No cover image available

The Oxford Handbook of the Mental Lexicon

Anna Papafragou (ed.) et al.

https://doi.org/10.1093/oxfordhb/9780198845003.001.0001

Published: 2022 **Online ISBN:** 9780191880292 **Print ISBN:** 9780198845003

CHAPTER

13 Infants' Learning of Speech Sounds and Word-Forms

Daniel Swingley

https://doi.org/10.1093/oxfordhb/9780198845003.013.6 Pages 267–C13.P75

Published: 14 February 2022

Abstract

How do infants start learning their native language? This chapter reviews the conventional understanding of this problem, illustrated by a review of the most important studies in this area, and suggests that this conventional understanding mischaracterizes the problem infants solve and the developmental process by which they solve it. Recent experimental and modeling work from several labs suggest new ways to think about the beginnings of language learning and the emergence of the lexicon.

Keywords: perception, categorization, learning, infant language, words, phonemes

Subject: Semantics, Psycholinguistics, Linguistics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

13.1 Introduction

WE speak to babies from the moment they are born. To the infant, the sound of the human voice is already familiar from prenatal experience (e.g., DeCasper, Lecanuet, Busnel, Granier-Deferre, and Maugeais, 1994; Kisilevsky, Hains, Lee et al., 2003), but after birth, the voice plays a new role. Babies see us leaning in toward them, looking in their eyes, and singing, cooing, or speaking. They notice how our own vocalizations can be timed to the infant's actions, and how our faces move when we talk (Guellaï, Streri, Chopin, Rider, and Kitamura, 2016). What do they think we are doing when we speak? They probably think the melody is the message, as Fernald (1989) put it: infants seem to resonate to the emotional meaning of our intonation contours, at least for some intonation patterns, probably without having to learn to do so. Parents are adept at using the infant-directed speech register to keep infants' attention and modulate their emotional state (e.g., Fernald, 1992; Lewis, 1936; Meumann, 1902; Spinelli and Mesman, 2018). To the extent that young babies reflect on language at all, they might begin with the hypothesis that speech is similar to cuddling—auditory rather than tactile, but nonetheless primarily a personal and intimate source of emotional support.

Of course, this hypothesis is decidedly incomplete as a description of what speech does, covering almost none of what we think of when considering language. In learning a particular language, additional speech features come to the fore: sets of consonants and vowels, their nature when appearing in context, the longer and more complex sequences that form words, and most generally the partitioning of phonetic variation into its myriad causal sources. Infants make progress in all of these areas during their first year, and also
p. 268 make substantial headway in learning aspects of the meanings of ↳ many early words. Here, we outline what is known, and what is still unknown, about how this process typically unfolds in early development.

13.2 Infants' categorization of speech sounds

By common consensus, this story begins in 1971 with the publication of a paper by Eimas, Siqueland, Jusczyk, and Vigorito examining 1- and 4-month-old infants' discrimination of the syllables /ba/ and /da/. They chose to test infants on these syllables for an interesting and theoretically significant reason. Researchers at Haskins Labs had developed a procedure for synthesizing consonant-vowel syllables, allowing precise control over the syllables' acoustic features. They found that they could simulate the phonetics of the long voice onset time appropriate for stop consonants like /p/ or /t/ by manipulating two features of the resonances (formants) leading into the following vowel: silencing the first formant, and replacing the second and third formants with noise (Liberman, Delattre, and Cooper, 1958). The short voice onset time appropriate for a /b/ could be synthesized similarly, by effecting these alterations for a shorter period (thereby allowing voicing to begin more closely following the release of the consonant). Remarkably, gradually increasing the voice onset time between the two did not gradually increase the likelihood that a given syllable would be heard as voiceless; instead, listeners consistently perceived the syllable as 'ba' until voice onset time reached about 25 ms, at which point most responses went to 'pa.'

This phenomenon, in which perceptual discrimination is governed by the stimulus sitting either left or right of a boundary point on a continuum, rather than by the acoustic distance along that continuum, came to be known as *categorical perception* (Liberman, Harris, Hoffman, and Griffith, 1957; Repp, 1984). The early Haskins papers acknowledged uncertainty regarding where the boundaries might come from: perhaps they were innate, or perhaps they were learned; if learned, mature listeners may once have been highly sensitive to many decision boundaries in phonetic space and learned to disregard some of them; or rather quite poor at the start and learned to sharpen discrimination at these boundaries (Liberman, Harris, Kinney, and Lane, 1961).

Eimas et al. (1971) aimed to address this question of origins, testing infants of one and four months of age. They used a recently developed habituation technique (Siqueland and DeLucia, 1969), in which infants' sucking on a pacifier was rewarded, in this case by presentation of syllables; once the initial reward-prompted increase in sucking rate tailed off (habituation), the habituation syllable was replaced by another, with a voice onset time 20 ms different from the initial one; or it was kept the same, for infants in the control group. When the habituation and test stimuli straddled the ~25 ms boundary (at 20 and 40 ms), sucking rates rebounded; when the stimuli sat on the same side of the boundary (at -20 and 0 ms or at 60
p. 269 and 80 ms), sucking rates continued to sink after the ↳ change, just as much as in the no-change control condition. Eimas et al. concluded, " ... the means by which categorical perception ... is accomplished may well be part of the biological makeup of the organism and, moreover ... must be operative at an unexpectedly early age" (p. 306).

The usual interpretation of the Eimas et al. (1971) result is that languages tend to settle into a position where at least some of the phonological distinctions required for differentiating words come naturally to the auditory system, presumably reducing the burden on learning and helping make speech interpretation more accurate in the mature state. These nonlinearities in discrimination may not really be adaptations for

spoken language per se, because similar nonlinearities are found in other mammals who do not produce consonants (e.g., Kuhl and Miller, 1978). This is why these perceptual boundary phenomena are viewed as influencing how languages evolve over historical time, rather than as human evolutionary adaptations to some fixed set of phonetic standards.

Additional patterns of selective sensitivity were found in subsequent studies of infants, showing that the alignment between naive speech perception and language structure is not unique to the voicing distinction in stop consonants. For example, the difference between /d/ and /ɡ/ at the start of a syllable is signaled primarily by changes in vocal tract harmonics (the second and third formants) during the transition from the consonant into the vowel. Eimas (1975) used synthesized syllables that adults identified as either [dæ] or [ɡæ] to test their discriminability to 2- and 3-month-olds, and found again that infants distinguished the sounds that were considered different by adults, but did not distinguish sounds that adults rated as falling within the [dæ] category.

Similar conclusions were drawn from other early studies showing that infants treat speech variation in ways that would apparently facilitate linguistic categorization. For example, a major part of the difference between /b/ and /w/ is the speed of the formant transitions into the vowel: fast for /b/, slower for /w/. What counts as "fast" or "slower" depends on speaking rate. If the speaker is talking quickly, a /b/'s transitions must be especially speedy, or the consonant may be interpreted as /w/; if the speaker is talking slowly, the transitions of a /w/ must be even slower. It follows, then, that a /w/ near the boundary can be turned into a /b/ just by making the syllable shorter to signal a faster speaking rate (Liberman, Delattre, Gerstman, and Cooper, 1956).

Is this relationship learned through extensive exposure to language? Probably not. Using habituation methods, Eimas and Miller (1980) found that 2–4-month-olds also were sensitive to syllable length in categorizing a consonant as [b] or [w] in the same manner as adults. Thus, infants detected a change from a syllable with a transition of 16 ms to a syllable with a transition of 40 ms only when the syllable was short (when those transition times are consistent with a change from [b] to [w]) and not when the syllable was long (two [b]s). Likewise, infants detected a change from a syllable with a 40 ms transition to one with a 64 ms transition only when the syllable was long, and not when it was short. Once again these effects are probably not specific human adaptations for language (for example, birds have shown the same effect with human syllables; Welch, Sawusch, and Dent, 2009), but they point to the fact that the perceptual similarity
p. 270 space ↳ in human infants and adults is peculiar in some of the same ways, an alignment that surely facilitates the intergenerational transfer of linguistic conventions.

Through the 1970s and early 1980s a wave of studies using precisely controlled speech materials followed up on the Eimas et al. (1971) study (and also early experiments by Morse, 1972; and Moffitt, 1971), revealing infants' ability to detect subtle linguistically-relevant phonetic distinctions (e.g., Werker and Lalonde, 1988). Many of the initial studies were essentially infant versions of several of the speech perception experiments performed at Haskins Labs (Liberman, Cooper, Shankweiler, and Studdart-Kennedy, 1967). These had characteristic similarities: very simple (often synthesized) stimuli, precise parametric manipulation of auditory cues, and discrimination or simple categorization as the response measure.

