Racing to Segment?
Top-Down vs. Bottom Up in Infant Segmentation

Elena Tenenbaum and James Morgan
Brown University

1. Introduction

Explaining how infants learn words is one of the central problems in language acquisition. Learning words requires the ability to recognize them in real time, and spoken word-form recognition is an exceedingly complex skill. Because the speech signal is ephemeral, processing is subject to considerable time-pressure. Furthermore, words occur as part of the continuous flow of speech. If words were presented in isolation, word-form recognition, and hence word learning, would be greatly simplified, but in reality, infants most often hear phrases of concatenated words (van de Weijer, 2001; Aslin, Woodward, La-Mendola & Bever, 1996 but see Brent & Siskind 2001). Therefore, to identify the phonological forms of words, infants must learn to segment those words from fluent speech. Understanding how infants segment word-forms from speech during the early stages of language acquisition may thus illuminate the means by which infants bootstrap their way to parsing speech into meaningful components.

As adults we (subjectively) recognize words effortlessly and instantaneously. With access to a vast lexicon, adults can rely on their knowledge of familiar words to map lexical items onto the speech stream. Words may be recognized even before they are completely pronounced (Marslen-Wilson & Welsh, 1978), and listeners can use such recognition to forecast the end of the current word and, hence, the onset of the following word (Cole & Jakimik, 1980). In most research to date, however, it has been assumed that infants lack the requisite lexical knowledge and must therefore rely on alternative means for processing continuous streams of speech. A number of studies have shown that infants are sensitive to a variety of cues in the speech stream including stress (Jusczyk, Houston, & Newsome, 1999), statistical information (Saffran, Aslin & Newport, 1996), and prosodic boundaries (Christophe, Gout, Peperkamp & Morgan, 2003). In keeping with the tradition in this literature, we use “top-down” to refer to

*This work was supported by NIH grant HD32005 to JLM. We thank Melanie Soderstrom and Katherine White for input and recordings, Lori Rolfc, Megan Blossom and Glenda Molina for subject recruitment and testing and Sandy Waxman for her suggestion about priming addressed in the discussion.
segmentation involving preexisting lexical knowledge and “bottom-up” to refer to segmentation based on statistical analysis of properties of the speech signal.

American English-learning infants are better at recognizing words with trochaic (strong-weak) stress patterns (the dominant pattern in English) than words with iambic (weak-strong) stress patterns in fluent speech. Jusczyk, Houston, & Newsome (1999) demonstrated that whereas 10-month-olds can recognize both trochaic and iambic words, 7.5-month-olds can recognize only trochaic words. In a language such as English in which 90% of all lexical (content) words maintain stress on the first syllable (Cutler & Carter, 1987), using an initial strategy in which word onsets are equated with strong syllables is a reasonable and efficient means of segmentation for the naïve learner, but as with most bottom-up cues, it is not a fool-proof method. Jus czyk et al. also showed that younger infants were more likely to segment a weak-strong word such as guitar into the more standard strong-weak pattern by conjoining the strong stress with a following weak stressed syllable such as is, resulting in incorrect segmentations such as taris. Thus, infants must rely on other cues as well to supplement the incorrect segmentation they would achieve if they focused their attention entirely on stress.

Saffran, Aslin and Newport (1996), showed that by 8 months infants may exploit the statistical tendencies of a language for segmentation. Alternating high and low probability transitions between syllables can be used to identify which units will continue a word (high probability transitions) and which mark a word boundary (low probability transitions). Saffran et al. demonstrated that after only two minutes of exposure to an artificial language lacking stress or pho-notactic cues, 8-month-old infants were able to distinguish words from non-words based solely on the transitional probabilities between the syllables. It is clear from the minimal length of familiarization in this study that infants are extremely adept at recognizing the statistical probabilities of transitions and are likely making similar calculations when attending to their native language.

Infants are aware of prosodic boundaries within utterances and may also be incorporating this information when segmenting words from speech. Christophe, Gout, Peperkamp, & Morgan (2003) trained 13 month old infants to turn their heads upon hearing a target word and demonstrated that they cease to respond when the word crosses a prosodic boundary. Thus, by 13 months, infants recognize prosodic breaks as word boundaries and do not attempt lexical segmentation across such breaks.

All of the studies discussed thus far have focused on individual cues that infants may use in segmenting fluent speech. In reality, natural speech presents multiple cues simultaneously. In a follow-up study to Saffran et al., Thiessen and Saffran (2003) presented infants with two conflicting bottom-up cues to examine the interaction between stress and statistical information. Infants were familiarized with two minutes of an artificial language in which the words based on statistical probabilities bore weak-strong stress patterns (the non-dominant pattern for American infants). The authors found that infants at 7 months focused on the statistical cues for segmentation, ignoring the stress patterns. At 9
months, infants shifted to reliance on the stress cues, segmenting the strong-weak words in spite of the conflicting statistical information available to them. Infants’ dependence on particular cues for segmentation thus shifts across development.

