. Indeed, looking paradigms have revolutionized the field of developmental psychology. However, there has been a longstanding uneasiness with the difficulties involved in interpreting infant looks (e.g. Aslin 2007; Cohen 2001). My own personal view is that looking paradigms are absolutely invaluable in the study of infant perception and cognition. However, it is important that the results of these studies be interpreted very cautiously, especially when examining the processing of higher-level structures of language (see Kooijman, Johnson and Cutler 2008, for a related discussion).

In the past, one common approach to compensating for the weaknesses of looking procedures was to run many experiments addressing the same issue and trying to control for as many possible alternative explanations as possible. For example, there were fifteen (fifteen!) experiments in the original paper claiming that English-learning infants use lexical stress to identify word boundaries in fluent speech (Jusczyk, Houston and Newsome 1999). Certainly, even in the 90's, reporting 15 experiments in a single paper was out of the ordinary. But nowadays, one is hard-pressed to find a paper on infant language learning with more than even 2 or 3 experiments. This would be fine if we were all interpreting our results rather cautiously, but even when we try to be cautious sometimes it is hard to imagine all the different possible explanations for the results we obtain (not to mention the pressure to publish exciting findings, and to do so very quickly). I am not suggesting that papers with 2 or 3 infant experiments should not be published, but certainly in the best case scenario multiple follow-up studies and replications would also be published in order to ensure the generality of the conclusions drawn in the original set of studies.

Studies examining infants' sensitivity to grammatical constructions provide a good example of the shortcomings (and frustrations) of looking time data. For example, toddlers have been shown to listen longer to utterances containing grammatical morphosyntactic dependencies (e.g. she is walking the dog) than utterances containing ungrammatical dependencies (e.g. she can walking the dog; Höhle, Schmitz, Santelmann and Weissenborn 2006; Santelmann and Jusczyk 1998). This has been interpreted as evidence that infants have learned (or are sensitive to) this dependency. The results of artificial language studies have further suggested that these types of dependencies could be learned by tracking non-adjacent dependencies

in speech (Gomez and Gerken 2000). Indeed, more recent studies have actually shown that the dependencies that are most strongly marked by statistical cues in the input are the very dependencies that toddlers first show evidence of knowing in looking time studies (van Heugten and Johnson 2010). All of this seems to be strong convergent evidence that young children learn morphosyntactic dependencies by tracking non-adjacent dependencies between syllables in their linguistic input. However, there is another slightly less exciting explanation.

Perhaps infants like to listen to things that sound familiar. Things that occur frequently in the environment sound familiar, thus things that are statistically frequent in the environment attract longer looking times (see van Heugten and Johnson 2010; van Heugten and Shi 2009, for a related discussion). How does this differ from saying that toddlers have developed sensitivity to discontinuous dependencies by tracking the statistical relationship between non-adjacent elements in their language input? In fact, whether or not it differs depends entirely on whether you adopt a conservative or rich interpretation of the looking time studies. A rich interpretation would credit the child with sophisticated grammatical knowledge, whereas a more conservative interpretation would simply credit the child with picking up on a pattern frequently heard in the input. Most researchers reporting that infants are sensitive to non-adjacent (or discontinuous) dependencies in natural speech word their findings very carefully. Saying that an infant is 'sensitive to discontinuous dependencies' makes no theoretical claims regarding the underlying nature grammatical knowledge of the child. Can tracking discontinuous dependencies actually be the learning mechanism that serves as the main driving force behind the acquisition of abstract grammatical constructions in English? Or are input distributions and statistics just triggers that help infants work out which of the many possible structures of human language that they are encountering? Maybe the language competency possessed by children is just item based, and isn't really based on abstract grammatical knowledge at all (e.g. Tomasello 2000)?

It would be helpful if in the future, language researchers would more clearly spell out their assumptions regarding the knowledge driving infants' looks. Is it grammatical knowledge? Is it statistical knowledge? Is there even a difference between grammatical and statistical knowledge? If there is a difference between the two, then we need to develop an explanation for how and when children shift from statistical to grammatical knowledge, which will require a way to test the underlying nature of children's sensitivity to linguistic patterns. If developmentalists propose no shift away from the initial statistical knowledge, then they must be held accountable for explaining studies demonstrating that adults have abstract linguistic knowledge.