Later studies with infants retained these features, but as time passed, the field's emphasis transitioned from measurement of specific acoustic determinants of phone identification, into evaluation of whether infants could distinguish a broad range of consonants or vowels of various languages (Aslin, Jusczyk, and Pisoni, 1998). The main conclusion was simple: within the earliest months of life, infants can distinguish the speech sounds of any language, if those speech sounds are clearly realized. Researchers have continued in this line of work, though present-day researchers tend to use more varied speech samples. The phenomenon of early speech-sound discrimination is robust enough that exceptions are newsworthy (e.g.,

Narayan, Werker, and Beddor, 2010) and spark empirical efforts to rescue the broader generalization (e.g., Sundara, Ngon, Skoruppa et al., 2018).

In addition to discriminating many sounds, young infants readily group some speech sounds together despite substantial acoustic variation. For example, Marean, Werner, and Kuhl (1992) found that 2- and 3-month-olds trained to respond to a change from [a] to [i] from a synthesized male voice generalized this response to a synthesized female voice, suggesting an early-emerging ability to track criterial features of vowels over variation in other acoustic properties. The generality of this result is not clear; subsequent studies examining generalization of a familiarized syllable or a word from one talker to instances from another talker, one emotional state or another, and so on, has revealed a mixed picture in somewhat older infants (e.g., Bergelson and Swingley, 2018; Singh, 2008). At present we are not in a position to quantitatively characterize young infants' naive similarity space well enough to say whether infants have a general predisposition to weigh especially heavily formant values or other features that serve to differentiate speech sounds crosslinguistically.

Of course, over time infants move beyond this initial state. After all, they are learning the language or languages of their environment, and languages vary in the categories they use and where the category boundaries fall. One might imagine that infants would simply get better and better at resolving their native-language categories, while retaining their initial capacity for sound discrimination, but this is not what happens: as infants develop, improvements in native-language speech-sound categorization are accompanied by decrements in categorization of nonnative speech sounds.

p. 271 Two foundational experiments demonstrating these early changes are Werker and Tees (1984) and Kuhl, Williams, Lacerda, Stevens, and Lindblom (1992). Both used a conditioned headturn method in which infants were rewarded with an audiovisual display that was available only when a soundtrack of repeated syllables changed to another, different syllable. Because infants had to turn to see the reward, headturns indexed infants' detection of the change. Werker and Tees (1984) found that English-learning 6–8-month-olds nearly all succeeded in learning this contingency for a pair of unfamiliar stop consonants (drawn either from the Salish language Nthlakampx, or from Hindi); by 10–12 months, infants only rarely succeeded, despite performing well with English consonants. Thus, unless sounds are present in the infant's language environment, they may become difficult to distinguish.

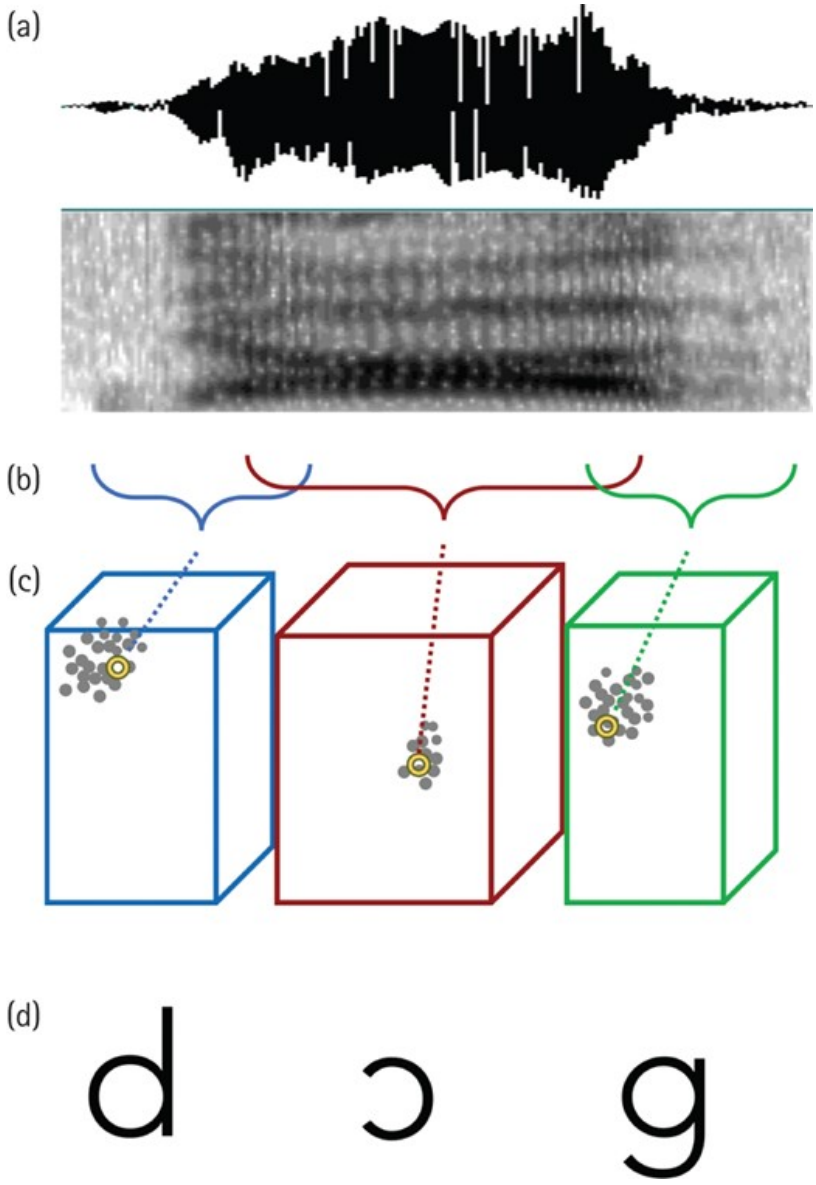
Kuhl et al. (1992) examined what is, in a way, a mirror image of this effect, showing that by 6 months of age, variant instances of a given native-language vowel category can become more difficult to differentiate. Compared to Swedish infants, English-learning infants were not as good at discriminating altered versions of English [i] from one another; but English infants were better than Swedish infants at telling variants of Swedish [y] apart. This might seem paradoxical: why would experience with language make infants worse at noticing variation within a familiar vowel category? (It would be surprising, for example, if dog experts were inferior to cat experts in noting an uncharacteristic bark in a terrier.) Nonetheless, Kuhl et al.'s result makes sense functionally: if a word has an /i/ in it, it does not matter exactly what shade of /i/ it is, and so English learners should eventually learn to collapse irrelevant differences within the category. Computational models that make intuitively sensible assumptions about the learning process have suggested ways this result could come about in infants even though they are not being explicitly trained to differentiate vowels (e.g., Guenther and Gjaja, 1996). Similar acquired-equivalence effects have been found in other domains, often paired with enhanced discriminability effects at category boundaries (Goldstone, 1998; for a review and a model, see Feldman, Griffiths, and Morgan, 2009). In the domain of speech, the phenomenon of reduced discriminability near the center of a category relative to the periphery became known as the "perceptual magnet effect," based on the metaphor of the category prototype being a magnet attracting other things toward it (Kuhl, 1991).

Further experiments testing infants' discrimination of native and nonnative speech-sound contrasts supported the patterns revealed in these initial studies: in general, infants whose native language uses a phonetically "close" pair of sounds will be able to distinguish clear instances of those sounds throughout development; infants whose native language does not use this pair of sounds will begin to show reduced discrimination starting between 6 and 12 months of age (reviews, e.g., Jusczyk, 1997; Kuhl, Conboy, Coffey-Corina et al., 2008; Werker, 2018). This decline in performance with nonnative sounds appears to take place somewhat sooner for vowels than for consonants, with evidence of nonnative discrimination failures at 6–8 months old in vowels (Bosch and Sebastián-Gallés, 2003; Polka and Werker, 1994) and analogous declines in consonants about three months later (e.g., Rivera-Gaxiola, Silva-Pereyra, and Kuhl, 2005; Segal, Hijli-
p. 272 Assi, and Kishon-Rabin, 2016; Tsao, Liu, and Kuhl, 2006). Note, though, that this ↳ observation, while frequently cited, is not yet conclusively supported (Tsuji and Cristia, 2014), and this pattern of results would not imply that vowels are easier to learn; it might instead imply that consonants take longer to unlearn.

