Speech Perception in Infancy:  
A Foundation for Language Acquisition

Janet F. Werker and Judit Gervain

Abstract
We discuss the development of speech perception and its contribution to the acquisition of the native language(s) during the first year of life, reviewing recent empirical evidence as well as current theoretical debates. We situate the discussion in an epigenetic framework in an attempt to transcend the traditional nature/nurture controversy. As we illustrate, some perceptual and learning mechanisms are best described as experience-expectant processes, embedded in our biology and awaiting minimal environmental input, while others are experience-dependent, emerging as a function of sufficient exposure and learning. We argue for a cascading model of development, whereby the initial biases guide learning and constrain the influence of the environmental input. To illustrate this, we first review the perceptual abilities of newborn infants, then discuss how these broad-based abilities are attuned to the native language at different levels (phonology, syntax, lexicon etc.).

Key Words: phonological/prosodic bootstrapping; experience-expectant and experience-dependent mechanisms; neonates; phoneme perception; word learning; linguistic rhythm; perceptual attunement; rule learning; word order; epigenetics

Key Points
1. Newborns have sophisticated speech perception abilities. They prefer language over complex speech analogs, discriminate rhythmically different languages, prefer to listen to their native language, detect word boundaries, and discriminate most phonemes of the world’s languages. These perceptual abilities lay the foundations for subsequent learning.

2. Perceptual biases present at birth prepare the infant to learn the sound system of any of the world’s languages, and perceptual learning, using the statistical properties of the input speech, helps establish the repertoire of native speech sound categories.

3. Infants start to segment the continuous speech stream into constituent words during the second half of the first year. Initially, they seem to rely primarily on statistical information such as the co-occurrence patterns of phonemes and syllables, which they can extract using general-purpose, universal learning mechanisms.

4. Toward the end of the first year, as they gain more experience with the native language, infants increasingly rely on language-specific cues to word segmentation, such as stress, phonotactics, or phoneme allophony.

5. As they extract an increasing number of potential word forms from the input, infants start to associate these with concrete, perceptually available objects in the environment. This initial associative labeling is the first step toward word learning and the lexicon. Soon after the first birthday, the main cognitive focus of the infant changes from phoneme to word learning. This functional reorganization allows infants to use phoneme-level representations, ignoring other acoustic details that are not relevant for lexical distinctions.

6. During this process, the two levels (i.e., the presentations of word forms and of objects)
interact. Naming objects acts as an invitation to form conceptual categories for these objects, while evidence of a meaning contrast can strengthen the establishment of language-specific phoneme categories.

7. Sensitivity to language structure appears early on. At birth, infants are able to detect identity relations, and they show evidence of rule learning and structural generalizations in the second half of the first year.

8. The perception of the acoustic and phonological properties of speech also plays a role in acquiring language structure, as certain morphosyntactic constructions have phonological and prosodic correlates. Infants are able to exploit these correlations to bootstrap abstract grammatical structure from perceptually available cues.

9. As an example of prosodic bootstrapping, infants are able to use cues such as intensity, pitch, and duration to establish an initial, rudimentary representation of the basic word order of their native language(s), which has well-defined prosodic correlates.

10. Overall, early language acquisition can be characterized by an epigenetically determined set of initial biases, on which specific language experience can build.

Introduction

Acquiring language is one of the most profound feats of human childhood. Although the potential for language acquisition is deeply embedded in our biology, it is the native language or languages that the child will acquire. Long before they understand or utter their first word, children are listening to and watching speakers of their native language. In this chapter we review evidence showing how initial biases and perceptual tuning work together to help launch language acquisition. We situate the evidence within an epigenetic perspective (Gottlieb, 1976). We use this term in an attempt to progress beyond the nature/nurture controversy: genetic underpinnings constrain the range of capabilities that can be manifest, yet the environment is equally important via selection and timing of activation and inactivation of genes that code ultimately for behavior. Using terms introduced by Greenough, Black, and Wallace (1987), we use the label “experience-expectant” to characterize those capabilities that are simply waiting for almost guaranteed environmental input to be set (e.g., some aspects of language development or the maturation of vision over the first few months of life), and we use the label “experience-dependent” to characterize those capabilities that emerge only as a function of learning (e.g., the ability to write or read or the acquisition of the vocabulary of a specific language).

To explain the link between initial biases, experience, and the emergence of an abstract, rule-governed language system, we argue for a cascading model of development. The biases present at the beginning or any subsequent juncture in development direct attention and processing and constrain the influence environmental input can have. In this way, changes in perceptual sensitivities that occur across time play a central role in the emergence of the language-specific representations and structures used in the native language.

We begin our characterization with an examination of the perceptual biases evident in the newborn infant for attending to speech, for discriminating acoustic/phonetic detail, and for apprehending structure. Because the peripheral auditory system is fully developed by 24 to 28 weeks of gestation and some of the properties of language can be processed in utero, by birth the infant has already had several weeks of experience with the native language. We review the ways in which this prenatal listening experience has changed perception, and consider it as an example of “experience-dependent” processes. Core aspects of language processing remain constant, however, even across wide varieties of language input. These will also be described. To capture the biological fact that even these tightly conserved initial sensitivities reflect the joint action of genes and gene activation/inactivation, we will use the term “experience-expectant” to characterize them, but it should be noted that this set of capabilities is largely akin to that referred to as “innate” by others.

From these beginnings, we review the literature on infant language perception across the first months of life, focusing on three broad aspects of language: phonetics/phonology, word learning, and the bootstrapping effects of speech perception on morphosyntactic structure.

Newborns’ Speech Perception Abilities: Experience-Expectant and Experience-Dependent

At birth, human infants show remarkably efficient language-learning propensities that likely involve a combination of both experience-expectant and experience-dependent processes. Newborns prefer to listen to speech over similarly complex
nonspeech sounds (Vouloumanos & Werker, 2004, 2007a). If the speech is filtered to render it more like that which is heard in utero, newborns no longer show a preference (Vouloumanos & Werker, 2007b). Moreover, they do show an equal preference for rhesus monkey calls—a signal they have never before heard—as they do for human speech (Vouloumanos, Hauser, Werker, & Martin, 2010), and prefer both over nonspeech. Although these two sets of findings cannot rule out a role for listening experience in utero, they do provide compelling evidence that the neonatal preference for speech is not merely based on familiarity, raising the possibility that the human infant is broadly prepared for sounds that are speechlike. By 3 months, human infants show a robust preference for human speech over monkey calls (Vouloumanos et al., 2010), revealing experience-based learning after birth.