Looking time measures will be helpful in addressing these issues, but the use of convergent measures, more clearly laid out theoretical assumptions, and cleverly designed experiments will also be necessary. In my lab, we are trying to begin to address these issues by presenting toddlers with strictly familiar patterns (i.e. statistically common in the input, but ungrammatical) versus grammatical items (statistically less common, but nonetheless fully grammatical). Preliminary work in this area suggests that toddlers' knowledge of discontinuous dependencies might very well not be grammatical in the adult sense of the word (van Heugten and Johnson 2011). More work in this area will surely emerge in the near future as researchers

move beyond demonstrating the remarkable statistical learning abilities of infants and move towards developing more comprehensive theories of language development.

3. Closing Comments

The title of this chapter poses a question: Are infant statisticians up to performing the job of bootstrapping language? In large part, the answer to this question might depend on how you define the 'job' of language acquisition. Clearly, infants are statisticians of some sort, as study after study has demonstrated how exquisitely attuned they are to statistical patterns in their environment. And many aspects of language are reflected in its statistical structure (especially if one assumes adult-like representations of the input). But when it comes to language acquisition, it is not yet entirely clear what information infants need to learn, what units of information they are tracking, or what calculations they might be performing over these units. And the looking time data we often used to address these questions is frustratingly ambiguous (and tempting to over-interpret). Indeed, we do not yet really know whether the task of artificial language learning is quantitatively or qualitatively different from natural language learning. As much as I appreciate the beauty of a well-designed artificial language study, I must admit that I fear that the language learning task faced by a child in an artificial language learning study differs both qualitatively and quantitatively from the task faced by a child learning a language in the real world. At the same time, artificial language learning studies allow us to test hypotheses in the laboratory that we cannot realistically test in the real world.

So how shall we, as a field, move forward towards discovering an answer to this difficult question? My advice would be to keep chipping away at the many challenges to distributional models, and to try to better define the role distributional learning plays in language acquisition. I would also recommend looking to the information-rich patterns in the speech signal as another potentially important cue to language structure (e.g. see Endress and Mehler 2009; Johnson and Seidl 2008; Johnson and Tyler 2010, for a discussion of the potential importance of utterance level prosody in early language acquisition). I also look forward to seeing more work focused on understanding the basic fundamentals of how infant speech perception differs from adult speech perception, and how experience sculpts infant language learners into adult language users. By better understanding these basic issues, we will be in a better position to judge the feasibility of many distributional models proposed based on artificial language and corpus studies.

In closing, regardless of which level of linguistic structure we focus on, the language acquisition 'problem' is remarkably far from being solved. At this point, it is impossible to say with any certainty whether or not infant statisticians are capable of bootstrapping language. Researchers are still at the early stages of grappling to understand what implications our discovery of infants' statistical learning abilities should have on our theories of language acquisition. The last 15 years have been such an exciting time to be involved in language acquisition research. As researchers continue to tackle the challenges posed in this chapter, I expect the next 15 years to be equally exciting.

References

Aslin, Richard N.   2007   What's in a look? *Developmental Science*, 10: 48-53.

Aslin, Richard N. and Elissa L. Newport   2008   What statistical learning can and can't tell us about language acquisition. In J. Colombo, P. McCardle, and L. Freund (eds.), Infant Pathways to Language: Methods, Models, and Research Directions. Mahwah, NJ: Lawrence Erlbaum Associates.

Aslin, Richard N., Jenny R. Saffran and Elissa L. Newport   1998   Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9:* 321–324.

Batchelder, Eleanor   2002   Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83: 167–206.

Bates, Elizabeth and Jeffrey L. Elman   1997   Learning rediscovered: A perspective on Saffran, Aslin, and Newport. *Science*, 274: 1849-1850.

Bonatti, Luca L., Marcela Peña, Marina Nespor and Jacques Mehler   2006   How to hit Scylla without avoiding Charybdis : comment on Perruchet, Tyler, Galland, and Peereman. *Journal of Experimental Psychology: General.* 135: 314-326

Bortfeld, Heather and James L. Morgan   2010   Is early word-form processing stress-full? How natural variability supports recognition. *Cognitive Psychology*, 60: 241-266.