Infants' adaptation to their native language is also characterized by quantitative *improvements* in categorization performance for familiar speech sounds. For example, Kuhl, Stevens, Hayashi et al. (2006) tested English learners' responses to [ɹa] and [la] using a headturn method and found that 6–8-month-olds averaged 64% correct, whereas 10–12-month-olds averaged 74% correct. While this kind of improvement is perhaps not surprising, it does exclude any theory holding that perceptual development is just a matter of weeding out pre-existing category boundaries that are not relevant in the local language. The priority and eminence of the Eimas et al. (1971) study and the common characterization of young infants as "universal learners" or "universal perceivers" might give the impression that infants are born with a highly reticulated phonetic perceptual space, and that development could consist of collapsing most of these distinctions to eventually arrive at the mature native speaker's state. Such a view is not really credible: mature speakers of different languages or dialects implement the "same" sounds in measurably different ways, which would imply a greater number of innate boundaries than there are sounds in all languages, and would still require a procedure for selecting among them. Indeed, innate facilitation of the voicing boundary in English stop consonants is probably more the exception than the rule. In the case of vowels, for example, categorical perception effects are weak. The vowel space is continuous, and languages impose vowel categories upon it (e.g., Kronrod, Coppess, and Feldman, 2016; Swoboda, Morse, and Leavitt, 1976; but see Kuhl, 1994 for a different view). Uncontroversially, these categories must be detected via a data-driven learning process.

One way to characterize the developmental transition we have described so far is the illustration in Figure 13.1. The newborn infant hears continuous speech (a), breaks it down into a sequence of consonants and vowels (b), and projects each token into a multidimensional, speech-specific similarity space (c). By 12 months of age, this phonetic space is divided into discrete categories, or reference points, to which experienced speech sounds are matched as they are heard. There is some debate about how closely these phonetic categories line up with the phonemes of the language and how readily they should be expected to serve the phoneme's role as defining lexical contrast (Swingley, 2016; Werker and Curtin, 2005). If the categories can work as phonemes, the infant would have achieved a significant linguistic goal: conversion of the continuous acoustic signal into a set of discrete representations suitable for concatenation into distinct, identifiable words (d).

Figure 13.1



Complete analog to digital conversion in infants: tempting, but unsupported.

This illustration is both incomplete and misleading in some important respects. It is incomplete in ignoring suprasegmental features of speech, for example, although infants certainly encode and learn about accent, tone, prosodic phrasing, and so forth, and these features interact with phonetic categorization in complex ways. Worse, though, the conception of the initial state shown in Figure 13.1 is that infants can easily
p. 273 identify which portions of the speech signal correspond to phones to be analyzed and learned. Figure 13.1 ↳ also implies that by 12 months of age, infants correctly and exhaustively account for the speech signal in terms of phonological units ready for service in representing and differentiating words. These assumptions are debatable, as we will see.

Several studies have attempted to determine whether infants represent speech in terms of sequences of consonants and vowels. One approach has been to use number: if infants interpret speech in terms of phones, they might detect when a sequence of four-segment words (like *rifu, iblo* ...), changes to a list of six-segment words (*suldri, kafest* ...). But 4-day-olds do not respond to this change, though they do respond to changes from two-syllable words to three-syllable words (Bijeljac-Babic, Bertoncini, and Mehler, 1993). Another approach has been to present infants with sequences of syllables that exemplify some regularity,

like rhyming, and see if they prefer such a list over one without any such regularity. Experiments of this sort have yielded mixed results. Jusczyk, Goodman, and Bauman (1999) found that 9-month-olds preferred lists of consonant-vowel-consonant (CVC) syllables matching in onset consonant and in onset consonant and vowel, but not in the rhyme (VC). Hayes and Slater (2008) found that 3-month-olds preferred onset-matching syllables over more miscellaneous syllables. Infants can be *trained* to detect changes to a series of rhyming syllables even if the change is only to alter the vowel or the final consonant (Hayes, Slater, and Brown, 2000; Hayes, Slater, and Longmore, 2009), but this does not necessarily require that infants interpret these CVC syllables as comprising three parts. Because these experiments each used quite various sets of syllables in setting up the tested regularities, infants' recognition of the patterns may implicate a similarity comparison that did not require generalization specifically over segmental units.

p. 274 Some more precise attempts at this question have also yielded mixed results. Jusczyk and Derrah (1987), using a sucking method like Eimas et al.'s (1971) found that infants who were habituated to the sequence [bi, bo, bɜ, ba] did not show a greater response to the addition of [du] than to the addition of [bu]. Eimas (1999) obtained concordant results using a looking-preference habituation method. These authors argued that infants represent the information that distinguishes all of these syllables, but, on the basis of parsimony, hypothesized that infants do not also segment syllables into consonants and vowels. The opposite result was found by Hochmann and Papeo (2014) who indexed surprise with a pupillary dilation response, and found evidence that infants recognized the distinctiveness of the added consonant despite the variation in the following vowels, a variation that can be quite subtle relative to the changes induced by context. At this point, this line of research taken together does not present a clear picture.

A number of studies have suggested that infants can learn phonetic generalizations ("phonotactics") based on specific features (like being a fricative, having a labial place of articulation, and so forth), but in many cases these do not necessarily implicate a segment-level representation. For example, an English-learning infant could be put off by the Dutch consonant sequence [kn] (as in *knie*, knee), without representing the two consonants as distinct sounds (Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk, 1993). Other studies that might be harder to explain without implicating segments involve infants of 9 or 10 months old (e.g., Chambers, Onishi, and Fisher, 2011; for a skeptical overview and meta-analysis, Cristia, 2018). In addition, putting aside any questions of interpretation, it is important to recognize that all of these experimental studies make exclusive use of very short utterances delivered in a hyperarticulated register that is only sometimes characteristic of words in infant-directed speech. Still, based on these studies it seems that at least some of the time, infants capture subsyllabic units from speech, and that by 9 months old or so, they have enough of a sense of their language's phonological patterns to treat rare phonetic chunks as unfamiliar.

Finally, the fact that infants learn the speech-sound categories of their language might imply a segmentation of the signal into units the size of those categories. But it might not: we can be familiar with a thing and not realize that its parts are parts, or give them any significance. A holistic understanding is not necessarily vaguer than an analytic one. Hypotheses in this domain are testable in specific cases. If, for example, infants build categories of vowel-nasal sequences that conflate the two sounds (and that are distinct from the category they are building for the vowel on its own in a non-nasal context), they would show stronger habituation or surprise effects across tokens within that vowel-nasal pairing than across tokens with non-nasal codas (in the manner of Hochmann and Papeo, 2014, for example).

Given the evidence summarized above, it does not seem necessary to assume that during the first year, perhaps even in the first half of the first year, infants spontaneously break down the speech signal into consonants and vowels. They might do so, perhaps even most of the time, but this appears to be an open question. There are hybrid alternatives to full segmental decomposition that are plausible, in my view,
p. 275 despite having no detectable currency in the literature. For example, infants might have innate or ↳ very early-developing parsing skills that lead them to draw boundaries at areas of salient phonetic change—

dividing fricatives from everything else, or continuants from stops, and so forth. Perhaps they interpret vowel-glide sequences in much the same way we think of diphthongs: complex sounds with trajectories built in. Even in the case of mature native speakers, borderline cases are not uncommon. English speakers might assert that there is a /w/ in *you wonder* but not in *you under*, but it seems a stretch to assume that infants would share these intuitions in an adult-like way before having learned to do so. These questions about parsing and representation have not been addressed much but may be amenable to investigation with existing techniques (see Martin, Peperkamp, and Dupoux, 2013, for discussion; see also Magnuson and Crinnion, this volume).

13.3 Hypotheses about phonetic category learning

When infants do isolate consonants or vowels, how do they learn the categories to assign them to? We know more about the developmental timing of this learning than we know about the learning process. An early assumption that children learn speech sounds by observing patterns of contrast in the lexicon (Trubetskoy, 1939/1969) became difficult to sustain once the precocious nature of early learning became clear. If knowledge of words like *boat* and *goat* were needed to differentiate /b/ and /ɡ/, then either stop consonants must be learned much later than the infant perception studies showed, or infants must have a large enough vocabulary to support these distinctions by 6–12 months of age. Neither proposal seemed plausible. Lexically driven theories require additional capacities for phonetic learning anyway, because knowing that *boat* isn't *goat* doesn't itself reveal what phonetic features are criterial for the distinction.