One major shift that infants must make involves incorporation of lexical knowledge as in adult top-down segmentation. Though the field is somewhat divided on the extent to which adult processing relies on top-down knowledge of the lexicon, it is clear that our knowledge of words contributes substantially to our ability to recognize word boundaries. In the split between explicit segmentation models (focusing on bottom-up cues such as stress) and serendipitous segmentation models (based on strictly top-down sequential word recognition or competition between word candidates), it appears that the latter plays a substantial role in adult segmentation (Cutler, 1996).

Top-down segmentation requires a lexicon from which to draw knowledge of familiar items. Until recently, it was thought that infants lack the requisite knowledge to make use of the top-down cues in the speech stream. However, Bortfeld, Morgan, Golinkoff & Rathbun (2005) demonstrated that as early as 6 months, infants are capable of using at least some lexical knowledge for segmentation. Bortfeld et al. showed that the presence of an infant’s own name or a familiar term such as Mommy aids in the recognition of a novel word immediately following that familiar word in fluent speech. Presence of a phonotactically similar word such as Tommy on the other hand did not result in facilitation of recognition of the target word. This was the first study to indicate top-down processing of language segmentation at such an early stage of development. It further suggested that infants are capable of segmentation prior to the 7.5 month point suggested by Jusczyk and Aslin (1995).

If infants are capable of top-down segmentation at this early stage, we might expect them to rely on these cues as quickly as possible, especially given that bottom-up cues are generally less reliable than top-down information. Bottom-up cues are highly variable and often absent entirely from natural speech (Johnson & Jusczyk 2001). Furthermore, modeling of infant language processing has suggested a potential superiority of top-down cues for segmentation. In fact, Brent (1997) developed a model that relies entirely on top-down segmentation for development of a lexicon.

In Brent’s model, INCDROP, word recognition begins with utterance-length phrases and gradually segments familiar units to form a lexicon. For instance, beginning with a phrase, getit, (as in the abbreviated version of the question do you get it?) an infant confronted with the short utterance yougetit could recognize the word unit you without ever having to hear the word in isolation or appealing to signal-based mechanisms. According to Dahan and Brent (1999), “Segmentation and word discovery during native-language acquisition may be driven by recognition of familiar units from the start, with no need for transient bootstrapping mechanisms.”

Though this reliance on top-down cues works for models of infant language acquisition, it appears not to reflect the reality of the approaches taken by infants.
This may be due in part to the simplifying assumptions made by the models about the invariance of words in speech. In reality, infants seem to be relying on a constellation of bottom-up cues for segmentation even after they are capable of top-down segmentation. Previous work has examined the interaction between multiple bottom-up cues. Here, we examined the interaction between top-down and bottom-up cues in infant segmentation. We know that top-down and bottom-up cues rely on very different information and therefore are likely processed independently initially, but what ultimately leads infants to decide where to segment words? How do the bottom-up cues based on low-level statistical analysis of acoustic information in the speech stream and top-down cues based on prior knowledge of lexical items interact in segmentation? Are the cues processed in parallel such that whichever cue provides the information the fastest is relied on for identifying word boundaries? Or are the cues integrated so that the presence of two cues has an additive effect and leads to faster segmentation than when only a single type of cue is present?

To investigate the interaction between these approaches to segmentation, we examined the relative speed with which infants recognized words in fluent speech given the presence or absence of top-down and bottom-up cues. To do this, we measured the speed with which infants recognized words in four conditions. The conditions were formed by crossing presence or absence of a familiar name before the target word (the top-down cue) with the presence or absence of a prosodic boundary (the bottom-up cue). The bottom-up prosodic cue mirrored that of Christophe et al. (2003) and the top-down cue of highly familiar word preceding the target item was based on the work by Bortfeld et al. (2005). In this study, we presented infants with pictures of the target referent and a distracter and measured the speed with which infants oriented to the target referent upon hearing the word in a fluent sentence.

If top-down and bottom-up cues are processed in parallel, we would expect to see one cue or the other driving the segmentation. Once the faster cue has resolved the segmentation problem, the presence of a second cue would have no effect on the speed of recognition. Conversely, if the cues are integrated, we would expect the condition in which two cues are present to be faster than either of the single cue conditions. The condition with neither a top-down nor a bottom-up cue served as a baseline for the speed of recognition measure.