Brent, Michael and Timothy Cartwright   1996   Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61: 93–125.

Byrd, Dani M. and Toben H. Mintz   2010   Discovering Speech, Words, and Mind. Wiley-Blackwell Publishing.

Cairns, P., Richard Shillcock, Nick Chater & Levy, J.   1997   Bootstrapping word boundaries: A bottom-up corpus-based approach to segmentation. *Cognitive Psychology*, 33: 111–153.

Chambers, Kyle E., Kris H. Onishi and Cynthia Fisher   2003   Infants learn

phonotactic regularities from brief auditory experience. *Cognition*, 87: B69–B77.

Chomsky, Noam   1957   *Syntactic structures.* The Hague, NL: Mouton.

Christophe, Anne, Guasti, T., Marina Nespor, Emmanuel Dupoux, & Van Ooyen, B.
1997   Reflections on phonological bootstrapping: Its role for lexical and syntactic
acquisition. Language and Cognitive Processes, 12: 585-612.

Cohen, Leslie B.   2001, April   *Uses and Misuses of Habituation: A Theoretical
and Methodological Analysis*. Symposium paper presented at the Society for
Research in Child Development Meeting, Minneapolis, MN.

Conwell, Erin & James L. Morgan   2007   Resolving grammatical category
ambiguity in acquisition. In H. Caunt-Nulton, S. Kulatilake, & I. Woo (Eds.),
Proceedings of the 32th annual Boston University Conference on Language
Development: Vol. 1 (pp. 117-128). Somervill, MA: Cascadilla Press.

Curtin, Suzanne, Toben Mintz and Morten Christiansen2005   Stress Changes the
representational landscape: Evidence from word segmentation. *Cognition*, *96*: 233-
262.

Curtiss, Susan   1977   *Genie: a psycholinguistic study of a modern-day "wild
child"*. Boston: Academic Press.

De Vries, Meinou, Padraic Monaghan, Stefan Knecht & Pienie Zwitserlood   2008
Syntactic structure and artificial grammar learning: the learnability of embedded
hierarchical structures. *Cognition*, 107: 763-774.

Endress, Ansgar D. and Jacque Mehler   2009   Primitive Computations in Speech
Processing. *Quarterly Journal of Experimental Psychology*.

Frank, Michael. C., Goldwater, S., Mansinghka, V., Griffiths, T. & Tenenbaum, J. (2007).
Modeling human performance on statistical word segmentation tasks. In D. S.
McNamara & G. Trafton (eds), Proceedings of the 29th Annual Meeting of the
Cognitive Science Society, 281–86. Mahwah, NJ: Lawrence Erlbaum.

Gervain, Judit, Marina Nespor, Reiko Mazuka, Ryota Horie & Jacques Mehler   2008
Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic
study. *Cognitive Psychology,* 57: 56–74

Gómez , Rebecca & Lou Ann Gerken   2000   Infant Artificial Language Learning
and Language Acquisition. *Trends in Cognitive Sciences, 4:* 178-186.

Gómez, Rebecca   2002   Variability and the detection of invariant structure.
Psychological Science, 13: 431-436.

Gómez, Rebecca & Jessica Maye   2005   The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7: 183-206.

Goodsitt, J.V., James L. Morgan & Patricia Kuhl   1993   Perceptual strategies in prelingual speech segmentation. *J Child Lang*, 20: 229–52.

Gow, David W.   2002   Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, 28: 163-179.

Harris, Zellig   1955   From phoneme to morpheme. *Language, 31:* 190–222.

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.

Höhle, Barbara, Michaela Schmitz, Lynn Santelmann & Jurgen Weissenborn   2006 The recognition of discontinuous verbal dependencies by German 19-month-olds: Evidence for lexical structural influences on children's early processing capacities. Language Learning and Development, 2: 277-300.

Houston, Derek M & Peter W. Jusczyk 2000   The role of talker- specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception & Performance, 26*: 1570-1582.

Houston, Derek M., Lynn Santelmann & Peter W. Jusczyk   2004   English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes*, 19: 97-136.