Consequently, the dominant explanation for infant phonetic learning abandons the lexicon and relies on distributional learning over experienced speech tokens. Distributional learning here refers to inducing categories by detecting clusters of similar sounds. The premise of this proposal is that for each sound of a language's phonology, spoken instances of that sound will be similar to one another, and separate from members of other categories. In principle, statistical modes in a set of perceptual experiences can be detected without labeled training data (Duda and Hart, 1973). For example, if a sample of vowels with a first formant of about 450 Hz includes half with a second formant below 2,000 and half with a second formant above 2,250, there is a basis for inferring that there are two separate categories in that region of phonetic space.

Kuhl et al.'s (1992) finding of language-dependent prototype structure in learned vowel categories was hypothesized to come about through this sort of unsupervised distributional clustering (Guenther and Gjaja, 1996; Kuhl, 1992; Lacerda, 1995). The hypothesis makes sense: how can one learn a category prototype structure, if not by attending to surface distributions of phonetic features? Laboratory experiments with
p. 276 adults and with ↳ infants have shown that brief but concentrated exposure to instances of well-separated acoustic or phonetic categories can modify listeners' interpretation of similar sounds (e.g., Francis, Kaganovich, and Driscoll-Huber, 2008; Goudbeek, Swingley, and Smits, 2009; Liu and Holt, 2015; Maye, Werker, and Gerken, 2002; Yoshida, Pons, Maye, and Werker, 2010). In one carefully studied case, infants were shown to have learned a native-language distinction better if their parent tended to articulate one of the sounds in a more acoustically distinct way (Cristià, 2011). Thus, this evidence suggests that distributional clustering is a learning mechanism that might explain early phonetic attunement.

Of course, distributional clustering can only succeed in yielding language-specific phonetic categories if the categories are present to be found in the child's linguistic environment. If instances of a given category are not particularly similar to one another, or if categories are close to one another relative to their spread, unsupervised distributional learning cannot work. Some early studies appeared to support the feasibility of distributional learning by describing the difference between infant-directed speech and adult conversation, and showing greater separation among vowel category centers in infant-directed speech. Further study has

produced a mixed picture, with some research suggesting that mothers clarify their speech (e.g., Bernstein Ratner, 1984; Kalashnikova and Burnham, 2018; Kuhl et al., 1997), and others suggesting that they do not (e.g., Bard and Anderson, 1983; Cristià and Seidl, 2014). This discrepancy may mean that increased clarity in parental vowel production varies according to aspects of the context, the sampling methods, the child's age, and various features of the parent.

For present purposes, though, our question is a bit different–not so much how infant-directed speech may be special, but to what degree an infant might learn speech sounds from it. Answering this question requires an estimate of what information infants might extract from each instance of a sound (and which instances count), and a guess about how the infant's mental clustering algorithm works. In general, researchers have started from the assumption that infants extract formant values from all instances of vowels they hear, and cluster them in a way that can be approximated by either statistical clustering models (like *k-means* or hierarchical cluster analyses) or more complex computational models (Vallabha, McClelland, Pons, Werker, and Amano, 2007). Almost all such studies have been unable to show that distributional information in infant-directed speech is adequate for category learning of the sort apparently demonstrated by infants. The vowels overlap too much (Adriaans and Swingley, 2017; Antetomaso et al., 2017; Jones, Meakins, and Muawiyath, 2012; Swingley and Alarcon, 2018). In general, these models do not just fail; they are appallingly bad. For example, Swingley and Alarcon (2018) found that the basic morphology of a clustering solution was so arbitrary, it was usually significantly altered by sampling random sets of 99.5% of the data rather than the full 100%.

The exception to this pattern, presented by Vallabha et al. (2007), did show successful learning of vowel categories. The speech data the model was given were not measurements of vowel tokens, but samples drawn from Gaussian distributions whose parameters were derived from recordings of mothers talking to their infants. The speech was not free conversation, but mothers reading nonce words from a storybook,
p. 277 where ↳ many of the words were phonologically similar to one another (*peckoo, payku, kibboo, keedo …*). Under these conditions is it very likely that mothers produced relatively hyperarticulated speech. This result suggests then that unsupervised learning models do not fail because they have in-principle flaws; they fail because ordinary parent-infant conversation is phonetically messy.

Why do infants succeed where our own efforts fail? The main contending explanations are these: (a) we are using the wrong phonetic characterization; (b) we are measuring the wrong set of cases; (c) we are neglecting helpful regularities at other levels of description.[1] On current evidence, each of these is plausible. Concerning the phonetic characterization, it is widely understood that the usual technique of describing vowels by measuring their first and second formants at the midpoint, and sometimes adding the vowel's raw duration, only offers an approximation of the full information provided in a vowel. There are many other features, such as change in formant structure over time, spectral tilt, creak, and pitch movement, as well as features outside of the auditory domain entirely, such as visual features in the talker's face (Teinonen, Aslin, Alku, and Csibra, 2008).

It is sometimes suggested that measuring these features and adding them to categorization models would make the models more successful. This may be true (though it is difficult to be sure, as these are hard to measure in large natural corpora). It is also likely that categorization models would be more successful if they incorporated more of infants' natural biases in the interpretation of speech (e.g., Eimas and Miller, 1980). A barrier to this implementation is that we do not have a complete accounting of these biases—the infant studies that have been done have been more like demonstration projects or existence proofs than like reference manuals. So perhaps speech perception is easier for infants than it seems, because of the information they have native access to.

It is also possible that our models inflate the difficulty of learning speech sounds because of the assumption that all speech tokens drive the learning process. Infants may attend to some instances more than others

(perhaps because they are unusually salient: louder, sitting on pitch peaks, longer …), and perhaps the ones they attend to exemplify their categories more clearly. If this is the case, the random-looking explosions of points that make up typical speech first-formant x second-formant plots present too pessimistic a picture of the learning problem. For the purpose of recognition, every missed sound or word is an error; for the purpose of learning, relying on just a few very good tokens might be a perfect strategy. Adriaans and Swingley (2017) tested this idea by comparing the separability of vowels that had been independently labeled as being emphasized by the mother, and vowels that had not been so labeled. Certainly, the "focused" vowels were significantly more distinct from one another, and showed indications of hyperarticulation. This being said, the focal vowels still overlapped considerably, suggesting that attending only to prominent vowels does not make the variability problem go away.

p. 278 A third possibility is that infants make use of contextual information outside the segment for helping to identify those segments. As discussed above, one version of this is a very old idea: the notion that children's knowledge of the different meanings of minimal-pair words could inform children about phonological distinctions. The problem with this idea for explaining infant development is that infants were supposed to start learning the meanings of words between 9 and 12 months of age or so, after the age at which experiments showed that they had begun to learn some of their language's speech-sound categories. More recent research has indicated two ways around this sensible developmental objection: first, perhaps infants are already building a meaningful lexicon by midway through the first year, and if so, these words may be numerous and diverse enough to contain minimal or near-minimal pairs that in ensemble could drive phonetic learning. Second, infants might come to represent chunks of speech corresponding approximately to words, and rely on these chunks (the "protolexicon"; Swingley, 2005b) to serve as recognizable islands that could form the basis of phonetic generalizations. I will return to this possibility after characterizing what infants know of words.

13.4 Infants learning words

In the 1980s and through the 1990s, researchers interested in what infants know about language diversified in questions they asked. Does language help infants learn categories? (e.g., Balaban and Waxman, 1997). Do infants relate speech sounds to talking faces? (Kuhl and Meltzoff, 1982.) Can newborns tell one language from another? (Mehler, Jusczyk, Lambertz et al., 1988). Can infants learn abstract rules? (Gomez and Gerken, 1999; Marcus, Vijayan, Rao, and Vishton, 1999). Much of this work was possible because of innovations in the methods used to assess infant cognition in the domain of language. By far, the most influential among these has been the Headturn Preference Procedure.

In one early study, Fernald (1985) seated infants in a three-sided booth containing a loudspeaker on the left and right, and a light on the left, in front of the infant, and on the right. On each trial, one of the lateral lights was illuminated. When infants turned to look at it, speech was played from the speaker on that side. Fernald (1985) found that four-month-olds turned more to a side that played speech in an infant-directed speech register (with the pitch contours, elongations, and positive affect typical of speech directed to infants) than to the side that played speech in an adult-directed conversational register. Later studies adapted the method so that infants' time to orient to a particular trial type became the dependent measure, with side of presentation randomized (Hirsh-Pasek et al., 1987; see also Colombo and Bundy, 1981).

