Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants

Mohinish Shukla[a,1], Katherine S. White[b], and Richard N. Aslin[c]

[a]Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, NY 14627; [b]Department of Psychology, University of Waterloo, Waterloo, ON, Canada N2L 3G1; and [c]Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, NY 14627

Edited by E. Anne Cutler, Max Planck Institute for Psycholinguistics, Heilig Landstichting, The Netherlands, and approved February 23, 2011 (received for review November 23, 2010)

Human infants are predisposed to rapidly acquire their native language. The nature of these predispositions is poorly understood, but is crucial to our understanding of how infants unpack their speech input to recover the fundamental word-like units, assign them referential roles, and acquire the rules that govern their organization. Previous researchers have demonstrated the role of general distributional computations in prelinguistic infants' parsing of continuous speech. We extend these findings to more naturalistic conditions, and find that 6-mo-old infants can simultaneously segment a nonce auditory word form from prosodically organized continuous speech and associate it to a visual referent. Crucially, however, this mapping occurs only when the word form is aligned with a prosodic phrase boundary. Our findings suggest that infants are predisposed very early in life to hypothesize that words are aligned with prosodic phrase boundaries, thus facilitating the word learning process. Further, and somewhat paradoxically, we observed successful learning in a more complex context than previously studied, suggesting that learning is enhanced when the language input is well matched to the learner's expectations.

statistical learning | language acquisition | lexical development | intonational phrase

Acquiring a language includes learning mappings from sounds (or signs) to meanings. However, words—the principle units of meaning—are not given directly in the input, but are embedded in a speech signal whose structure is governed by grammatical processes operating at multiple levels. One of the primary steps in language acquisition, therefore, is to discover the sound sequences that define words. However, as any adult confronted with a foreign language can attest, it is hard to perceive unfamiliar speech as sequences of words. Additionally, the language learner must also discover what the words refer to, a particularly tricky problem given the innumerable possible referential features in the world (1, 2). Nevertheless, by 6 mo of age, infants have spontaneously extracted and begun to understand their first words, including highly frequent items such as "no", "Mommy", and the child's own name (3).

Here we provide evidence that 6-mo-olds can rapidly extract a statistically defined, novel auditory word form from running speech and simultaneously map it onto a visual referent in an array of objects. Moreover, we find this dual process of word segmentation and referent mapping only when the statistically defined words are aligned with phrasal prosodic constituents, a universal structural property of natural languages. These findings build on three key results from past research: (*i*) 7- to 8-mo-old infants can extract statistically defined syllable sequences from fluent speech as candidate auditory word forms (4, 5), (*ii*) by 14 mo, infants can reliably map isolated auditory word forms onto visual referents (6–8), and (*iii*) by 17 mo, toddlers can extract auditory word forms on the basis of syllable statistics and subsequently map them onto candidate visual referents (9). We demonstrate all of these behaviors simultaneously in infants as young as 6 mo of age. Further, we find that prosody plays a central role in these processes.

Background

Infants' early learning capabilities must be sufficiently general to handle the variation manifest in the languages of the world. Previous work on word segmentation in infants has thus focused on general distributional strategies. Given that infants are sensitive to the syllable as a unit of speech (10), computing statistical relations between syllables has been proposed to be a general mechanism that can parse sequences, generating as potential candidate words those syllable sequences that have high statistical coherence (4, 11). This strategy has the advantage of not requiring specific linguistic knowledge, instead relying on general cognitive mechanisms that track distributions over linguistic and nonlinguistic stimuli (4, 11, 12).

However, speech is more than a mere string of words; it is an organization of prosodic units, ranging from syllables to entire utterances (13). In prosodic theory, constituents at a higher level of the hierarchy are made up of units at lower levels. Of critical importance for the task of word segmentation, the larger, phrasal prosodic constituents are composed of word-like units, such that the onsets and offsets of phrasal constituents are also the beginnings and ends of words. Critically, as early as 2 mo of age, infants are sensitive to prosodic phrases (14, 15), which are marked by acoustic cues like pitch lowering and durational increases (13, 16). In light of the alignment of words with the edges of prosodic phrases, attention to acoustic cues signaling these phrases could serve as a potentially powerful aid to segmentation (ref. 17, for a computational model of this in adult speech). Indeed, previous studies have shown that even 10-mo-old infants can use prosodic information to constrain their search for known word-form candidates (18). However, to date, studies exploring the potential interaction between distributional and prosodic cues in speech segmentation have been limited to lexical (i.e., word-internal) prosody, where, perhaps unsurprisingly, infants rely on distributional cues at an earlier stage than language-specific, word-stress cues (19).