Neonates’ brains also appear to be specialized for language, as they show an increased response in the left auditory areas to normal, forward-going speech, but not to backward speech (Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002; Peña et al., 2003). This lateralization is seen in both structural and functional measures and is similar to the brain organization found in most (right-handed) adults, whose language areas are predominantly located in the left hemisphere (Dehaene-Lambertz, Hertz-Pannier, Dubois, & Dehaene, 2008). This initial predisposition for human language sets the stage for learning by allowing infants to maximize their exposure to speech.

Importantly, the evidence for structural differences in the posterior part of the left superior temporal cortex so early in life (Dehaene-Lambertz et al., 2008) has been tied to differences in gene expression in the perisylvian areas of the brain (Sun, Collura, Ruvolo, & Walsh, 2006). This may, in concert with other developmental processes, lead to structural differences that support the early development of cortical specialization. The fact that these differences in gene expression predate even prenatal experience with languages helps elucidate what we mean by an epigenetic approach. We are not claiming that specific experience with language is required to set up the perceptual biases seen. Rather, we are emphasizing that even genes whose expression is tightly constrained nevertheless require enabling conditions—in this case likely endogenous not only to the organism but perhaps also to the cell—to be expressed. We believe this point is important to make because it underscores how nature and nurture must be simultaneously considered to describe how biological systems work, thus illustrating how counterproductive it is to pit nature and nurture against one another.

Newborns also show evidence of prenatal learning. They show a specific preference for their native language(s)—that is, for the language(s) their mothers spoke during pregnancy (Mehler et al., 1988; Moon, Cooper, & Fifer, 1993). And they show a preference for their mother’s voice (DeCasper & Fifer, 1980). But yet again, experience-expectant processes play a role as well. Newborns go beyond simple familiarity with a language, as they can discriminate languages they never heard before if those are rhythmically different (Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998). A baby born into a French family can discriminate English from Japanese, for instance. This discrimination ability relies on the different rhythmic properties (proportion of vowels, variability of consonant cluster duration, etc.) of the languages, and it is preserved even if phonological cues other than rhythm (e.g., segment identity) are removed from the stimuli (Ramus, 2002; Ramus, Hauser, Miller, Morris, & Mehler, 2000; Ramus & Mehler, 1999). Recent evidence reveals that bilingual newborns do not show a preference for listening to one of their languages over the other even if the languages are from different rhythmical classes (Byers-Heinlein, Burns, & Werker, 2010). They do, nonetheless, retain the ability to discriminate them (Byers-Heinlein et al., 2010). The experience-dependent listening preference ensures attention to each language, while the remaining experience-expectant ability to discriminate the two languages may help bilingual infants separate the two systems from the start (Ramus, 2002; Ramus & Mehler, 1999; Ramus et al., 2000). Bilinguals growing up with two rhythmically similar languages might use cues other than rhythm to tell their languages apart, as Catalan–Spanish bilinguals have been found to discriminate the two languages at about 4 months of age (Bosch & Sebastián-Gallés, 1997, 2001).

Recent work shows the same evidence of experience-dependent processes complementing experience-expectant ones in neural organization. In a recent study, we extended the work by Dehaene-Lambertz and colleagues (2002) and Peña and colleagues (2003) reviewed three paragraphs above showing a specialized response for forward versus backward speech in the newborn brain, but found it to be particularly strong for the native language—that is, the language heard in utero or immediately following birth (May, Byers-Heinlein,
Gervain, & Werker, 2011). In our initial work, infants were tested on English versus Tagalog, a Filipino language. Moreover, filtered speech was used, raising the possibility that the signal was not sufficiently veridical to fully engage language processing. However, we are finding the same pattern in an ongoing study with English and Spanish (May, Gervain, Carreiras, & Werker, 2011).

Neonates are also sensitive to a number of more subtle phonological cues within a language. They can detect the acoustic correlates of word boundaries (Christophe, Dupoux, Bertoncini, & Mehler, 1994), as they respond differentially to phoneme sequences that are different only in the presence or absence of a word boundary (e.g., the phoneme sequence /mati/ from the single French word mathématicien [“mathematician”] or spanning two words in panorama typique [“typical view”]). They are also sensitive to the position of stress within a word and can discriminate words with different stress patterns (Sansavini, Bertoncini, & Giovanelli, 1997). Furthermore, they can categorize function words and content words on the basis of their different acoustic properties (Shi, Werker, & Morgan, 1999), since function words are phonologically minimal (short, bear no stress, have reduced vowels, etc.; Morgan, Shi, & Allopenna, 1996).

Adult perception organizes speech into hierarchical units: phonological phrases consisting of phonological words, which in turn comprise phonemes, etc. We do not completely understand what units infants use to represent speech. The existing evidence suggests that they are able to represent syllables and weigh this unit more strongly than the phoneme when representing larger units (Mehler, Dommergues, Frauenfelder, & Segui, 1981; Segui, Dupoux, & Mehler, 1992). In these experiments, infants distinguished words that consisted of two and three syllables, but an equal number of phonemes, yet failed to show discrimination when presented with words consisting of an equal number of syllables, but a different number of phonemes. This, of course, does not mean that newborns and young infants are unable to discriminate individual phonemes. In fact, the contrary is true. As the next two sections will discuss, newborns and young infants have remarkable phoneme-discrimination abilities. However, they seem to be using different units for different purposes.

As recent brain imaging data suggest (Gervain, Macagno, Cogoi, Peña, & Mehler, 2008), newborns are also sensitive to the structural regularities in speech. They exhibit increased brain activity in response to simple patterns, such as immediate and identical repetitions of the same syllables within words (e.g., mubaba, penana, etc.), but they cannot detect distant repetitions (e.g., bamuba, napena, etc.). (For a more detailed discussion, see the section entitled “Learning Abstract Regularities: Rules and Perceptual Mechanisms” and Fig. 31.2.)