Infants turned out to be willing to look longer to hear some kinds of speech than to hear others. For the most part, studies revealed a preference for listening to speech samples that had more features of the native
p. 279 language, indicating that infants had ↳ learned from their experience. Several studies exploited this to evaluate whether infants would listen longer to lists of words that they might have heard, as opposed to infrequent or invented (nonce) words. Hallé and de Boysson-Bardies (1994) were the first to show just that,

in 11- and 12-month-olds, inspiring a series of follow-up studies testing whether this preference would be maintained if the words were altered phonologically (Hallé and de Boysson-Bardies, 1996, Poltrock and Nazzi, 2015; Swingley, 2005a, Vihman, Nakai, De Paolis, and Hallé, 2004, Vihman and Majorano, 2017). The motivation for testing "mispronunciations" like this is to learn whether infants' knowledge of frequent word-forms should be viewed as vague, retaining only gross acoustic aspects of words; or more detailed, retaining sufficient phonetic information to cause phonological deviations to reduce familiarity.

At 11 months, the familiar-word preference often disappears when the stimulus words are mispronounced: infants might gaze at a blinking light longer to hear 'dog' but not 'tog.' In some studies, infants also reveal a preference for canonically realized words over deviant pronunciations of those words. These changes in response seem most reliable in stressed syllables, and less consistent when less prominent portions of words are altered. This suggests that either infants only accurately represent the more salient parts of words, or less salient changes interfere less with whatever motivates infants to listen longer to familiar words. At five months, the familiar-word preference has been shown using the infant's own name (Mandel, Jusczyk, and Pisoni, 1995). Experiments altering the phonological form of the name have yielded a more complex pattern of results, with infants detecting some changes and not others, in ways that may depend on the language being learned (Bouchon, Floccia, Fux, Adda-Decker, and Nazzi, 2015; Delle Luche, Floccia, Granjon, and Nazzi, 2017). What these studies show is that by 11 months and perhaps earlier, infants, in their daily experience with language, hear some frequent words and store them in memory with a level of fidelity that would be adequate for differentiating similar-sounding words as the language requires. These studies do not provide much of a basis for estimating how many words this amounts to, however; infants might know a few dozen word-forms, or they might know several hundred.

How do infants find these words to begin with? Most utterances that infants hear contain more than one word, so short of assuming each utterance is a word, infants need a way to isolate them. Researchers have tried to characterize infant word-finding primarily using brief training procedures. Infants are first presented with a set of sentences, like a very short story, that contains several instances of a given word, such as 'bike.' Then, a series of test trials evaluates infants' listening times to repetitions of that word ('bike ... bike ... ') or another word ('feet ... '). In most studies, materials are counterbalanced across infants, so some children's familiar target serves as other children's unfamiliarized distracter. If, across the sample, most children listen longer to the familiarized word, it shows that they were able to recognize the match between the word as it appeared in the sentence, and as it appeared in isolation. This, in turn, would require that infants pull that word out of the sentence to begin with, and remember it for a brief period. The procedure can also be done in reverse, starting with isolated words and testing on different passages (Jusczyk and Aslin, 1995).

p. 280 The introduction of this method sparked an explosion of studies about infants' detection of words in continuous speech (for a review, see Johnson, 2016). Some of these studies asked about memory and representational format. For example, Jusczyk and Hohne (1997) exposed 8-month-olds to words read from a storybook over a period of two weeks, and then tested them in the lab after a delay of another two weeks. Infants revealed a preference for common words from the storybook (e.g., *jungle*, *python*) relative to control words matched for syllable number (e.g., *lanterns*, *camel*).

Other home-exposure/delay studies using these methods have yielded some variations: 7.5-month-olds perform better when the familiarization recordings use an infant-directed speaking style, if the familiarization is done from audio recordings played without coordinated social engagement (Schreiner, Altvater-Mackenson, and Mani, 2016). Keren-Portnoy, Vihman, and Fisher (2019) found that 12-month-olds could recognize home-exposed words, but only when those words were uttered as one-word utterances (in exposure and at test). The latter study suggests a less generous picture of infants' word-segmentation ability. The authors propose that previous studies may have highlighted the trained words by presenting them in a voice other than the mother's, making them more memorable at test. There are more variables in

play here than there are datasets, but these conclusions seem safe: first, infants do show durable memory of some of the words that they hear in fairly typical at-home contexts, even when they have little or no information about what the words mean. Second, utterance position and talker identity probably play a role in determining which words children will recall (at least when recall is measured using this preference measure).

The bulk of the headturn preference studies in this domain have been dedicated to characterizing infants' capacities for speech segmentation, identifying which features of the speech signal they interpret as cohesive. Several studies have shown that infants do not group together portions of speech that fall on either side of a prosodic boundary, like a clause boundary; furthermore, words that are aligned with these boundaries are easier for infants to detect (Seidl and Johnson, 2006; see also Hirsh-Pasek et al., 1987). This is true even in infants as young as 6 months (e.g., Johnson, Seidl, and Tyler, 2014; Shukla, White, and Aslin, 2011). Defining exactly what counts as such a boundary to an infant, and then characterizing where these boundaries actually fall in infant-directed speech, would help provide quantitative estimates of how much these boundaries reduce ambiguity about word boundaries in the speech signal.

A second way infants might discover words is by identifying chunks of speech whose component parts appear together in sequence more often than might be expected (which suggests they may exist together as an independent unit), or whose component parts seem to appear together only infrequently (which suggests they may include a boundary). Considering the Friederici et al. (1993) result mentioned above, an English learner hearing /kn/ might well consider there to be a boundary between these two sounds (as in 'walk now'), whereas a Dutch learner might instead consider them an onset (as in 'knoop,' *button*). Experimental tests of this possibility have consistently shown that infants extract words more easily when the contexts favor their segmentation phonotactically. For example, in English the sequence /nɡ/ is rare within words. On
p. 281 ↳ the other hand, /ŋɡ/ and /ft/ are quite common within words ('finger,' 'lefty'). Likewise, /fh/ is rare within words, appearing only in compounds like 'wolfhound.' Consequently, a word like /ɡæf/ should fairly leap out of its context in 'bean gaffe hold,' and melt comfortably into 'fang gaffe tine.' Indeed, 9-month-olds detect 'gaffe' more readily in the 'bean … hold' context, suggesting that they not only have some sense of the phonotactic facts, but also use them to pull words out of their context (e.g., Mattys and Jusczyk, 2001).

Infants also appear to learn language-particular prosodic generalizations that can help with speech segmentation. The best known of these is the "trochaic bias," or the tendency to assume that stressed syllables (or syllables with full vowels: Warner and Cutler, 2017) start words. In Germanic languages like English, bisyllabic words tend to have stress on the first syllable, and mature listeners have a bias against interpreting weak-strong syllable pairs as words (Cutler, 2012). Studies of English-learning infants show that infants in headturn preference experiments more readily recognize strong-weak bisyllables like 'kingdom' than weak-strong bisyllables like 'beret' (e.g., Jusczyk, Houston, and Newsome, 1999). Given that only some languages exhibit this phonological regularity, it is probably learned through experience with the language rather than being an innate default. Infants in other language environments use regularities present in their language (e.g., Nazzi, Mersad, Sundara, and Iakimova, 2014).

How could infants learn consonant-sequencing regularities and typical stress-pattern regularities at the word level, before knowing words? Is there a generic mechanism that could gain infants a foothold regardless of language? Certainly, phonotactic regularities concerning words might be guessed at using utterance edges or other innately available boundaries. If a given consonant cluster comes at the start of an utterance, it's a good bet as a word-initial sequence. Models that learn phonotactic regularities in this way perform well in locating word boundaries in phonologically transcribed corpora in English (e.g., Daland and Pierrehumbert, 2011).