Further, although much work has demonstrated infants' use of general statistical learning mechanisms in the service of speech segmentation, the linguistic status of the sequences extracted using such mechanisms is not entirely clear. Whereas such sequences are preferred (compared with sequences of lower statistical coherence) in subsequent sentential contexts (ref. 5, at 8 mo) and for subsequent mapping onto referents (ref. 9, at 17 mo), the possibility remains that the extracted sequences have no initial linguistic status and are preferred in subsequent tasks due to their enhanced (statistical) salience or familiarity. In addition, previous evidence for the mapping of even single novel words onto referents before the age of 12 mo has been mixed (6, 8, 20). Studies that do show that such young infants are capable of learning the referents of words typically present to-be-learned words as well segmented, isolated

Author contributions: M.S., K.S.W., and R.N.A. designed research; M.S. and K.S.W. performed research; M.S. analyzed data; and M.S., K.S.W., and R.N.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: mohinish.s@gmail.com.

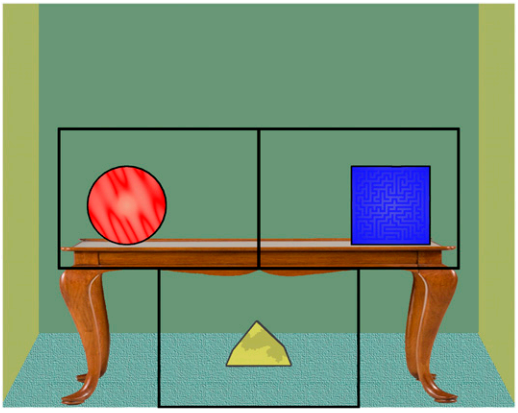

Fig. 1. Screen shot from the experiment, showing the setup of the objects. The three observation windows are overlaid as black outline rectangles.

tokens, embedded in a rich social context (8). However, isolated words form a small proportion of the input to infants (21, 22).

In the present study, we present 6-mo-old infants with prosodically and statistically organized sequences in a simultaneous segmentation and mapping task. Thus, we come closer to assessing the linguistic status of the segmented auditory word forms on the basis of how infants assign these word forms to plausible referents. It may seem counterintuitive that such young infants, who have trouble mapping a single sound sequence onto a single visual object, would find the task easier if they were required to simultaneously segment a statistically coherent word form from fluent speech and map this word form onto one of several visual objects. However, this "simpler is easier" intuition may be misguided. If there are innate constraints on learning, then linguistic input that best matches these implicit expectations may be more easily processed (23, 24). That is, if infants are highly attentive to prosodic cues in the input and expect speech input to be prosodically organized, then a segmentation and mapping task may be easier when such expectations are met. Evidence for such innate constraints is quite strong. In the first month of life, infants prefer intonated speech to syllable lists, show a right ear and left-hemisphere brain advantage for processing speech-like stimuli, and can discriminate nonnative languages across, but not within, prosodically defined linguistic families (25–28).

Given that words typically do not occur in isolation, and that infants prefer intonated speech, we tested the prediction that these circumstances would be maximally effective for extracting auditory word forms from continuous speech and mapping these onto visual referents. That is, word forms embedded in intonated, fluent speech might, somewhat paradoxically, be easier to apprehend than word forms presented in isolation, as in previous studies (20).

Current Research

We asked if 6-mo-olds could segment an auditory word form (a nonce word) from short "utterances" and simultaneously associate it with one of several objects. (An utterance—a string of one or more words bounded by silence—is the highest level of the prosodic hierarchy.) To control the properties of the visual display, infants saw short video animations. The display consisted of a target object (a red circle) and two distracter objects (a blue square and a yellow triangle, Fig. 1), arrayed around a table. In each of nine training trials, infants heard two utterances, as the target object moved along the table (Fig. 2). Each utterance consisted of a pentasyllabic string of the form $x\boldsymbol{AB}yz$, where $\boldsymbol{AB}$ is the target nonce word, whereas the syllables x, y, and z vary. The target nonce word was, statistically speaking, the best possible word-like candidate: (*i*) The transitional probability from $\boldsymbol{A}$ to $\boldsymbol{B}$ was 1.0, and all other transitional probabilities between syllables were <1.0, and (*ii*) the conditional probability of $\boldsymbol{AB}$ given the visual scene was 1.0, thereby creating a perfect association between the statistically coherent word form and the visual target object. [The transitional probability (TP) from an element x to an element y is given by frequency (xy)/frequency (x); it reflects the normalized probability that x is followed by y. Previous work has established that infants are sensitive to TPs in fluent speech (e.g., ref. 4). In this study, the word is not just the sequence with the highest internal TP, but is also the most frequent bisyllable and is the bisyllable with the highest mutual information. For the present purposes, we treat these various statistical metrics as equivalent.]