These perceptual abilities allow very young infants to start cracking the linguistic code by providing efficient mechanisms to keep track of their target language(s) and obtain a first parse of the continuous input into constituent units such as syllables, words, utterances, etc. This lays the foundations for subsequent learning.

From Broad-Based Acoustic/Phonetic Perception to Language-Specific Phonetic Perception

Infants begin life with broad-based sensitivities to phonetic contrasts. For example, infants discriminate similar-sounding vowels (e.g., /i/ and /I/ as in the words “bear” and “bit”) and similar-sounding consonants (e.g., the /b/ vs. /d/ distinction in “big” and “dig” and the /b/ vs. /p/ distinction as in “bat” and “pat”). More importantly, the perceptual sensitivities shown by young infants map onto the phonetic differences that are used in adult languages. At 1 to 4 months infants discriminate the same-sized acoustic difference more easily if the two stimuli on which they are being tested are from either side of a phonetic category boundary used in an adult language (e.g., /b/ vs. /p/) than they do if the two stimuli are from within the same phonetic category (e.g., two physically distinct instances of /p/; Eimas, Siqueland, Jusczyk, & Vigorito, 1971). Moreover, there is organization to the phonetic space of young infants. Young infants will treat as equivalent different pronunciations of the same consonant or vowel when spoken by different voices (Dehaene-Lambertz & Gliga, 2004; Kuhl, 1979; 1980), paying more attention to the phonetic category boundary than to the change in speaker.

The sensitivities shown by young infants map onto an “experience-expectant” pattern, with an organization that supports the learning of the phonetic categories of any particular language. During the first several months of life, “experience-dependent” listening functions to attune these sensitivities in a number of ways. The most commonly described pattern is one of maintenance/decline. This was first illustrated by Werker and Tees (1984), who showed that English infants aged 6 to 8 months can discriminate a difference between two Hindi “d”
sounds that adult English speakers can no longer discriminate. By 10 to 12 months of age the English infants no longer maintained sensitivity to the distinction, although Hindi-learning infants did. This pattern has now been replicated many times with a variety of other consonant and vowel distinctions (Gervain & Werker, 2008), using both behavioral and event-related potential (ERP) tasks, with the reorganization in vowel perception seen at a slightly younger age than it is for consonants (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994). Moreover, the pattern has even been shown to extend to sign language (Baker, Golinkoff, & Petitto, 2006; Palmer, Fais, Golinkoff, & Werker, 2012; Wilbourn & Casasola, 2007). Of interest, bilingual infants ultimately maintain sensitivity to the phonetic distinctions used in each of their native languages (Bosch & Sebastián-Gallés, 2003; Burns, Yoshiida, Hill, & Werker, 2007; Sebastián-Gallés & Bosch, 2009; Sundara, Polka, & Molnár, 2008). Those consonant and/or vowel distinctions that are used in the native language are not only maintained but are also sharpened, with listening experience improving (Kuhl et al., 2006; Sundara, Polka, & Genesee, 2006) and realigning (Burns et al., 2007) discrimination.

Finally, recent research suggests that acoustic salience may also be important. In illustration, although infants divide the nasal consonant space without experience and can discriminate /ma/ from /na/ at a very young age, listening experience may be required to induce discrimination of the acoustically similar and fairly rare (i.e., nonsalient) distinction between /n/ and /ng/ in initial position. This distinction is used in Filipino. At 4 and 6 months, neither English nor Filipino infants can discriminate /na/ from /nga/. By 10 months discrimination is evident in Filipino infants but not in English infants (Narayan, Werker, & Beddor, 2010).

How might infants learn the phonetic categories of their native language? It has been argued that just as in other types of perceptual learning, infants use a form of similarity matching (Best & McRoberts, 2003; Kuhl, 2004). Basically, if Language A has a single phonetic category in an acoustic space where another language, say Language B, has two categories, infants learning Language A may begin to collapse the two categories into the single intermediate category, whereas infants in Language B will continue to maintain the distinction. To empirically test this type of explanation, Maye, Werker, and Gerken (2002) conducted an artificial language-learning study in which infants were familiarized to either a unimodal (mimicking Language A above) or bimodal (like Language B) set of stimuli from along an acoustic continuum and then tested on their ability to discriminate the endpoints. Following only 2.4 minutes of exposure, infants in the bimodal but not the unimodal condition continued to discriminate the two endpoints, indicating that only the infants in the bimodal condition continued to treat the input as two categories. This pattern has now been replicated with a new continuum and has been shown to help infants generalize to a different phonetic distinction, and to help induce discrimination of an otherwise difficult contrast (Maye, Weiss, & Aslin, 2008). Of importance, although distributional learning is effective at 6 to 8 months of age, and continues to be so across the lifespan (Maye & Gerken, 2001), distributional learning is measurably less effective by 10 months of age and beyond, after infants have already begun to establish native phonetic categories (Yoshida, Pons, Maye, & Werker, 2010a). At this time social interaction appears to facilitate learning (Kuhl, Tsao, & Liu, 2003), perhaps due to its impact on attention (Yoshida et al., 2010a).

In summary, the research on phonetic perception in infancy indicates a substantial level of organization prior to significant postnatal listening experience. The perceptual biases that are evident prepare the infant to learn the sound system of any of the world’s languages, and perceptual learning, using the statistical properties of the input speech, helps establish the repertoire of native speech sound categories. In the section entitled “The Beginnings of Meaning: Associating Word Forms with Objects” we review the literature on how perceptual tuning to the phonetic categories of the native language might support recognizing and learning the words of the native language. But before infants can recognize familiar words and learn new words, they have to be able to extract words from the ongoing stream of speech. Hence we turn to a discussion of word segmentation in the next section.

Segmenting, Representing, and Recalling Word Forms

One of the most challenging tasks a young infant faces is to break the continuous speech stream into its constituent units. Speech does not have the systematic and reliable cues to word boundaries that we find in written language: not every word boundary is signaled by pauses, and words rarely appear in isolation. Therefore, a growing body of research has been directed to understanding how babies might
find the word forms in ongoing speech. Two types of cues have been proposed. Universal cues rely on the statistical properties of speech, which are believed to work in all natural languages in approximately the same way, so these cues can guide infants in segmenting speech from early on. Language-specific cues require some learning about the phonological properties of the target language, and are thus assumed to be operational at a later stage of speech segmentation and word learning.