2. Method
2.1 Participants

Twenty 18-month-olds were included in the study (mean age: 547 days, range: 532–564 days, 10 male, 10 female). All subjects were full term normally developing infants from monolingual English language environments in the Providence area. An additional 15 infants were tested and discarded from analysis for the following reasons: 1 for foreign language input, 8 for fussiness, 2 for overall latencies greater than two standard deviations from the mean, 4 for insufficient data points. To help ensure that infants were attending to the task,
trials were omitted if an infant did not look to the target at all over the course of the trial or if the looks were too fast (shorter than 200ms after onset of the target word indicating that the saccade had been initiated prior to target word presentation) or too slow (greater than two standard deviations from the infant’s own mean latency).

2.2 Stimuli

Because our measure required infants to look to the referents of familiar words, sixteen of the most frequently spoken words by 18-month-olds were chosen from the UCSD Lexical Developmental Norms Database (Dale & Fenson, 1996). Surveys from the participants’ caregivers confirmed the familiarity of these words averaging 3.6 on a scale in which 4 was most familiar.

For each of the sixteen words, four sentential contexts were created. This resulted in sixteen sentence sets and a total of sixty-four sentences. Syntactic structure varied across the sentence sets. As in the example shown in Table 1, each sentence set was composed of one sentence from each of the four conditions. Here, sentence (a) contains both a top-down cue (familiar name) and a bottom-up cue (prosodic contour) for segmentation. Two sentences, (b) and (c) contain only one type of cue while (d) contains neither a prosodic boundary nor a familiar name.

All stimuli were recorded in a sound-treated booth by a female speaker from upstate New York. Individual recordings were made for each of the sixty-four sentences and then cross spliced such that within a sentence set, each of the four condition-dependent sentence beginnings were attached to the same target word and sentence ending. The sentence endings were chosen evenly from the four conditions. Importantly, pauses preceding the target word were deleted and any pause after the target word would be present in all four conditions.

The mean length of the sentences was 2869ms (range: 2148ms-3987ms). The mean length of the trial time from onset of the target word until the end of the sentence was 1737ms (range: 1246ms-3266ms). The mean length of the word preceding the target word was 799ms in the conditions with a prosodic

<table>
<thead>
<tr>
<th>Prosodic Boundary</th>
<th>Familiar Name</th>
<th>Unfamiliar Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Prosodic Boundary</td>
<td>b. Four mommies’ cars are fun to drive.</td>
<td>d. Four dobbies’ cars are fun to drive.</td>
</tr>
<tr>
<td></td>
<td>a. For mommies, cars are fun to drive.</td>
<td>c. For dobbies, cars are fun to drive.</td>
</tr>
</tbody>
</table>
2.3 Procedure

An intermodal preferential looking procedure (IPLP) was used to test the infants. Infants were seated in a testing booth on the lap of a caregiver who wore aviator headphones playing loud music to mask the experimental stimuli. At eye-level for the infants at a distance of approximately 90cm were two screens mounted on a peg board with a small light between them. Below the light, a video camera was situated such that the experimenter in an adjacent room could monitor and record the experimental sessions over a closed-circuit video system. Centered behind the pegboard was a speaker that played the auditory stimuli at a conversational level (75 dB).
Trials consisted of three phases: familiarization, salience and test. All phases began with a blinking light to draw the infant’s attention to midline. Once the infant successfully oriented to the light, the experimenter would call for the trial and the infants were presented with a picture and label of the target referent. The target was named once and the image remained on the screen for 3s. This was then repeated for the distracter (target and distracter presentations were counterbalanced for order and side of presentation). After familiarization, there was a salience phase with no audio presentation in which both pictures remained on the screen for 5s. The purpose of the salience phase was to remind the infants of the target location. Finally, there was a test in which the sentence containing the target word would play and both the target and distracter referents would appear simultaneously at the onset of the target word within the sentence. Tests lasted until the offset of the sentence.

The dependent measure in this experiment was the latency of look in test from the onset of the target word until the infant looked to the appropriate referent. The speed with which infants orient to the target referent indexes the relative speed with which they have recognized the word in the sentence. We also calculated overall looking time across the trials to the referent of the target. This allowed us to determine the certainty with which infants had recognized the target word. Had the infants been unsure of what they heard, they would be more likely to shift their attention between the two pictures.