Each utterance was recorded as two intonational phrases without an intervening pause (Fig. 3 and *Materials and Methods*). Typically, an intonational phrase (IP) is "… the domain of a perceptually coherent intonational contour, or tune" (ref. 29, p. 210). From Fig. 3 *A* and *B* it can be seen that the two IPs correspond to two intonational (pitch) contours and that there is no pause between the two IPs. Critically, for one group of infants (n = 12) the nonce word [$\boldsymbol{AB}$] was aligned with the boundary between the phrases ([$x\boldsymbol{AB}$] [yz]), whereas for the other group (n = 12) it straddled the boundary ([$x\boldsymbol{A}$] [$\boldsymbol{B}yz$]), as can be seen from Fig. 3 *A* and *B*.

The short training phase was immediately followed by a test phase that was identical for all infants. In each test trial (Fig. 4), infants saw all three objects looming on the screen. Once the infant looked at the screen, we collected baseline proportion of

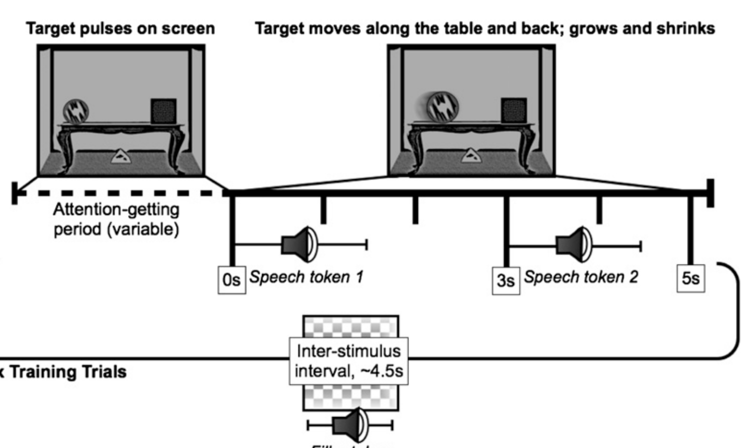

Fig. 2. Timeline of the training trials. The target object is a patterned, red, circular object (Fig. 1). The duration of the auditory stimuli is shown with respect to the timeline.