Hanulikova, Adriana, James McQueen & Holger Mitterer   2010   Possible words and fixed stress in the segmentation of Slovak speech. *Quarterly Journal of Experimental Psychology*, 63: 555 -579.

Johnson, Elizabeth K.   2003   Word segmentation during infancy: The role of subphonemic cues to word boundaries. Unpublished Doctoral Dissertation, The Johns Hopkins University.

Johnson, Elizabeth K.   2005   English-learning infants' representations of word-forms with iambic stress. *Infancy,* 7: 95-105.

Johnson, Elizabeth K.   2008a   Dutch infants' segmentation of diminutive word forms: An exploration of morpheme stripping and lexical embedding. Paper presented at the 16th International Conference on Infant Studies, Vancouver, Canada.

Johnson, Elizabeth K.   2008b   Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. *Journal of the Acoustical Society of America*,

123: EL144-EL148.

Johnson, Elizabeth K. & Peter W. Jusczyk   2001   Word segmentation by 8-month-olds: When speech cues count more than statistics. Journal of Memory and Language, 44: 548–567.

Johnson, Elizabeth K. & Peter W. Jusczyk   2003a   Exploring statistical learning by 8-month-olds: The role of complexity and variation. In D. Houston, A Seidl, G.Hollich, E. Johnson, & A. Jusczyk (Eds.) *Jusczyk Lab Final Report* (pp. 141-148).

Johnson, Elizabeth K., & Peter W. Jusczyk   2003b   Exploring possible effects of language-specific knowledge on infants' segmentation of an artificial language. In D. Houston, A Seidl, G.Hollich, E. Johnson, & A. Jusczyk (Eds.) *Jusczyk Lab Final Report* (pp. 141-148).

Johnson, Elizabeth K., Peter W. Jusczyk, Anne Cutler & Dennis Norris   2003   Lexical viability constraints on speech segmentation by infants without a lexicon. *Cognitive Psychology,* 46: 65-97.

Johnson, Elizabeth K. & Amanda Seidl   2008   Clause segmentation by 6-month-olds: A crosslingusitic perspective. *Infancy,* 13: 440-455.

Johnson, Elizabeth K. & Amanda Seidl   2009   At 11 months, prosody still outranks statistics. *Developmental Science,* 12: 131-141.

Johnson Elizabeth K. & Michael D. Tyler   2010   Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13: 339-345.

Johnson, Elizabeth K., Joost Van de Weijer & Peter W. Jusczyk   2001   Word segmentation by 7.5-month-olds: Three words do not equal one. BUCLD 25: Proceedings of the *25th Annual Boston University Conference on Language Development* (Volume 2: 389-400). Cascadilla Press, Somerville, MA.

Johnson, Elizabeth K. & Marieke van Heugten   in press   Infant artificial language learning. In N.M. Seel (Ed.), *Encyclopedia of the Sciences of Learning.* Heidelberg: Springer-Verlag.

Johnson, Keith   2004   Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium (pp. 29–54). Tokyo, Japan: The National International Institute for Japanese Language.

Jusczyk, Peter W   1997   The discovery of spoken language. Cambridge, MA: MIT Press.

Jusczyk, Peter W., Derek Houston & Mary Newsome   1999   The beginnings of

word segmentation in English-learning infants. *Cognitive Psychology*, 39: 159–207.

Jusczyk, Peter W., David Pisoni, D. & James Mullennix   1992   Some consequences of stimulus variability on speech processing by 2-month-old infants. C*ognition*, 43: 253–291.

Kirkham, Natasha Z., Jonathan Slemmer & Scott P. Johnson   2002   Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition, 83:* B35–B42.

Kooijman, Valesca, Elizabeth K. Johnson & Anne Cutler   2008   Reflections on reflections of infant word recognition. In Friederici, Thierry (ed.), In book entitled *Early Language Development: Bridging Brain and Behaviour, Series Trends in Language Acquisition Research.* (pp. 91-114). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Kuhl, Patricia   1979   Speech perception in early infancy: Perceptual constancy for perceptually dissimilar vowel categories. *Journal of the Acoustical Society of America*, 66: 1168-1679.

Mattock, Karen, M. Molnar, Linda Polka & Denis Burnham   2008   The developmental course of lexical tone perception in the first year of life *Cognition*, 106, 3: 1367-1381.