The most widely recognized statistical cue to segmentation is conditional probability—that is, the probability of a unit (syllable, phoneme, etc.) to appear before or after another unit. Phonemes or syllables within the same word predict each other with a higher conditional probability than two units spanning a word boundary (Harris, 1955; Shannon, 1948). For instance, in the sequence pretty baby, the forward-going transitional probability1, a type of conditional probability measure, is higher between the syllables pre and tty than between tty and ba, since the word pretty can be followed by a large number of words, making it hard to predict the first syllable of the next word. Building on a seminal study by Hayes and Clark (1970), where adults were shown to be able to segment a continuous stream of artificial speech analogs on the basis of conditional probabilities, Saffran, Aslin, and Newport (1996) found that 8-month-old infants have the same ability, successfully segmenting a continuous artificial language into its constituent words using statistical cues after only a few minutes of exposure. This discovery gave rise to a large number of subsequent studies that aimed to explore the nature of this statistical segmentation mechanism (Bonatti, Peña, Nespor, & Mehler, 2005; Endress & Bonatti, 2007; Fiser & Aslin, 2002, 2005; Hauser, Newport, & Aslin, 2001; Nazzi, Paterson, & Karmiloff-Smith, 2003; Newport & Aslin, 2004; Newport, Hauser, Spaepen, & Aslin, 2004; Peña, Bonatti, Nespor, & Mehler, 2002; Saffran & Wilson, 2003; Thiessen & Saffran, 2003; Toro & Trobalon, 2005; Yang, 2004).

Interestingly, this statistical learning ability is triggered by specific aspects of the signal such as its continuous, nonsegmented nature. In a study with adults Peña and colleagues (2002) showed that adults readily compute statistics over adjacent and nonadjacent syllables, using this information to find words in the input, if the speech stream is continuous, but extract and generalize structural regularities if the speech stream is segmented by subliminal pauses. Recently, Marchetto and Bonatti (under revision) have shown that 12-month-old (but not 7-month-old) infants have the same sensitivity to the properties of the input stream, finding words in continuous streams, but extracting structural regularities when the input is presegmented. This might reflect a division of labor between mechanisms used to learn words and those used to acquire the morphological structure of a language, since these are intimately related, especially in morphologically rich languages, where learning a word might require a prior morphological decomposition into constituent morphemes.

As infants gain increasing knowledge about their native language, they start using some of the phonological cues correlated with word boundaries within the target language. At least three such cues have been identified: stress, phonotactics, and allometry. In English, for instance, content words typically have a strong–weak (trochaic) stress pattern (e.g., ‘doctor). Consequently, infants might use a stressed syllable at the beginning of a bisyllabic unit as a cue to segmentation. Indeed, infants have been shown to develop sensitivity to the stress patterns of their native language between 6 and 9 months (Jusczyk & Aslin, 1993; Morgan, 1996; Morgan & Saffran, 1995). The Metrical Segmentation Strategy (Cutler, 1994; Cutler & Carter, 1987), which relies on stressed-based cues, has been shown to underlie 7.5-month-old English-learning infants’ recognition of familiar words. Jusczyk and colleagues (Jusczyk, Houston, & Newsome, 1999) have shown that when familiarized with trochaic English words (e.g., ‘doctor, ‘candle), 7.5-month-olds prefer passages containing these words over passages that do not contain them. This preference is specific to the trochaic word form. Passages containing the first strong syllables of the words (e.g., dock, can) did not induce any preference. Infants are also able to apply these cues to an ongoing speech stream to extract words. When presented with a continuous stream of consonant vowel (CV) syllables where every third syllable was stressed, 7- and 9-month-olds treated as familiar only those trisyllabic sequences that had initial stress (strong weak weak). Infants showed no recognition of trisyllabic sequences that were not trochaic (WSW or WSS; Curtin, Mintz, & Byrd, 2001). Another prediction of the metric segmentation strategy is that weak–strong (i.e., iambic) words (e.g., gu-tar) might initially be missegmented. Jusczyk, Houston, and colleagues (1999) found exactly this pattern of results.

Phonotactics (i.e., the regularities guiding possible and impossible phoneme combinations in a
language) provide another language-specific cue to segmentation. Knowing that the sequence /br/ is frequent in the initial positions of English words, while /nt/ is common in final positions, can help infants detect word boundaries. Saffran and Thiessen (2003) tested the acquisition of phonotactic constraints using segmentation as the experimental task. During the familiarization phase, they taught infants a rule concerning permissible syllable forms (words consisted of only CV or only CVC syllables; e.g., boda and bikrub, respectively) and a rule relating to permissible consonantal distributions (voiced stop consonants in syllable onsets, unvoiced stops in codas; e.g., dakdot, or vice versa). They found successful segmentation in the test phase in both cases. Using a different approach, Mattys, Jusczyk, Luce, and Morgan (1999) explored how 9-month-old English-learning infants’ knowledge of English phonotactics helps them posit word boundaries. They familiarized English-speaking infants with nonsense CVCCV words, in which the CC cluster is either frequent word-internally but infrequent across word boundaries (e.g., /tʃk/), or vice versa (e.g., /tʃt/). Infants segmented the nonsense words into two monosyllables where the CC cluster was infrequent word-internally and frequent across word boundaries. No segmentation was observed for the other type of CC clusters. Taken together, these results suggest that 9-month-old infants can use their knowledge of the native phonotactics to assist them in word segmentation (Mattys & Jusczyk, 2001a, 2001b).

A third language-specific segmentation cue comes from the distributions of allophones in different positions within words. In English, for instance, aspirated stop consonants appear in the initial positions of stressed syllables (Church, 1987), and their unaspirated allophones appear elsewhere. This distribution makes aspirated stops good cues to word beginnings. It is not implausible that infants might use such cues for segmentation, because infants as young as 2 months are able to discriminate the different allophones of a phoneme (Hohne & Jusczyk, 1994). Indeed, Jusczyk, Hohne, and Bauman (1999) have shown that at 9 months, infants are able to posit word boundaries (e.g., night rates vs. nitrates) based on allophonic and distributional cues, and at 10.5 months, allophonic cues alone were found to be sufficient for successful segmentation.