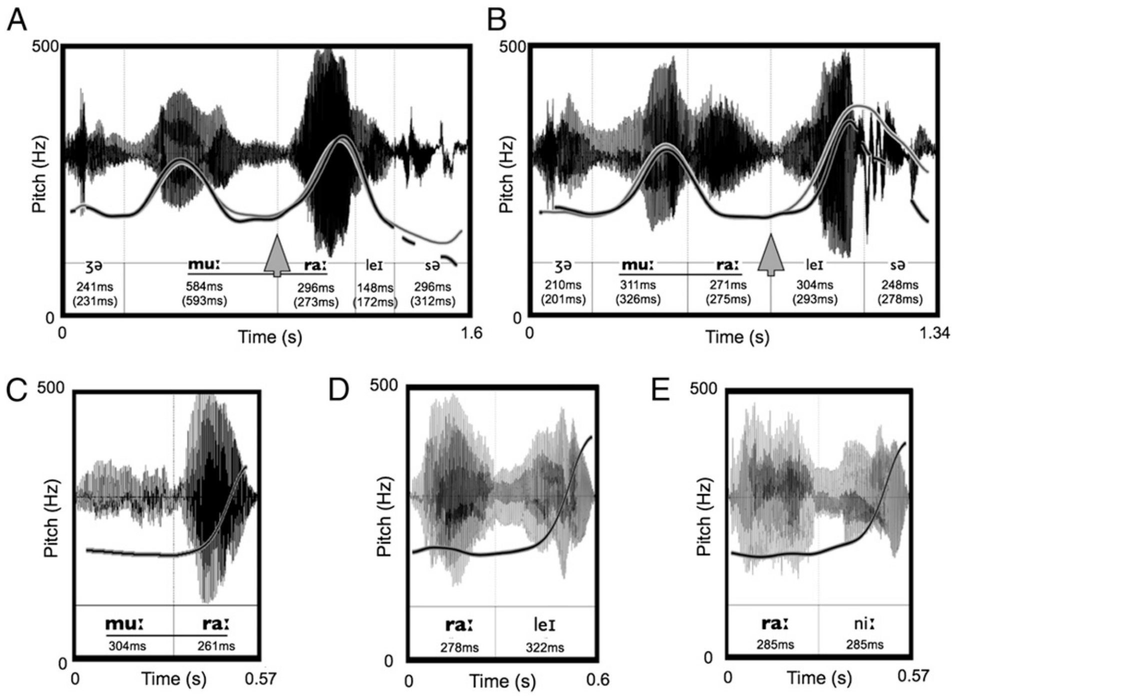

Fig. 3. (*A* and *B*) Speech waveforms corresponding to single tokens of training utterances from the (*A*) IP-straddling word condition and (*B*) IP-aligned word condition. Pitch and duration characteristics are also given. Arrowheads mark the location of IP boundaries. The pitch track with dark shading corresponds to the displayed waveform, whereas the pitch track with light shading is the mean across all four tokens in that condition. Similarly, durational measures for the syllables in *A* and *B* are given for the displayed token, with means across all tokens in parentheses (Table 1). (*C–E*) Speech waveforms, pitch tracks, and syllable durations for the three single tokens of the bisyllabic test items (the high-TP word and two bisyllabic part words). The nonce "word" (*C*) is perceived by adults as being more similar to its occurrence in the IP-straddling word condition (*A*) than in the IP-internal word condition (*B*) (*Materials and Methods*).

looking to each of the three objects. Then, we played two tokens of either the [***AB***] word or one of the two [***B**y*] part-word foils and measured changes in looking over baseline (*Materials and Methods* and Fig. 4). These test tokens were well-formed utterances, prosodically distinct from the training tokens (see Fig. 3 *C–E*).

Results and Discussion

During the 1.5-s baseline period in test trials, overall proportion of looking to the target (median 0.69) and to the distracters (median 0.12) was nonnormally distributed, and infants spent significantly greater time looking at the target compared with the distracters (Wilcoxon signed-rank test, $Z = 8.36$, $P < 0.0001$). Proportion of looking to the target window alone during baseline showed no effect of prosodic condition or the test items. Thus, during baseline, infants showed the same looking pattern regardless of test stimulus or familiarization group.

The primary dependent measure of interest was the difference in proportion of looking between the remaining 2 s of each trial (the critical period, Fig. 4) and the baseline period. For this measure, we found significant effects for the nonce word but not the part words, indicating that infants extracted the frequent, high-TP (statistically coherent) unit from the speech streams (*Data Analysis* and Fig. 5). However, the pattern of looking differed between the two groups of infants. For the change-in-looking measure on nonce word test trials, the presentation of the high-TP word form caused significantly more looking to the target over the distracters in the IP-aligned condition ($P = 0.005$). In contrast, in the IP-straddling condition, presentation of the high-TP word form caused significantly more looking to the distracters compared with the target ($P = 0.011$, see *Materials and Methods* for further details), resulting in a significant three-way interaction ($P = 0.0026$) between *word prosody*, *bisyllable type* (word or the two part words), and *observation window* (corresponding to the object locations).

These results demonstrate that 6-mo-old infants can succeed at the seemingly complex task of simultaneously extracting potential word-like auditory units from short sentences and associating them with on-screen visual referents after only limited exposure in a laboratory context. Consistent with many previous studies (4, 5, 19), infants were sensitive to the statistical properties of speech, showing significant modulation of their looking behavior only for the high-TP nonce word. However, the prosodic organization of speech appeared to play a vital role in what referent the infant associated with this bisyllable: When it was aligned with an intonational phrase, infants associated it with a simultaneously moving object. In contrast, when the statistical information and prosodic information were misaligned, infants ultimately mapped the high-TP nonce word to the nontarget objects. In this case,

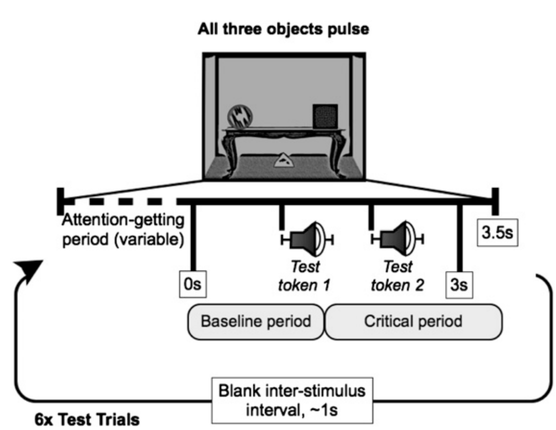

Fig. 4. Timeline of the test phase trials.