Another potential generic mechanism that could help group together elements of a word is to evaluate whether those elements tend to appear together frequently, where "frequently" could refer to some

absolute definition (is this pair of syllables, AB, more common than most pairs of syllables?) or a more complex conditional definition (when A occurs, does B usually follow, and vice versa?). Laboratory studies using "artificial languages" and, less often, carefully tailored natural-speech recordings, have used familiarization–preference designs to test this possibility. These studies have shown that infants are capable of computing both the absolute and the conditional kinds of frequency (Aslin, Saffran, and Newport, 1998; Goodsitt, Morgan, and Kuhl, 1993; Pelucchi, Hay, and Saffran, 2009; Saffran, Aslin, and Newport, 1996). Although most of this research has tested infants of about 8 months, recent evidence suggests that newborns perform similar frequency computations over simple syllabic stimulus materials (Fló et al., 2019).

These possibilities give us a plausible narrative for word-form discovery. First, infants detect high-frequency elements at readily identifiable, prosodically defined edges, and high-probability transitions
p. 282 from one element to another. This yields a stock of familiar ↳ phonetic forms, or "protolexicon" (Swingley, 2005b). Because the protolexicon is made up of elements that were identified using imperfect heuristics, its fidelity to the language is rather poor at first, containing numerous word clippings and spurious portmanteaus (e.g., Loukatou, Moran, Blasi, Stoll, and Cristia, 2019; see also Saksida, Langus, and Nespor, 2017). But it may be just correct enough to feed discovery and refinement of additional heuristics (such as the trochaic bias in English; Swingley, 2005b; Thiessen and Saffran, 2003), which in turn support additional growth and refinement.

As this growth proceeds, words (or hypothesized proto-words) can aid in finding additional words, through "segmentation by default" (Cutler, 1994), that is, if a word is identified for sure, whatever comes after it is the start of another word (Brent and Cartwright, 1996). Infants use this strategy from a young age, according to Bortfeld, Morgan, Golinkoff, and Rathbun (2005), wherein infants hearing their name (or 'mommy') in a sentence were able to extract the following word, but were otherwise unsuccessful (see also Altvater-Mackensen and Mani, 2013; Shi and LePage, 2008). A difficulty with segmentation by default, of course, is that words are embedded in other words; the child might hear the familiar syllable 'can,' identify it as the word *can*, and assume that 'teloupe' must be its own word. Anecdotes about children protesting that they needn't behave because they are already 'being have' fit this picture. Ultimately, determining how often this heuristic should be useful is a matter for computational models of corpora.

Researchers have crafted a range of computational learning models inspired by infants' performance in laboratory experiments and on *a priori* theoretical considerations. We cannot review them in any detail here, but Loukatou et al. (2019) and Saksida et al. (2017) provide some quantitative comparisons. Such models are critical for evaluating the real-life utility of the capabilities infants reveal in lab demonstrations. The authors of these models concede that they generally make utopian assumptions about the infant's ability to interpret every phone in the input correctly, where "correctly" means "spoken like a dictionary" (see Figure 13.1). This is usually justified by appealing to the infant speech-categorization experiments reviewed above, and by the assumption that infant-directed speech, being hyperarticulated relative to adult conversation, does not suffer the same crushing levels of reduction (Warner, 2019). Another concern about most presentations of the computational models is that they are evaluated on their ability to produce correct segmentations, rather than their ability to mimic infants' performance (that is, a great model that finds all the words is probably a very poor characterization of real human infants). Both of these problems have clear origins: canonical-pronunciation corpora are used because alternative corpora are not yet available, and gold standard evaluation is done because we do not know the actual contents of infants' protolexicons. These are hard problems. We will point out a few partial solutions in Section 13.7 ("Quantitative analysis … "; see also Creel, this volume, and Magnuson and Crinnion, this volume).

What is the developmental role of infants' speech segmentation? It was widely assumed in the 1990s and 2000s that infant word-finding at around 8 months old was a precursor to word learning. Words whose forms were already known might be easier to learn as real words later on, when infants were older and
p. 283 beginning to build a ↳ true, meaningful lexicon (Graf Estes, Evans, Alibali, and Saffran, 2007; Hay,

Pelucchi, Graf Estes, and Saffran, 2011); or words whose forms were familiar might have more robust phonological representations (Swingley, 2007a). And of course, enlarging the protolexicon would yield a larger database of phonological patterns, which could in turn improve the quality of speech segmentation.

The picture that emerges from all we have discussed thus far is that infants are competent learners of phonetic structure at multiple levels. They can learn categories of sounds, they can detect how those sounds tend to co-occur within syllables, and they can learn stretches of sound that in many cases line up with words of the native language. Once they are familiar with these words, they can use them in turn to locate more words in running speech. By 12 months, they are primed and ready for word learning. Of course, implicit in the discussion of these studies is the assumption that infants start learning language by solving the conversion of speech into the correct sequence of consonants and vowels, and then using these segmental units as the elements from which syllables are counted and words are constructed.

13.5 Words and speech sounds

This tale of the statistically prodigious but phonology-driven infant turned out to have a little flaw and a bigger flaw. The little flaw is the one mentioned earlier: unsupervised learning of speech-sound categories solely from distributions of infant-directed speech tokens seems difficult, and might be impossible. The bigger flaw is that word meaning seems to come into the developmental picture much earlier than previously supposed. We will take these up in turn.

Infant speech-segmentation research has shown that infants learn word-sized chunks of speech at around the same time they are learning speech sounds. As we have seen, maternal speech does not seem to support speech-sound categorization by presenting sounds in phonetic clusters. Could forms from the protolexicon somehow aid in the discovery of phonetic categories? Perhaps variability in the environments of speech sounds could help render those sounds distinct from their neighbors. For example, analysis of a speaker in the Buckeye corpus of adult-adult conversation (Pitt, Dilley, Johnson et al., 2007) showed that vowel pairs like [ɛ-æ] do not form a bimodal distribution in the space formed by [duration, first formant, and second formant]. A clustering algorithm would not isolate two categories. However, the [ɛ] sound is vastly more likely to be followed by [n] than [æ] is; indeed, in the sample measured by Swingley (2007b), the presence of [n] as a coda consonant cleanly separated the [ɛ] tokens from the [æ] tokens. Analysis of phonotactic and phonetic distributions reveals several such cases, driven mainly by accidents of lexical distribution and lexical frequency (and sometimes by phonological rules, like the English ban on syllable-final lax vowels). Wherever these distributional differences are statistically strong, they might help indicate to infants a difference in category identity.

p. 284 However, while there is evidence of 9-month-olds learning phonotactic rules, there are more abundant data on somewhat younger infants learning word-forms, which suggests the question: could word-forms themselves point infants to speech-sound categories? Consider the position of an infant learning Spanish and unsure whether /i/ and /ɛ/ are the same or not. The instances of these sounds overlap substantially, although as expected the /i/s tend to exhibit a higher second formant and lower first formant than the /ɛ/s. Based on these data, the evidence to the child that there are two categories is slim at best.

If the child were familiar with some words, such as *quieres* and *mira* (with [i] in the first syllable), and *bueno* and *esta* (with [e]), his or her representation of those words, though it derives from multiple instances, might more closely resemble an average or other central tendency of those instances. It appears that clustering over these averages is more successful than clustering over the tokens that gave rise to them (Swingley and Alarcon, 2018). The proposal, then, is that infants learn words and refine speech sounds at the same time, with their first guesses about word-forms providing an additional source of constraint on

phonetic category boundaries (Swingley, 2009; for a fuller discussion and a computational model, Feldman, Griffiths, Goldwater, and Morgan, 2013).

That infants might use contexts in this way is supported by laboratory studies in which meaningless phonetic contexts help shape infants' categorization of speech sounds. For example, Thiessen (2011) found that 15-month-olds familiarized with repetitions of words like 'dabo' and 'tagu' (distinct contexts for [d] and [t]) were more likely to succeed in a difficult minimal-pair word learning task contrasting 'da' and 'ta,' than children familiarized with 'dabo' and 'tabo.' Feldman, Myers, White, Griffiths, and Morgan (2013) took this a step further, testing much younger children (8-month-olds) on a vowel discrimination task. In a familiarization phase, some children heard the vowels [a] and [ɔ] in distinct phonological contexts (like [ɡuta] and [litɔ]), and others heard these vowels in a minimal-pair context (like [ɡuta] and [ɡɔta]). All infants were then tested on discrimination of [ta] and [tɔ], and on this task, only the infants who had been familiarized to these vowels in phonologically distinct contexts discriminated them.