The Beginnings of Meaning: Associating Word Forms with Objects

In addition to extracting word forms from continuous speech, infants also face the (possibly related) challenge of associating these word forms with meanings to build a full vocabulary. To achieve this, infants need to be able to ultimately form stable, detailed, and permanent representations of the word forms they segmented out from speech, and they need to have the ability to link these forms to entities in the world (i.e., to treat word forms as having a referent). This requires sophisticated conceptual abilities, such as categorization, understanding referentiality, solving the induction problem for meaning (Quine, 1960), etc. These are issues that we are not discussing here. Rather, we will concentrate on the simplest and developmentally earliest case when a word gets reliably associated with a perceptually available, concrete object. Importantly, while learning the word forms and selecting the right meanings have always been regarded as challenging for learners, associating already-learned words with already-existing meanings/concepts was for a long time regarded as fairly straightforward, and especially not involving too much interaction or readjustment between the two levels. In the past decade, however, new evidence emerged showing both that labeling can induce object categorization that doesn’t happen in the absence of labels (Waxman, 2001; Waxman & Lidz, 2006) and that the presence or absence of objects influences the amount of phonetic detail that infants use to represent word forms (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002; Yeung & Werker, 2009).

One of the first studies to suggest that simply learning language-specific phonetic representations might not be sufficient for directing (associative) word learning was Stager and Werker (1997). The authors used a minimal phonetic contrast, bih and dib, and tested whether infants detected this contrast in a labeling and a discrimination situation. Infants at 14 months were able to discriminate the two words when those were presented with a non-nameable (non-object-like) visual display (a checkerboard), so the context required discrimination, whereas they failed to distinguish the two words in a labeling context, when each was first associated with a different object, and then the labels were switched around. In fact, 14-month-olds, but not 8-month-olds (Stager & Werker, 1997) or 17-month-olds (Werker et al., 2002), failed to notice this minimal switch, but even 14-month-olds succeeded when the two labels were more distinct (i.e., lif and neem).

The authors interpreted these results as a sign of functional reorganization, as the main cognitive focus of the infant changes from speech perception
to word learning. During the developmental process, language-specific representations are employed differently in different contexts. In an associative word-learning task the infant needs to pay attention to the details of both the word form and of the object and also establish an association between them. But at this early stage of word learning, although able to discriminate language-specific phonetic differences, infants might not yet give these differences any more weight than they give to other properties of the words, such as the affect in the voice, the gender of the speaker, etc. And indeed, there is evidence that in both word recognition and early word-learning tasks, infants can fail to treat as equivalent two instances of the same word if they are spoken in a different voice, different affective tone, etc. (Newman, 2008; Singh, Morgan, & White, 2004; White & Morgan, 2008). Perhaps one reason, then, that minimal pair associative word learning takes up so much of infants’ cognitive resources is that they encode all the properties of the word and try to hold them all in memory. The result is a U-shaped developmental curve, with 8-month-olds, who are not yet in the word-learning phase, succeeding in the task by performing a more simple associative discrimination task, 14-month-olds, in transition toward robust word learning, failing, and 17-month-olds, experienced word learners, succeeding again.

Further confirmation for the cognitive load hypothesis comes from several recent studies showing that if the cognitive load is decreased, infants are more likely to succeed even at 14 months in associating different objects with minimally different labels. As Fennell and Werker (2003) have shown, 14-month-olds have no problem noticing a switch from the known word ball to the known word doll. This confirms the suggestion that it is the acquisition of object-label associations that underlies 14-month-olds’ difficulty with fine phonological information, not a general inability to represent such information.

Indeed, if the infants are prefamiliarized with the objects at home for several weeks and hence do not need to learn both the words and the objects in the experimental situation, they succeed at using the phonetic details in the words in the switch task (Fennell & Werker, 2004). Infants also succeed if they are offered a visual choice at test, hence reducing the memory demands (Yoshida, Fennell, Swingley, & Werker, 2009). And finally, infants succeed if the phonetic differences are more acoustically salient (Curth, 2010; Curtin, Fennell, & Escudero, 2009), or if they have been taught the words in a task that highlights which sound differences are important (Thiessen, 2007).

Still unexplained from the cognitive resource limitation hypothesis is why infants are able to succeed at 17 months and beyond. We (Curtin & Werker, 2007; Werker & Curtin, 2005) and others (Nazzi & Bertoncini, 2003; Newman, 2008) have suggested that by 17 to 18 months, language-specific phonetic sensitivities have become organized into more stable, “phonological” representations that have begun to act much like the phoneme categories adults use to guide word learning and word recognition. With such a stable representation, infants’ attention is weighted toward the phonetic, allowing them to more easily summarize across irrelevant information such as gender and affect in the voice, and attend instead to the criteria phonemic difference.

There is increasing research to support this hypothesis. One recent study (Dietrich, Swingley, & Werker, 2007) compared English and Dutch toddlers on their ability to learn the minimal pair words “tam” and “taam” that differ in only a vowel length distinction. Of importance, vowel length is phonemic (used to contrast meaning between two words) in Dutch but not in English. Although not phonemic, the difference between a long and short vowel is so acoustically salient that even English-learning infants and toddlers can discriminate the difference in perceptual discrimination tasks that do not involve meaning (Mugitani et al., 2009). However, if stable phonological representations rather than phonetic differences guide word learning at 18 months, only Dutch infants should use the vowel length difference in a word-learning task. And indeed this is what was found using the switch task (Fig. 31.1). Dutch infants succeeded but English infants failed. As an important control, it was shown that English infants aged 18 months nonetheless succeeded when a vowel difference that is phonemic in English was used (“tam” vs. “tem”). In a more recent study it was shown that by the same age, 18 months, English infants will generalize across different accents to treat different pronunciations of the same vowel as equivalent (Best, Tyler, Gooding, Orlando, & Quann, 2009). Thus, by 18 months, infants use phonological representations, perhaps abstract phoneme categories that are part of the grammar of the native language, rather than simple perceptual sensitivities, to guide word learning.