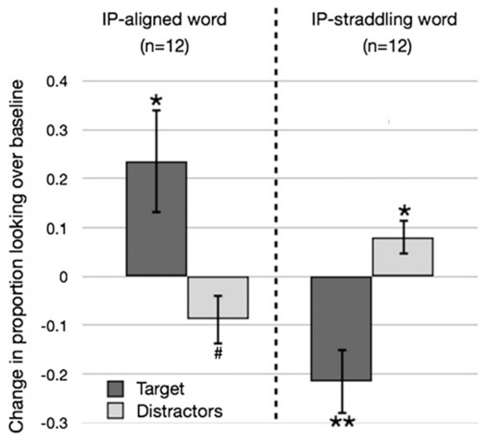

Fig. 5. Change in proportion of looking over baseline (±SEM) to the target and distracter objects upon hearing the nonce word in the two prosodic conditions (*n* = 12 each). Positive values indicate an increase in proportion of looking over baseline. $^{**}P < 0.01$; $^{*}P < 0.05$; $^{\#}P = 0.1$.

infants apparently failed to map the statistically coherent bisyllable onto the target object in training. When the bisyllable was then encountered in the test phase as a well-formed utterance, uninterrupted by a prosodic phrase boundary, infants appear to have treated it as a novel word, mapping it onto the distracters, presumably by a strategy such as mutual exclusivity (30).

These results are consistent with previous findings from adults, who also show better segmentation of nonce words that are internal to (artificially constructed) prosodic phrases compared with nonce words that straddle a phrasal prosodic constituent boundary (31). Whereas in adults such preferences could be due to experience with the language, the present findings with young infants suggest a much more fundamental processing bias.

With adult participants in artificial language learning paradigms, we can explicitly ask them whether they treat the experimental stimuli as "language" and the extracted syllable sequences as "words". With infants, we can only make inferences from the pattern of results about whether our auditory word forms are treated as words. Word learning is thought to entail more than just associating auditory and visual patterns (32). For example, we might associate a mewing sound with a kitten, but the mewing sound would not be a word referring to the kitten. It has nevertheless been proposed that auditory word forms might be treated as part of a multimodal percept that defines an object (33). That is, the word form "kitten" might be one perceptual feature of an associative, multimodal representation of a kitten that includes other perceptual features such as mewing or paws and whiskers.

There are reasons to believe that even if our results are not conclusive evidence for full-blown lexical "reference" (32), they at least entail mechanisms that go beyond such simple perceptual associations. We found that infants in the IP-internal word group associated the high-TP test word to the target object, but that infants in the IP-straddling group did not. However, the stimuli were designed so that, perceptually, the intonation pattern of the test items was more similar to the high-TP bisyllable in the IP-straddling training exemplars than in the IP-internal training exemplars (Fig. 3 and description of stimuli in *Materials and Methods*). Therefore, from a strict perceptual, associative viewpoint, the multimodal test stimulus was a better match to the multimodal percept in the training phase for infants in the IP-straddling word condition, where mapping was not observed. We therefore suggest that infants do not rely solely on associations based on multimodal perceptual similarities, but are constrained to form associations between auditory word forms and visual objects in a manner that is conducive to rapidly acquiring lexical items.

Our findings thus lend support to arguments that prosodic cues, which signal constituent edges, are critical for acquiring word forms and grammatical patterns in infants and adults (34–40).

Conclusions

Our results suggest that the cognitive capacities of infants are appropriately constrained and that language acquisition is most rapid when the structure of the linguistic input is well matched to these constraints (ref. 41, for a similar proposal). Thus, prosodically organized input may be an essential feature for optimal word learning. Moreover, our results show that simpler input is not necessarily easier to learn, and that past failures to document sophisticated capabilities in young infants could be due to an absence of the rich structural context expected by the infants' learning mechanisms.

The perception of prosodic phrases in speech has parallels to the perception of phrases in nonlinguistic domains, such as music (42, 43), and appears to be related to basic perceptual grouping principles (refs. 44 and 45, for review). To the extent that the language faculty is built up from or relies on phylogenetically preexisting cognitive capacities such as these grouping principles or the nature of memory and attentional processes more generally, studying them can offer insight into the system that enables language acquisition and the structure of language itself. In this report we have demonstrated that learning can be significantly influenced by the nature of the input, suggesting that a better understanding of infants' predispositions can lead to a better understanding of the mechanisms of language acquisition.

Materials and Methods

Participants. We tested 24 6-mo-old infants (13 males, ages ranging from 165 to 192 d). Two additional infants were excluded, one for fussiness and one for equipment failure. One additional infant was replaced to ensure normally distributed data in all conditions (*Data Analysis*). Infants were randomly assigned to one of the two *word prosody* conditions (see below).