This idea turns on its head the minimal-pair mechanism of establishing contrast. But in a sense, the ideas are similar. Vowels as instances populate the phonetic space too uniformly to be readily clustered. Words populate the phonetic space in a lumpy, nonrandom way, such that many words of the early lexicon are identifiable (if quite imperfectly) even before the precise phonetic bounds of their constituent units are defined; as a result, words can serve as identifying contexts for their components. On such an account, minimal pairs are predicted to make learning harder at first, because the infant does not have a strong basis for differentiating them before their meanings are known; by hypothesis, if similar contexts bring categories together, minimal pairs would count as identical contexts. But once the members of a pair can be distinguished (by aspects of meaning, or perhaps by cues to syntactic category), minimal pairs should be helpful in guiding children to the right phonological analysis.

Indeed, infants can use semantic evidence to guide their attention to phonetic distinctions, including
p. 285 distinctions that are not used lexically in the native language. ↳ This phenomenon has been demonstrated in a series of studies by Yeung (Yeung and Nazzi, 2014; Yeung, Chen, and Werker, 2014; see also ter Schure, Junge, and Boersma, 2016). For example, Yueng and Werker (2014) familiarized 9-month-olds to two unusual objects, labeling each one with either the word [ɖa] or the word [d̪a] (i.e., contrasting in the nonnative Hindi retroflex and dental /d/, which English-learning 9-month-olds do not discriminate). This consistent sound-to-word pairing, as opposed to no such pairing or an inconsistent one, led infants to differentiate these consonants. What this suggests is a set of interdependent relationships between phonetic categorization, the growth of the protolexicon, and the emerging lexicon (Werker and Curtin, 2005).

13.6 Word meanings

Is there a developmental transition from initial reliance on a protolexicon of phonetic word-forms, into a "true" lexicon of words with semantic content? A key question is when infants begin to link words and meanings. Certainly, infants seem ready to detect connections between utterances and things in the world. When infants are shown pictures of objects and hear a word repeated along with the pictures, hearing the word seems to bind together these objects into a category in the infant's mind; likewise, hearing different words applied to distinct objects seems to set the objects apart. These phenomena were first demonstrated in children 9–12 months old (e.g., Plunkett, Hu, and Cohen, 2008; Waxman and Markow, 1995; Xu, 2002) but have been extended to infants as young as 3–4 months (e.g., Ferry, Hespos, and Waxman, 2010; for a review, Perszyk and Waxman, 2018). These studies would seem to rule out any account in which words are simply carriers of emotional prosody by 3–4 months.

Laboratory training studies have shown that it is possible to teach 6–7-month-old infants the connection between a novel word and a picture or an object (e.g., Gogate and Bahrick, 2001; Shukla et al., 2011).

However, a popular argument holds that the referential uncertainty in children's language environments prevents children from learning word meanings until they have developed sufficient skills of social cognition to understand the intentions behind communicative acts. If such skills are not in place before about 9 months, that age should also mark the onset of word understanding (e.g., Tomasello, 2001; Bloom, 2001). On this line of thinking it is usually assumed that laboratory word learning is either unrealistically simple, or reflects something more pedestrian, like audiovisual association, which is then argued to not qualify as word learning.

Early experimental tests attempting to evaluate infants' knowledge of the meaning of actual words, learned through daily life and not lab training, showed little evidence of word comprehension before 12 months (e.g., Thomas, Campos, Shucard, Ramsay, and Shucard, 1981). Tincoff and Jusczyk (1999) showed that 6-month-olds would look at a video of their mother when hearing the word 'mommy,' and at their father when hearing 'daddy,' but it was not clear how broadly to generalize this result given that those words are probably proper names for infants. Still, even infants not understanding reference ↳ and restricted to a meaning-free protolexicon might detect that certain word-forms appear in particular (proto)lexical contexts, distinct from others, in the same way that Latent Semantic Analysis and similar approaches represent word meaning (Landauer and Dumais, 1997). In principle, this might provide infants a leg up in connecting words to broad semantic categories.

p. 286

Elika Bergelson and I aimed to test this "proto-semantics" idea, and set about a prolonged attempt to develop an anticipatory-eye-movement categorization procedure for this purpose. Being unsuccessful, in the meantime we embarked on a control experiment using a better-established language-guided looking method that tests whether children understand words. Pairs of images were presented on a screen, and parents named one of the images aloud. Infants' gaze was monitored to determine whether they would look more at the named picture. This study was intended to confirm that indeed 6–9-month-olds do not know what common words mean, and to lay out the developmental course of word comprehension over the 6–20-month period (Bergelson and Swingley, 2012). In this, we seem to have failed, because 6–9-month-olds showed evidence of at least partial understanding of several words.

Since that study appeared, a few studies have confirmed that by about 6–7 months of age, infants know at least a little about what some words mean (Bergelson and Swingley, 2015; 2018) and other studies have affirmed this in 9- or 10-month-olds (Nomikou, Rohlfing, Cimiano, and Mandler, 2019; Parise and Csibra, 2012; Syrnyk and Meints, 2017). On current evidence, infants' lexical knowledge is quite sketchy; Bergelson and Aslin (2017) found that 6-month-olds looked at named pictures when the alternative was semantically unrelated (see *car* and *juice*, hear 'car') but not when it was related (see *car* and *stroller*, hear 'car'). Perhaps 6-month-olds think 'car' is a decent word for a stroller, 'bottle' for a spoon, and 'juice' for milk. If so, this raises interesting questions about what the semantic contents of the 6-month-old lexicon. Are words linked to objects within broad situational contexts, rather than to specific object categories? Or are words linked to just a few salient features and are therefore over-inclusive? (Or do infants actually have more precise semantic representations, but smaller semantic mismatches drive their fixations less efficiently?) These questions merit further study.

What does this mean for the notion of the protolexicon? It is difficult to say what proportion of spoken language is comprehensible, even minimally, to infants halfway through their first year. Intuitively, it seems that infants probably remember many word-forms as familiar sequences of speech without knowing their referent. Swingley (2007a) used the Brent and Siskind corpus (2001) to estimate how many words a child might hear with high frequency in a period of three weeks and suggested that in this period a child would hear almost 1,000 words 50 times or more. This count is probably an overestimate, because it extrapolates to the whole day data from sessions when parents knew they were being recorded (Bergelson, Amatuni, Dailey, Koorathota, and Tor, 2019). Based on those more recent results, we might estimate that if an infant needs to hear a word 50 times to enter it into their protolexicon, they could reach 1,000 word-

forms in two or three months. This would likely exceed the stock of words to which they attach some semantic content.

p. 287 This speculative line of reasoning suggests that there is a period on early development in which the infant lexicon contains several words with detailed referential content (*mommy* vs. *daddy*, *hand* vs. *foot*; Tincoff and Jusczyk, 2012), dozens or perhaps a hundred words with some broad semantic (and possibly syntactic) features, and several hundred that would be recognized as familiar but that are not yet meaningful. If this is anywhere near the truth, it suggests a lexicon that mixes both meaningful words, and also primarily phonetic entries akin to those of the previously hypothesized protolexicon, with experience filling in increasing semantic detail over time. How that works is, of course, a large topic on its own (see Gleitman and Trueswell, this volume).

13.7 Quantitative analysis and the poverty of the stimulus

Learning is turning experience into knowledge. In their first year, infants learn a lot about their native language: they learn the basics of how it sounds, and they begin to build their vocabulary. As reviewed above, our strengths in characterizing this learning lie primarily in evaluating the hidden knowledge infants possess. In infants' daily lives, their mental categorization of speech sounds is not visible by parents; their recognition of a word as familiar causes no consistent behavioral response. Part of the excitement of research on early language development has come from revealing this hidden knowledge. This work has given us a developmental timeline. To take the two major developments we have focused on here, infants learn to categorize clear instances of their language's speech sounds (and presumably get better at categorizing more atypical instances too), and infants come to understand something about their first words, at around 6 months of age (and quickly make considerable progress from this shaky start). Along with a developmental timeline, his work has also spoken to individual differences to some degree. When we can measure variability in performance, this variability often turns out to be correlated with later measures of linguistic performance (e.g., Kidd, Junge, Spokes, Morrison, and Cutler, 2018; Kuhl, Conboy, Padden, Nelson, and Pruitt, 2005).