Mattys, Sven & Peter W. Jusczyk   2001   Do infants segment words or continuous recurring patterns? *Journal of Experimental Psychology: Human Perception and Performance*, 27: 644-655.

Maye, Jessica, Janet Werker & Lou Ann Gerken   2002   Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82: 101–111.

McMurray, Bob & Richard N. Aslin   2005   Infants are sensitive to within-category variation in speech perception. *Cognition*, 95: B15-B26.

McMurray, Bob, Michael Tanenhaus & Richard N. Aslin, R   2002   Gradient effects of within-category phonetic variation on lexical access, *Cognition, 86:* B33-B42.

McQueen, James M., Anne Cutler, T. Briscoe & Dennis Norris, D 1995   Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes*, 10: 309–331.

Mehler Jacques, Marcela Peña, Marina Nespor & Luca Bonatti   2006   The soul of language does not use statistics: reflections on vowels and consonants. *Cortex*. 42: 846-854.

Mintz, Toben   2003   Frequent frames as a cue for grammatical categories in

child directed speech. *Cognition, 90:* 91–117.

Monaghan, Padraic, Morten Christiansen & Nick Chater, N   2007   The phonological–distributional coherence hypothesis: Cross-linguistic evidence in language acquisition, *Cognitive Psychology,* 55: 259–305.

Morgan, James L. & Elissa L. Newport 1981   The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20:* 67–85.

Morgan, James L. & Jenny Saffran   1995   Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Dev,* 66: 911–936.

Newman, Rochelle S.   2009   Infant's listening in multitalker environments: Effect of the number of background talkers. *Attention, Perception & Psychophysics*, 71: 822-836.

Newport, Elissa L. & Richard N. Aslin   2004   Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48:* 127–162.

Norris, Dennis G., James M. McQueen, Anne Cutler & Sally Butterfield   1997   The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34: 191-243.

Nurayan, Chandan, Janet Werker & Patricia Beddor   2010   The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, 13: 407-420.

Pelucchi, Bruna, Jessica Hay & Jenny Saffran   2009a   Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80: 674–685.

Pelucchi, Bruna, Jessica Hay & Jenny Saffran   2009b   Learning in reverse: Eight-month-old infants track backwards transitional probabilities. *Cognition*, 113: 244-247.

Peperkamp, Sharon, R. Le Calvez, J.-P Nadal, & Emmanuel Dupoux   2006   The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition*, 101: B31-B41.

Perruchet, Pierre & S. Desaulty   2008   A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, 36: 1299-1305.

Perruchet, Pierre, Ronald Peereman & Michael Tyler   2006   Do we need algebraic-like computations? A reply to Bonatti, Peña, Nespor, and Mehler (2006) . Journal of Experimental Psychology: General, 135: 322-326.

Perruchet, Pierre, & Annie Vinter   1998   Parser: A model for word segmentation. *Journal of Memory and Language*, 39: 246–263.

Rost, Gwyneth & Bob McMurray   2009   Speaker variability augments phonological processing in early word learning. *Developmental Science, 12:* 339-349.

Rytting, C. Anton, Chris Brew & Eric Fossler-Lussier   2010   Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, 37: 513-543.

Saffran, Jenny R.   2001   The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44: 493–515.

Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport   1996   Statistical learning by eight-month-old infants. *Science, 274:* 1926–1928.

Saffran, Jenny R., Elizabeth K. Johnson, Richard N. Aslin & Elissa L. Newport   1999 Statistical learning of tonal sequences by human infants and adults. *Cognition*, 70: 27-52.

Saffran, Jenny R., & Erik D. Thiessen   2003   Pattern induction by infant language learners. *Developmental Psychology, 39:* 484–494.

Saffran, Jenny R., & D.P. Wilson   2003   From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4:* 273–284.

Sahni, Sarah D., Mark S. Seidenberg & Jenny R. Saffran   2010   Connecting cues: overlapping regularities support cue discovery in infancy. *Child Development*, 81: 727-736.

Santelmann, Lynn M. & Peter W. Jusczyk   1998   Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition, 69:* 105–134.