Stimuli. Auditory stimuli were produced by the second author (K.S.W.). For the training phase, we recorded two-IP pentasyllabic sequences that were either (Type *a*)/ʒə-muː-(#)-raː-(#)-leɪ-sə/ or (Type *b*)/ʒə-muː-(#)-raː-(#)-niː-sə/; (#) marks are possible IP boundary locations. Mean durations of the syllables and the pitch height at the two pitch peaks are given in Table 1 (also Fig. 3). All acoustic analyses were carried out in Praat (46); the pitch values were smoothed with a 10-Hz bandwidth. Between two groups of infants (*n* = 12 each), the (statistical) nonce word /muː-raː/ either straddled or was aligned with an IP boundary. Test bisyllables corresponding to the nonce word and two part words (/raː-leɪ/ and /raː-niː/) were separately recorded as single utterances with a rising intonation, making them prosodically distinct from the utterances (and IPs) during training (Fig. 3 *C–E*). To ensure that the test bisyllables were not prosodically more similar to the aligned familiarization

Table 1. Mean durations for the five syllables (in ms, ±SD) and the mean pitch at the two peaks (in Hz, ±SD) for training utterances from the two conditions

	Durations, ms					Pitch peaks, Hz	
	ʒə	muː	raː	leɪ/niː	sə	Peak 1	Peak 2
IP-internal word condition	200.5 (±22)	326 (±13)	275 (±21)	293 (±21)	278 (±28)	332 (±16.3)	394.5 (±21.9)
IP-straddling word condition	231 (±10)	593 (±13)	273 (±15)	172 (±18)	312 (±22)	291 (±4)	337 (±4)

words than the nonaligned words, adults (n = 20) were presented with excised portions of the training stimuli corresponding to the IP-aligned and IP-straddling nonce words and the test version of the nonce word (all band-pass filtered between 0.1 and 1.0 kHz). In a three-alternative, spot-the-odd-man task, adults rated the IP-aligned version of the nonce word as being most dissimilar (45% of the time), compared with the IP-straddling (25%) and the test (30%) versions (these proportions are different from chance, $\chi^2 = 7.8$, $P = 0.02$). This pattern of results shows that any preference for the prosodically aligned word at test cannot be due to greater acoustic similarity between the aligned word and the test stimuli. Further, inspection of pitch tracks (in Praat) corresponding to the stimuli confirmed the adult judgments: The test stimuli exhibit a rise in pitch on the second syllable; this second-syllable rise is similarly present in the IP-straddling nonce word during familiarization, but not in the IP-aligned nonce word (Fig. 3).

Visual stimuli consisted of a scene of a "room" with a table, around which three objects were arrayed (Fig. 1). The objects were ∼200 × 200 pixels (∼5°) in size. At various points during each trial, the objects were animated and moved with respect to the table. Animation was contingent on infant looking behavior, determined by custom-designed software (47) receiving input from the eye tracker.

Procedure. Infants were seated on a caregiver's lap. Caregivers listened to masking music over sound-attenuating headphones and wore a visor to block their view of the screen. Visual stimuli were presented on the 17-inch (1,280 × 1,024 pixels) monitor of a Tobii 1750 eye tracker. Each trial began with the red object looming to 105% of its initial size and back again (i.e., the object pulsed) over 1 s, while the sound of a drum repeated until the infant looked at the screen. Once the eye-tracker software detected the infant's gaze in an observation window surrounding the red object (Fig. 1), the object pulsed and moved 50 pixels along the table toward the center of the screen, pulsed once, and returned to its original location. Simultaneous with the movement, two tokens of five-syllable sentences (utterances), separated by 1.5 s of silence, were presented at a comfortable listening level (65 dB). The two sentences on each trial were either both Type *a* or one Type *a* and one Type *b*; over nine training trials, Type *a* sentences were twice as frequent as Type *b* sentences. In the intertrial interval, infants saw a full-screen looming checkerboard, while simultaneously hearing a single trisyllabic utterance of the form [/ʒə-X-sə/], where "X" was randomly chosen from /gaː/, /duː/, or /beɪ/. Therefore, the edge syllables [/ʒə/ and /sə/] had the highest syllable frequencies and the lowest mutual information, mimicking function words in language. In addition to providing a more language-like input, this manipulation also ensured that the target syllable sequences were not at the clearest perceptual edges, so that the observed effects can be attributed to infants' perception of utterance-internal prosodic phrasing.

In each (3.5 s long) test trial, all three objects loomed to maintain infants' interest. After 1 s of silence, two tokens (separated by 500 ms of silence) of either the word or one of the two part-word bisyllables were presented. The three test items were presented in two test trials each, for a total of six test trials. Although not all infants finished all six trials, each infant completed at least one trial for each test bisyllable.