How do infants develop these stable phonemic representations? We suggest it is with the establishment of a sizeable enough lexicon linking phonetic
categories to word forms that the infant is able to pull out—either through emergent statistical learning or through a more abstract generalization—the knowledge that certain sound differences matter more than others (Curtin & Werker, 2007; Werker et al., 2002; Werker & Curtin, 2005).

The above results suggest that the ability to map between labels and objects undergoes considerable reorganization at around 14 months of age. By that time, infants’ native phonetic categories are in place, so labeling seems independent of the establishment of the basic phonemic repertoire of the native language. Indeed, 8-month-olds appear unaffected by the presence of objects, and we reviewed earlier how simple statistical learning might be sufficient for establishing language-specific phonetic categories.

However, Yeung and Werker (2009) have recently shown that labeling might have a role to play in the establishment of the language-specific categories themselves. The idea that meaning contrasts might signal phoneme status (i.e., /p/ and /b/ are distinct phonemes in English, because words like pin and bin contrast in meaning) has been broadly accepted in the linguistic literature. However, this principle was believed to be an unlikely learning mechanism for phoneme contrasts because infants develop their native phonetic category repertoire before they have a sizeable lexicon. While this observation remains, in general, true, Yeung and Werker (2009) have shown that the presence of a nameable object can indeed induce the establishment of native-language phonetic categories in 9-month-old infants, an age at which the native repertoire is already taking shape but universal discrimination hasn’t been lost yet. The authors have shown that when a nonnative phonetic contrast is presented with labels in a consistent manner (i.e., one phone always associated with one object, the other phone with the other object), 9-month-olds were able to discriminate the two. They failed, however, when labeling was inconsistent (i.e., when both phones appeared with both objects), suggesting that the difference between the sounds did not indicate a contrast in meaning. These results indicate that labeling itself might influence the kind of speech sound representations that are being established, implying that speech perception and word learning might be more intimately related, and from the beginnings of the establishment of the native phonetic repertoire, than previously believed.

The connection between objects and speech also works in the opposite direction. As research by Waxman and Markow (1995) suggests, speech might act as an “invitation to form categories.” As a large body of research has shown (e.g., Fulkerson & Waxman, 2007; Waxman & Markow, 1995), infants treat linguistic labels, but not nonlinguistic sounds, as referring to categories of objects rather than just to the individual object the label was presented with. Infants as young as 6 months of age were able to form the noun category “dinosaur” when presented with visually different instances of dinosaurs paired with naming phrases (Look at the toma! Do you see the toma?). If, however, the dinosaurs were paired with tone sequences, no categorization occurred.

These results suggest a strong relation between speech perception and conceptual organization within the domain of word learning, especially during the initial stages when concrete nouns are learned.

Learning Abstract Regularities: Rules and Perceptual Mechanisms

A question that has gained significant theoretical importance in the past decade is how speech perception contributes to the acquisition of language structure. Above, we reviewed research suggesting that phonological categories likely emerge only in the second year of life, yet these categories do have a foundation in the experience-expectant categories present earlier in perceptual development. When experimental research on language acquisition began about 50 years ago, the accepted view in generative linguistics, as well as other approaches to language acquisition, held that the grammar of a language is largely independent of its phonetics, phonology, and prosody, so speech perception was
not believed to play an important role. This view changed in the 1990s, when it was discovered that infants might use the existing correlations between the phonological aspects of language and its structure, well known from language typology, to solve what Pinker (1984) termed the “linking problem.” The linking problem means that even if infants are born with experience-expectant mental representations to facilitate language learning (e.g., a universal grammar), these abstract characterizations of language structure use rules and discrete, abstract categories such as noun, verb, subject, or predicate, and cannot be linked directly to the continuous, physical speech (or sign) input. If, however, infants can somehow exploit the sound–structure correspondences that exist in language, they might be able to use them to bootstrap abstract language structure.

One of the first studies that directly addressed whether infants were able to extract underlying regularities from an artificial grammar was Gomez and Gerken’s (1999) work. In their study, Gomez and Gerken (1999) tested infants in a version of Reber’s classical artificial grammar-learning task (Reber, 1967, 1969), asking whether infants were able to extract regularities and whether they were able to generalize them (i.e., transfer them to a new vocabulary). The authors found that 1-year-old infants successfully discriminated grammatical from ungrammatical strings in several different experimental conditions, including those that required generalization by implementing the regularities in a new vocabulary at test.

Marcus, Vijayan, Rao, and Vishton (1999) used a different paradigm to show that 7-month-old infants were able to learn and generalize abstract rules containing identity relations (e.g., “wo fe fe,” [ABB] or “wo fe wo” [ABA]), taking this as evidence for infants’ ability to represent abstract variables and algebraic rules.

Since the auditory system is mature from about 24 to 28 weeks of gestation, sensitivity to identity relations and repetitions might be operational early on. Indeed, a brain imaging study with newborn infants (Gervain et al., 2008) found that the left temporal and frontal areas responded differentially to syllable sequences containing adjacent repetitions (“mubaba,” “penana”: ABB) as opposed to random sequences (“mubage,” “penaku”: ABC) from the very first trials of the study, suggesting that an automatic repetition-detecting mechanism might be at work (Fig. 31.2). This differential response increased over the time course of the experiment, especially in the prefrontal brain areas, indicating that some form of learning or memory-trace formation might take place. Interestingly, no such differential response was found when nonadjacent repetitions (ABA) were pitted against random ABC sequences.

These results suggest that newborns have experience-expectant perceptual processing mechanisms that allow them to detect and represent structural regularities in speech from the very beginning of linguistic development.