Still, in studying infant language development we are better at measuring knowledge than at measuring learning. This is not unusual in developmental psychology (e.g., Siegler, 2000). Our attempts at measuring learning, as opposed to knowledge, tend to take the form of training experiments in which some phonetic item, word, or pattern is presented over a short period (measured in seconds or minutes) with maximum density (most or all items being relevant to the pattern). The pattern of successes and failures is then informative about the capacities of infants in the tested age group; and patterns of correlation with real-world outcomes (like vocabulary size counts) are informative about the skills that bear on success.

p. 288 All the same, it remains difficult to make *quantitative* predictions about how the variations under study bear on the course of a child's progress toward mastery of his or her native language. To take a common example, many studies have placed infants in a learning situation where the materials are delivered either in stereotypically "infant-directed" speech, or in an "adult-directed" register. Typically, infants perform better with the infant-directed register. We conclude: something about the infant-directed register is working. What we cannot say from this is how much of a benefit it provides. Effect sizes in laboratory measures are not effect sizes outside the lab, because interventions in the lab are often not similar to real-world experience. (There are exceptions—for example, studies with natural, normal-density exposure, such as Kuhl, Tsao, and Liu, 2003.)

A consequence of this is that we have theoretical frameworks, rather than models designed to make quantitative predictions. In many cases, it is difficult to place competing frameworks on a footing that allows for direct comparison. Often, frameworks differ more in their *domain* of prediction than on the

outcomes they predict. This does not mean that frameworks are not useful; they are. To take two examples from our field's most accomplished scholars: The PRIMIR framework (Werker and Curtin, 2005) encourages consideration of the difference between phonetics and phonology, and exhorts us to be aware of the influence of task demands. These are critical reminders, and considering them helps clear up some puzzles in the literature. The Native Language Magnet: Expanded framework (Kuhl et al., 2008) proposes "native-language neural commitment" as an explanatory mechanism and encourages consideration of the neurological underpinnings of learning, knowledge, and on-line processing. These notions allow us to place a wide range of results in a common space for evaluation and consideration of a broad developmental picture. For many verbal models of this sort that conceptually integrate information from a many datasets, asking for quantitative predictions seems downright unfair, and asking which of two frameworks is correct feels like a category error.

Ultimately, we would like to make quantitative predictions, and understand learning at a finer grain. Doing this will require that we characterize the experience of the child. To repeat the slogan given earlier, learning is turning experience into knowledge. But what is that experience? Often, the lack of adequately annotated datasets means that we are reading "experience" off the characterizations of language given in grammars, phonetics-lab recording studies, and idealized corpora. A problem with this is that children's actual environments may present poverty-of-the-stimulus problems that we are not aware of, until we look (e.g., Bion, Miyazawa, Kikuchi, and Mazuka, 2013; Swingley, 2019). In many cases poverty-of-the-stimulus problems are quantitative: infants might have an in-principle "sensitivity" to feature *x*, but is this sensitivity good enough under day-to-day conditions? Are the conditions good enough for even perfect sensitivity to win the day? (According to the two studies just cited, even perfect measurement of vowel duration would not suffice for characterizing the phonological implications of vowel duration in Japanese, English, or Dutch.)

p. 289 One way to achieve a better quantitative understanding of early language development is to measure connections between the language environment and linguistic outcomes more precisely. To take one example, Swingley and Humphrey (2017) examined which features of words make them most likely to be learned. We started from the Brent and Siskind corpus of child-directed speech, and parent report checklists of vocabulary among those same children. Following Brent and Siskind (2001), we used regression analysis to evaluate which aspects of words in the corpus made them most likely to be reported as understood or said by the children. For each word on the CDI (Fenson et al., 1994), and for each child's corpus, we computed its frequency, its frequency in one-word utterances, its frequency sentence-finally, how much parents tended to speak that word with exaggerated duration, and a few other predictors such as the word's form class and concreteness. Among the results were two key findings: first, that overall frequency was by far the strongest predictor of whether a given child would be reported to know a given word; second, frequency in one-word utterances was also a predictor (just as Brent and Siskind claimed), and was therefore not merely a proxy for other measured variables (such as elongated duration) that correlate with appearance in isolation. The point here is not so much the results, but the method: using regression, it is possible to evaluate the relative importance of several aspects of the language environment on the learning of specific items. This kind of study can provide an important counterpart to laboratory word-teaching studies that generally can evaluate only one or two variables at a time, and in an unusual corner of the frequency-of-exposure x density-of-exposure space.

Similar studies are helping to differentiate theories of the infant's capacity for word segmentation. Larsen, Cristia, and Dupoux (2017) implemented several word-segmentation algorithms that have been proposed in the literature, using the Brent and Siskind (2001) corpus as the environment, and parental report data from the WordBank repository (2017) as the outcome measure. Larsen et al. found that the models with the "best" performance in extracting words (i.e., a gold standard determined by the language) were not the

models that most accurately predicted children's word knowledge. Results like this signal the importance of using child data rather than gold-standard perfection in evaluating models of learning.

For the near future the largest hurdle in making quantitative models of infant phonetic and word-form learning is the difficulty of simulating the infant's innate similarity space for speech (Dupoux, 2018), and the lack of annotated corpora of infant-directed speech. Ideally, a computational learning model should take as its input the speech signal itself, or rather this signal represented as the transformation effected on the signal by the neonatal speech perception system. This is a difficult and unsolved engineering problem (e.g., Jansen et al., 2013). In principle, a reasonably accurate model of infant speech perception, supported by the many experiments that have been done to date (and, undoubtedly, by others, designed to fill in the most important gaps in our knowledge), would help us to evaluate how much of the developmental course of early language learning can be attributed to the infant's processing of the information in the speech signal itself, and how much to other sources of information. Achieving this will require substantial effort in
p. 290 corpus collection and annotation, and in developing ↳ speech-engineering tools. Ideally, this should be done in parallel with quantitative characterization of infants' visual environments (e.g., Smith, Jayaraman, Clerkin, and Yu, 2018), since what infants see and hear obviously both contribute to word learning, and each may reinforce phonetic learning as well, via the lexicon.

13.8 Conclusions

When infants experience people, they experience people talking. Parents, siblings, family friends, strangers, and even some sounding objects engage in this familiar, intricate, emotion-laden activity, which is as much a part of the infant's early environment as smiles and milk. Infants are born recognizing the sound of speech, and very quickly, they capitalize on innate auditory abilities to characterize many aspects of the speech of their native community. In reviewing past research in this topic I have focused primarily on the questions and characterizations that have driven this work. The common narrative in which infants begin this process by clipping speech sounds into segments, statistically clustering those segments into categories according to the consonant and vowel inventories of their language, and then using those categories to build the vocabulary, appears to be too simple. The clipping is imperfect, the clustering may well depend on the precursors to the lexicon, and the early vocabulary may not be represented in terms of these discrete units, at least in part of the first year.

What will this narrative look like ten years from now? New approaches may well lead to different emphases and somewhat different questions. For example, infants probably discover their first word-forms when they find that one chunk of speech is especially similar to another chunk heard recently (per, e.g., Park and Glass, 2008). Are consonants and vowels implicated in this similarity comparison at all, as they are in current infant models? If they are not implicated in newborns, then when are they, and why? Is it related to the infant's vocal production? Word recognition is sometimes conceptualized as involving the activation of segmental categories, because language cannot be adequately explained without them (e.g., Pierrehumbert, 2016; Pisoni and Luce, 1987). But these categories cannot be recognized or learned without taking their context into account, because the specifics of their realization depend on many aspects of the context. Indeed, speakers often realize one sound by embedding a gesture toward it within another sound (Hawkins, 2010). Eventually, children must be able to interpret language not by slicing speech into segments and categorizing them, but by solving an *attribution* problem: for each perceivable aspect of the phonetic signal, what is its linguistic origin? Viewing speech perception as a sequence of categorization problems rather than an attribution or "blame assignment" problem may be a mistake (see Quam and Swingley, 2010, for discussion). Finally, evidence is building that infants' own articulations help organize their interpretation of others' speech (e.g., Vihman, 2017). We may find that the course of normal development depends critically on infants' own somatosensory representations (e.g., Beckman and Edwards, 2000; Choi,

p. 291 Bruderer, and Werker, 2019). ↳ If so, the corpora that we will need to create in order to model development adequately may come to involve not only microphones for parent and child, and head-mounted eyetrackers all around, but a scheme for measuring the infant's own articulatory movements. Answering these questions is will be a challenge, but the progress that research has made in the past several years suggests that quantitative explanations of the course of language development are reasonable goals to aim for.

Acknowledgment

This work was supported by NIH grant R01-HD049681 to D. Swingley.

Notes

1 There is also (d), we have mischaracterized the learning that infants show in our categorization experiments. This is worth taking seriously (Schatz, Feldman, Goldwater, Cao, and Dupoux, 2019).