Data Analysis. Data organization and the Lilliefors tests were performed in Matlab (MathWorks), and subsequent analyses were carried out in Data Desk 6.2 (Data Description). Observation windows were 500 × 350 pixels in size, surrounding each of the three objects (the larger widths accommodate the lateral movements of the objects during training). For each object (window), we computed the proportion of looks in that window relative to the total on-screen looks for two time periods (Fig. 4): (*i*) a baseline period from the start of the infant looking to the screen until 500 ms beyond the onset of the first bisyllable and (*ii*) a critical period, 2 s beyond the baseline time period. For the purpose of comparing equivalent-sized windows, looking proportions to the two distracter objects were averaged. Proportion of looking to the target window during baseline was submitted to a 2 (*word prosody*: IP-aligned or IP-straddling) × 3 (*bisyllable type*: word, frequent part word, infrequent part word) ANOVA, with subjects as a random factor, nested under *word prosody* (all groups were normally distributed, Lilliefors test, all $P > 0.05$). None of the main effects or interactions were significant for the baseline period (all $P > 0.2$).

The dependent measure for the critical analyses was the proportion of looking for a given window in the critical time period minus the proportion of looking for that window in the baseline time period. If the infant devotes a greater percentage of time to a given window after hearing a bisyllable, the dependent measure increases above zero; it decreases below zero if the infant looks away from that window upon hearing that bisyllable. The change in proportion of looking was averaged across trials for each bisyllable × window combination for each infant. Values for all groups but one were normally distributed, therefore, all of the data from the infant with the most extreme value in that group (Z-score >2) were replaced by data from another infant. Although the pattern of results does not change, we report the newer dataset, where all bisyllable × window groups are normally distributed (Lilliefors test, all $P > 0.05$). Two-tailed t tests showed that for the nonce word in the IP-aligned condition, proportion of looking to the target increased over baseline (+0.235, $P = 0.043$), whereas it decreased marginally for distracters (−0.088, $P = 0.1$). In the IP-straddling condition, proportion of looking to the target decreased (−0.216, $P = 0.007$), whereas it increased for distracters (+0.08, $P = 0.039$). The corresponding comparisons for the part words were nonsignificant (all $P > 0.3$).

An ANOVA for the critical dependent measure was performed, with infants as a random factor, nested within *word prosody*. *Bisyllable type* and *observation window* (target or distracter) were entered as additional within-subject variables. The only significant effect was a three-way interaction between *word prosody*, *bisyllable type*, and *observation window*, $F(2,158) = 6.162$, $\eta^2 = 0.058$, $P = 0.0026$). In post hoc (Scheffé) tests, for the IP-aligned word condition, target > distracters ($P = 0.005$), whereas for the IP-straddling word condition, distracters > target ($P = 0.011$).

ACKNOWLEDGMENTS. We thank Alyssa Thatcher and Rochester BabyLab members for help with testing infants, Johnny Wen for programming assistance, and Marina Nespor and Jacques Mehler for useful comments. Research was supported by grants to R.N.A. from the National Institutes of Health (HD-30782) and the J. S. McDonnell Foundation (220020096).

1. Quine WV (1990/1992) *Pursuit of Truth* (Harvard Univ Press, Cambridge, MA).
2. Gleitman LR, Gleitman H (1992) A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Curr Dir Psychol Sci* 1:31–35.
3. Tincoff R, Jusczyk PW (1999) Some beginnings of word comprehension in 6-month-olds. *Psychol Sci* 10:172–175.
4. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928.
5. Saffran JR (2001) Words in a sea of sounds: The output of infant statistical learning. *Cognition* 81:149–169.
6. Gogate LJ, Bolzani LE, Betancourt E (2006) Attention to maternal multimodal naming by 6- to 8- month-old infants and learning of word-object relations. *Infancy* 9:259–288.
7. Smith L, Yu C (2008) Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106:1558–1568.
8. Gogate LJ (2010) Learning of syllable-object relations by preverbal infants: The role of temporal synchrony and syllable distinctiveness. *J Exp Child Psychol* 105:178–197.
9. Graf Estes KM, Evans JL, Alibali MW, Saffran JR (2007) Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol Sci* 18: 254–260.
10. Mehler J, Bertoncini J (1981) Syllables as units in infant perception. *Infant Behav Dev* 4:271–284.
11. Saffran JR, Newport EL, Aslin RN (1996) Word segmentation: The role of distributional cues. *J Mem Lang* 35:606–621.
12. Saffran JR, Johnson EK, Aslin RN, Newport EL (1999) Statistical learning of tone sequences by human infants and adults. *Cognition* 70:27–52.
13. Nespor M, Vogel I (1986/2000) *Prosodic Phonology* (Foris, Dordrecht, The Netherlands).
14. Hirsh-Pasek K, et al. (1987) Clauses are perceptual units for young infants. *Cognition* 26:269–286.
15. Mandel D, Kemler Nelson DG, Jusczyk P (1996) Infants remember the order of words in spoken sentences. *Cogn Dev* 11:181–196.
16. Vaissière J (2005) *The Handbook of Speech Perception*, eds Pisoni DB, Remez RE (Blackwell, Malden, MA), pp 236–263.
17. Christiansen M, Allen J, Seidenberg M (1998) Learning to segment speech using multiple cues: A connectionist model. *Lang Cogn Process* 13:221–268.
18. Gout A, Christophe A, Morgan JL (2004) Phonological phrase boundaries constrain lexical access: II. Infant data. *J Mem Lang* 51:547–567.
19. Thiessen ED, Saffran JR (2003) When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Dev Psychol* 39:706–716.
20. Werker JF, Cohen LB, Lloyd VL, Casasola M, Stager CL (1998) Acquisition of word-object associations by 14-month-old infants. *Dev Psychol* 34:1289–1309.
21. Aslin RN, Woodward JZ, LaMendola NP, Bever TG (1996) *Signal to Syntax*, eds Demuth K, Morgan JL (Erlbaum, Mahwah, NJ), pp 117–134.
22. Brent MR, Siskind JM (2001) The role of exposure to isolated words in early vocabulary development. *Cognition* 81:B33–B44.
23. Gleitman L, Gleitman H, Landau B, Wanner E (1988) *Linguistics: The Cambridge Survey (Language: Psychological and Biological Aspects)*, ed Newmeyer FJ (Cambridge Univ Press, New York), Vol 3, pp 150–193.
24. Fernald A, Hurtado N (2006) Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Dev Sci* 9:F33–F40.
25. Mehler J, Bertoncini J, Barriere M, Jassik-Gerschenfeld D (1978) Infant recognition of mother's voice. *Perception* 7:491–497.