Learning Language-Specific Grammatical Structures: The Case of Word Order

Another perceptual mechanism that has been proposed to play a role in the acquisition of language structure is the auditory grouping principle known as the Iambic/Trochaic Law, originally proposed for music and nonlinguistic sounds. According to this law, sound sequences in which the elements contrast in intensity and/or pitch are perceived as forming units with initial prominence (trochees), while elements contrasting in duration form units with final prominence (iambics) (Hayes, 1995). This grouping principle has been claimed to help infants bootstrap word order (Nespor, Guasti, & Christophe, 1996; Nespor & Vogel, 1986), as there is a correlation between the type of grouping a language uses and its word order. Object–Verb (OV) languages, such as Japanese, Turkish, Basque, etc., have initial prominence in their phonological phrases, and it is realized as increased pitch and intensity, whereas VO languages like English, French, etc., implement phrase-final prominence using lengthening. This has been shown (Nespor et al., 2008) to hold true both across languages (e.g., French vs. Turkish) and within a language that has phrases with mixed word orders (e.g., German and Dutch, where the verb phrase [VP] shows OV or VO order depending on the syntactic context). The reason why this grouping principle is believed to be a particularly useful bootstrapping cue is because it is carried by universal acoustic cues (intensity/pitch vs. duration) automatically processed by the auditory system. Consequently, it might be operational from early on (Nespor et al., 1996).

Current research on the Iambic/Trochaic Law has focused on two issues: (1) whether the law operates at the level of the word, the phrase, or both, and (2) whether the law is inherent to the auditory system or whether it emerges as a consequence of experience with the native language(s).

The original formulation of the principle claimed that it is independent of language and applies to speech and nonspeech sounds alike (for summary,
see Hayes, 1995). Indeed, a study by Hay and Diehl (2007) found no difference in grouping preference between English- and French-speaking adults despite the fact that English has a predominantly trochaic word-level stress pattern, whereas French is iambic. In this experiment, the authors exposed participants to sequences of syllables or sine wave segments (i.e., speech or nonspeech) that alternated either in intensity or duration, and asked participants to indicate whether they heard a strong–weak (trochaic) or a weak–strong (iambic) pattern. Both English and French participants heard a strong–weak pattern when intensity was manipulated and a weak–strong pattern when the contrast was durational. The authors concluded that the Iambic/Trochaic Law was a general auditory principle. More recently, however, Iversen, Patel, and Ohgushi (2008) tested English and Japanese participants and found that while English speakers had the predicted grouping preferences for both the intensity contrast (trochaic) and the durational contrast (iambic), Japanese participants only showed a predicted trochaic contrast for the intensity-manipulated tone sequences but had no preference when the manipulation involved duration. Further analyzing their data, the authors showed that this null preference is due to a bimodal pattern of responses in the Japanese group. One sub-set of the Japanese participants had an unpredicted long–short preference. The authors concluded that this reflects a language-specific bias, indicating that

![Figure 31.2. Newborns can discriminate adjacent repetitions (ABB) from random speech sequences (ABC). A. Experimental procedure. B. Probe placement. C. Hemodynamic responses obtained for the ABB and ABC stimuli (plotting follows probe placement illustrated in B). Reproduced from Gervain et al. (2008) with permission.](image-url)
the Iambic/Trochaic Law emerges as a result of language experience.

The above-discussed studies seem to point in different directions, leaving the question about the origins of the grouping principle unanswered. However, two issues need to be considered. First, the two language pairs differ in the prosodic unit involved in the comparison. English and French have different rhythmic patterns at the word level, but both have VO (i.e., iambic rhythm) at the phrasal level. Japanese and English, by contrast, differ at the phrasal level, since Japanese has an OV-type (i.e., trochaic) prosody. Second, Iversen and colleagues (2008) found a bimodal response pattern, which means that one subset of their participants had an unpredicted preference for the durational contrast, but the other, though smaller, subset showed the expected short–long grouping preference. Taken together, the two adult experiments, as well as the above considerations, might be best understood as indicating that just as the iambic/trochaic law predicts, there is a universal and language-independent grouping preference assigning initial prominence to intensity contrasts and a final prominence to durational contrasts, as suggested by Iversen and colleagues’ (2008) English participants and the smaller subset of the Japanese participants. In addition to this language-independent perceptual bias, there might exist another, language-based grouping principle that mirrors the rhythmic patterns found in the native language, as Iversen and colleagues’ (2008) larger Japanese subgroup suggests. In this context, Hay and Diehl’s (2007) findings might further confirm the existence of a universal bias, or alternatively, they can be taken as evidence that if cross-linguistic grouping preferences differ, the relevant level of analysis is that of the phonological phrase, where English and French are similar, not that of the word, where these two languages differ.

The above hypothesis can be tested developmentally, as the inherent and the emergent views on the Iambic/Trochaic Law make different developmental predictions: universal, experience-expectant perceptual principles can be expected to appear earlier in development than experience-dependent language-specific biases. To date, only a handful of studies have been conducted with infants, and the results are not conclusive. The position of phrasal prominence as a cue was first tested with infants in the context of language acquisition by Christophe, Nespor, Guasti, and Van Ooyen (2003), who showed that 6- to 12-week-old infants were able to discriminate two languages that have opposite word orders and opposite prosodic patterns (French and Turkish), solely on the basis of the different locations of prosodic prominence in the two languages. This was achieved by resynthesizing natural utterances in the two languages, suppressing segmental identity and all phonological properties other than the location of the prosodic prominence. This study demonstrates that even very young infants can discriminate the relevant prosodic properties, although it doesn’t address the question of the origin of the principle directly.

This issue was recently investigated by a number of studies that arrived at conflicting results. Yoshida and colleagues (2010b) exposed English and Japanese infants to alternating sequences of short and long tones. At age 5 to 6 months, neither group showed a preference, but by 7 to 8 months, the English babies exhibited longer looking times to long–short test sequences as opposed to short–long ones, which can be interpreted as a novelty preference. Crucially, the Japanese infants showed the opposite pattern, although their preference was weaker than that of the English-speaking babies. Intensity was not tested in this study, but the results suggest that at least an iambic grouping preference for duration might emerge as a result of experience with language.

Hay and Saffran (2008) found similar results with regard to duration in a series of experiments testing how statistical learning (Saffran et al., 1996) and the Iambic/Trochaic Law might interact. They observed that 6-month-old English infants showed no preference when durational cues were pitted against statistical cues, whereas 9-month-olds relied more heavily on durational cues than statistical cues, irrespectively of whether syllables or tones were used as stimuli. A very different developmental trajectory emerged, however, for intensity as a cue. Both 6- and 9-month-old English babies were able to use intensity as a cue to word onsets in the segmentation task.