Sue Savage-Rumbaugh & William M. Fields   2000   Linguistic, Cultural and Cognitive Capacities of Bonobos(Pan Paniscus). *Culture and Psychology*, 6: 131-153.

Schmale, Rachel, Alejandrina Christià, Amanda Seidl & Elizabeth K. Johnson   2010 Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy,* 15: 650-662.

Seidl, Amanda & Eugene Buckley   2005   On the learning of arbitrary phonological rules. *Language Learning and Development*, 1: 289-316.

Shatzman, Keren B. & James M. McQueen   2006   Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science,* 17: 372-377.

Shi, Rushen, Anne Cutler, Janet F. Werker & Marisa Cruickshank   2006   Frequency and form as determinants of functor sensitivity in English acquiring infants. *Journal of the Acoustical Society of America*, 119: EL61–EL67.

Shockey, Linda   2003   *Sound patterns of spoken English*. Oxford: Blackwell Publishing.

Singh, Leher, James L. Morgan, Katherine S. White   2004   Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51*: 173-189.

Spinelli, Elsa, James M. McQueen & Anne Cutler   2003   Processing resyllabified words in French. *Journal of Memory and Language*, 48: 233-254.

Smith, Linda & Chen Yu   2008   Infants rapidly learn word- referent mappings viacross-situational statistics. *Cognition*, 106: 1558-1568.

Soderstrom, Melanie, Erin Conwell, Naomi Feldman & James L. Morgan   2009   The learner as statistician: three principles of computational success in language acquisition. *Developmental Science*, 12: 409-411.

Sundara, Megha, & Adrienne Scutellaro   in press   Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*.

Swingley, Daniel   2005   Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50: 86-132.

Terrace, Herb S., Laura-Ann Petitto, R.J. Sanders & T.G. Bever   1979   Can an ape create a sentence? Science, 206: 891-902.

Thiessen, Erik D., Emily Hill & Jenny R. Saffran   2005   Infant-directed speech facilitates word segmentation. *Infancy*, *7*: 53-71.

Thiessen, Erik & Jenny Saffran   2003   When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39: 706–716.

Thiessen, Erik D. & Jenny R. Saffran   2007   Learning to learn: Infants' acquisition of stress-based segmentation strategies for word segmentation. *Language, Learning and Development*, *3*: 73-100.

Thompson, Susan P. & Elissa L. Newport   2007   Statistical learning of syntax:

The role of transitional probability. *Language Learning and Development*, 3: 1–42.

Tomasello, Michael   2000   Do young children have adult syntactic competence? *Cognition*, 74: 209–253.

Toro, Juan M. & Joseph Trobalon   2005   Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, 67: 867-875.

Tsao, Feng Ming, Huei-Mei Liu & Patricia Kuhl   2006   Perception of native and non-native affricate–fricative contrasts: cross- language tests on adults and infants. *Journal of the Acoustical Society of America*, 120: 2285–2294.

Tyler, Michael & Anne Cutler   2009   Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, 126: 367-376.

Valian, Virginia & Seana Coulson 1988   Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27:* 71–86.

Van Heugten, Marieke & Elizabeth K. Johnson   2009   The robustness of infants' early word representations. *Journal of the Canadian Acoustical Association,* 37: 148-149.

Van Heugten, Marieke & Elizabeth K. Johnson   2010   Linking infants' distributional learning abilities to natural language acquisition. *Journal of Memory and Language, 63*: 197-209.

Van Heugten, Marieke & Elizabeth K. Johnson   April 2011   Is dependency acquisition by children grammatical or distributional in nature? Poster presented at the Society for Research in Child Development Biennial Meeting, Montreal, Canada.

Van Heugten, Marieke & Rushen Shi   2009   French-learning toddlers use gender information on determiners during word recognition. *Developmental Science*, 12: 419-425.

Vihman, Marilyn & Virve-Anneli Vihman   in press   From first words to segments: A case study in   phonological development. In E. V. Clark & I. Arnon, eds., How children make linguistic generalizations: Experience and variation in learning a first language. Trends in Language Acquisition Research. Amsterdam: Benjamins.

Werker, Janet F. & Richard C. Tees   1984   Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development,* 7: 49–63.

Yang, Charles D.   2004   Universal grammar, statistics, or both? *Trends in Cognitive Science*, 8: 451- 456.