26. Bertoncini J, et al. (1989) Dichotic perception and laterality in neonates. *Brain Lang* 37:591–605.
27. Ramus F, Hauser MD, Miller C, Morris D, Mehler J (2000) Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science* 288:349–351.
28. Peña M, et al. (2003) Sounds and silence: An optical topography study of language recognition at birth. *Proc Natl Acad Sci USA* 100:11702–11705.
29. Shattuck-Hufnagel S, Turk AE (1996) A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res* 25:193–247.
30. Markman EM, Wasow JL, Hansen MB (2003) Use of the mutual exclusivity assumption by young word learners. *Cognit Psychol* 47:241–275.
31. Shukla M, Nespor M, Mehler J (2007) An interaction between prosody and statistics in the segmentation of fluent speech. *Cognit Psychol* 54:1–32.
32. Waxman SR, Gelman SA (2009) Early word-learning entails reference, not merely associations. *Trends Cogn Sci* 13(6):258–263.
33. Sloutsky VM, Kloos H, Fisher AV (2007) When looks are everything: Appearance similarity versus kind information in early induction. *Psychol Sci* 18:179–185.
34. Morgan JL, Meier RP, Newport EL (1987) Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognit Psychol* 19:498–550.
35. Peña M, Bonatti LL, Nespor M, Mehler J (2002) Signal-driven computations in speech processing. *Science* 298:604–607.
36. Christophe A, Guasti MT, Nespor M, van Ooyen B (2003) Prosodic structure and syntactic acquisition: The case of the head-complement parameter. *Dev Sci* 6:213–222.
37. Seidl A, Johnson EK (2006) Infant word segmentation revisited: Edge alignment facilitates target extraction. *Dev Sci* 9:565–573.
38. Nespor M, et al. (2008) Different phrasal prominence realizations in VO and OV languages? *Lingue e Linguaggio* VII.2:1–28.
39. Shukla M, Nespor M (2010) *The Sound Pattern of Syntax*, eds Rochman L, Erteschik-Shir N (Oxford Univ Press, Oxford), pp 174–188.
40. Endress AD, Hauser MD (2010) Word segmentation with universal prosodic cues. *Cognit Psychol* 61:177–199.
41. Yang CD (2004) Universal grammar, statistics or both? *Trends Cogn Sci* 8:451–456.
42. Krumhansl C, Jusczyk PW (1990) Infants' perception of phrase structure in music. *Psychol Sci* 1:70–73.
43. Jusczyk PW, Krumhansl CL (1993) Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *J Exp Psychol Hum Percept Perform* 19: 627–640.
44. Creel SC, Newport EL, Aslin RN (2004) Distant melodies: Statistical learning of non-adjacent dependencies in tone sequences. *J Exp Psychol Learn Mem Cogn* 30:1119–1130.
45. Bregman AS (1990) *Auditory Scene Analysis* (MIT Press, Cambridge, MA).
46. Boersma P (2001) Praat, a system for doing phonetics by computer. *Glot Int* 5(9/10): 341–345.
47. Shukla M, Wen J, White KS, Aslin RN (2011) SMART-T: A system for novel fully automated anticipatory eye-tracking paradigms. *Behav Res Methods*, 10.3758/s13428-010-0056-6.