Similarly, Bion and colleagues (2011) also found that pitch contrasts resulted in the predicted (high–low) grouping of simple syllables in 7-month-old infants, while durational contrasts did not induce any grouping. Unlike infants, the Italian-speaking adults in this study showed both the predicted (high–low) grouping when syllables of alternating pitch were used, as well as the predicted (short–long) grouping when syllables alternated in duration.

Surprisingly, other infant studies found the opposite pattern of results. In Trainor and Adams’s (2000) study, 8-month-olds as well as adults treated
lengthening as a cue to the end of groups of tones, whereas intensity did not result in systematic grouping at either age.

Interestingly, in music perception longer duration and lower pitch have been shown to reliably indicate the final boundaries of tonal groups in infants as young as 4.5 months (Jusczyk & Krumhansl, 1993). This was true even when other factors, such as the presence or absence of a large durational or pitch discontinuity across boundaries, were controlled for. In this study, however, pitch and duration were used conjointly as cues to boundaries.

In an experiment that directly addressed the role of the grouping bias as a cue to word order, Gervain and Werker (2009; under revision) found evidence that 7-month-old bilingual infants exposed to one VO (English) and one OV (Japanese, Hindi, Punjabi, Farsi, or Korean) language were able to use pitch and duration as cues to word order. When exposed to an artificial grammar with VO (iambic) prosody, realized as a durational contrast, bilingual infants preferred test items with VO word order over test items with the opposite word-order pattern. When pitch (i.e., OV prosody) was used during familiarization, the opposite pattern was found: infants showed a preference for the OV order.

Taken together, these studies suggest that at around 7 to 9 months, language-specific grouping biases emerge, as suggested by our hypothesis above. However, results are relatively weak, and studies are inconsistent as to which of the two cues is more reliable. It seems clear, though, that at a younger age, roughly between 5 and 7 months, grouping preferences seems to be absent. This appears to contradict the first tenet of our interpretation of the adult data, namely that there exists an initial and universal grouping principle. However, the youngest ages from 0 to 4 months have not been tested in any of the studies. Thus, it is not impossible that the iambic/trochaic grouping principle follows a U-shaped developmental trajectory, with an initial universal bias that disappears before 5 months of age, and a new grouping principle that reemerges, partly as a result of language experience. Further research into newborns’ rhythmic grouping abilities is needed to clarify this issue.

Conclusion

Language acquisition begins long before the first word. In this chapter we have provided an overview of recent research on infant speech perception—initial biases, mechanisms of change, and the relation between perception and language use—as a means to better understand the ontogeny of human language. We have tried throughout to distinguish between those initial experience-expectant perceptual sensitivities and representations that are common to all human infants, irrespective of the precise prenatal or early experience encountered, and those experience-dependent sensitivities and representations that reflect the operation of perceptual tuning or learning. We argue that all initial biases, whether common to the species or influenced by prenatal listening experience, reflect an epigenetic process wherein endogenous (factors within the organism, such as trophic factors, the intercellular matrix, etc.) or exogenous (factors in the world that affect the developing child through ingestion, perception, etc.) variables influence the expression of genetic potential.

Our review makes clear that infants begin life with an epigenetically determined set of linguistic biases that form a foundation on which specific language experience can build. Within this constrained set, some language-specific tuning appears to require only limited exposure during a narrow window in development, and is then relatively permanently set (experience-expectant). Other patterns of change appear to require more exposure in order to change, and appear more open to continued learning even beyond the infancy period (experience-dependent). Enormous progress has been made on characterizing each of these, thus advancing considerably our understanding of the initial foundations that are in place to serve language acquisition.

What remains more controversial is just how changing perceptual sensitivities and early representations bootstrap language acquisition. Are the initial perceptual primitives the first bits of structure, which, when tuned by experience with the native language, constitute actual linguistic representations that contribute to an abstract linguistic system? Or are the initial sensitivities more broadly shared across perceptual systems, and only later usurped by the linguistic system for specific uses? The literature reviewed provides support for each. For example, in the realm of phonetic perception there seem to be initial sensitivities that set the range of phonetic distinctions that can be learned. According to one strand of data, distributional (statistical) learning leads to the emergence of language-specific perceptual categories that only after linking to meaning lead to a second-order generalization of phoneme categories that can then be used to direct word learning. Yet, according to another strand of data, learning about native-language phonetic categories
unfolds via acquired distinctiveness, wherein the pairing of sounds and objects, even before word learning, helps establish the category structure of each. Similarly, in the case of the iambic/trochaic law, there are different results from different studies. This suggests either a very general perceptual sensitivity to pitch or loudness in stress-initial groupings and to duration in stress-final position, which must then be specified for language use. Alternatively, it might suggest a different unfolding for these sensitivities within and outside the domain of linguistic stimuli. The challenge of future work is to tease apart these apparent differences, carefully investigating the generality, specificity, and utilization of initial and emergent representations. Only in this way can we fully understand not only when, but how, the changing perceptual system supports and directs the acquisition of language.

Questions for Future Research
1. How specific are initial perceptual biases and abilities to language? Are they part of the language faculty or are they shared across different perceptual and cognitive abilities?
2. How do these perceptual and learning processes interact with experience in multilingual environments, where the input contains several languages?
3. How is language organized in the newborn and infant brain? Does the attunement process from broad to narrower speech-perception abilities accompanied by functional reorganization in the infant brain?
4. How does the conceptual system interact with language acquisition (e.g., in word learning or early sentence processing)?
5. Is the prosodic bootstrapping mechanism contributing to the acquisition of word order a language-independent general auditory bias or is it the result of experience with the native language?

Notes
1. This work was supported by an ANR Jeunes Chercheurs Jeunes Chercheuses grant (n° 21373) to JG.
2. TP(A→B) = P(AB)/P(A), where TP(A→B) is the transition probability between adjacent units A and B. P(X) is the probability of occurrence of unit X.

References


Marchetto, E., & Bonatti, L. L. (under revision). Finding words in fluent speech by 7-12-month-old infants (pp. 1–11).