3. Results

Because this study was predicated on the assumption that infants knew the target words and their referents, we first considered overall looking time to the target to verify that infants had succeeded in recognizing and identifying the appropriate referents. The results in Figure 2 illustrate that the infants did spend more time looking at the referent of the target and were thus successful in recognizing and orienting to the target word (mean looking time to target: 11.4s ($SE = 0.5$), mean looking time to distracter: 8.5s ($SE = 0.3$), $t(19) = 3.4$, $p < .01$). This pattern of results held for all conditions in which at least one cue was present (two cues: target: 11.6s ($SE = 0.1$), distracter: 8.8s ($SE = 0.9$), $t(19) = 1.6$, $p = .1$; familiar item: target 13.5s ($SE = 0.9$), distracter: 8.5s ($SE = 0.7$), $t(19) = 3.5$, $p < .05$, prosodic cue: target: 10s ($SE = 0.8$), distracter: 7.3s ($SE = 0.7$), $t(19) = 2.5$, $p < .05$, neither cue: target: 10s ($SE = 0.9$), distracter: 9.6s ($SE = 0.7$), $t(19) < 1$, $p > .1$).

We predicted that the presence of cues to segmentation would facilitate faster recognition and therefore shorter latencies; the open questions here concerned the relative amounts of facilitation provided by top-down and bottom-up cues and whether two cues provide more facilitation than any single cue. Latency was analyzed in an overall 2x2 within-subject ANOVA and a series of planned comparisons addressing specific questions. As shown in Figure 3, there
was a main effect of prosodic cue on the overall latency results \( F(1,19) = 8.25, p = .01 \) but no main effect of familiar word preceding the target item \( F(1,19) < 1, p > .1 \). There was no significant interaction between familiar word and prosodic cue \( F(1,19) < 1, p > .1 \).

As predicted, infants were faster to orient to the referent of the target word with two cues (mean latency: 6.7s, \( SE = 0.7 \)) than when neither cue was present (mean latency: 9.5s, \( SE = 1.1 \)). In the single cue conditions, prosody alone (mean latency: 7.1s, \( SE = 0.6 \)) was not significantly different from the two cues condition \( t(19) < 1, p > .1 \). That is, infants were just as fast to orient to the target picture when there was only a prosodic break as when there was both a prosodic break and a highly familiar lexical item preceding the target word. The presence of a familiar item preceding the target in addition to the prosodic cue did not appreciably decrease the latency from the baseline condition with no cues (mean latency in familiar cue condition: 8.5s, \( SE = 0.6, t(19) < 1, p > .1 \)).

**Figure 2:** Overall looking time measured from onset of the target word to the end of the trial (error bars show standard error).

**Figure 3:** Latency of look to target referent from onset of target word (error bars show standard error).
These data suggest that given these two cues, the bottom up information facilitates early word recognition. Furthermore, the top-down information provided by a highly familiar lexical item immediately preceding the target word in the speech stream here did not contribute substantially to the speed with which infants recognized the target word. Though presence of a highly familiar lexical item before the target word did not result in appreciable differences in initial speed of recognition, it did appear to have an effect on overall looking time. That is, when a highly familiar word preceded the target item, infants seemed more certain of what they had heard and spent less time looking back and forth between the pictures resulting in a greater percentage of looks to the target across the trial. See Figure 4.

Differing from the latency measure, there was a main effect of presence of familiar item preceding the target word ($F(1,19) = 8.43, p < .01$) but no main effect in overall looking time of presence of a prosodic cue ($F(1,19) < 1, p > .1$) and again, no interaction between the cues $F(1,19) < 1, p > .1$). There was a significant difference in percent of the trial time that the infants spent looking to the target in the two cue condition (mean: 45%, $SE = 3.5\%$) and with no cues present (mean: 36%, $SE = 3.0\%$), $t(19) = 2.04, p = .05$. The fact that there was a main effect of prosodic break on latency but not on overall looking time and a main effect of familiar word on overall looking time but not on latency indicates that overall looking time is not determined by first look.

4. Discussion

In this study, we examined the effects of top-down and bottom-up cues on infant segmentation. As a measure of the success with which infants had recognized the target words in fluent speech, we demonstrated that the overall looking time to the target picture versus the distracter was greater when at least one cue to segmentation was present.
To measure the facilitation of word recognition by segmentation cues in fluent speech, we examined the latency of looks to the target referent from onset of the target word. Our data show that segmentation was no faster with two cues than with a prosodic break alone, suggesting that in this case, immediate word recognition in 18-month-olds was driven by the bottom-up cue, with no significant difference shown for presence of a top-down cue.

The presence of a familiar word preceding the target item, however, seemed to increase the certainty of the infants about what they heard: infants spent more time looking at the referent of the target word when that word followed a familiar item. One possible interpretation for these results is that the presence of a familiar item preceding the target word helped the infants to process the target word by demanding fewer resources for retroactively establishing what they had heard. This effect would likely appear later than an immediate cue such as the prosodic break. Alternatively, the presence of an unfamiliar item (dobbies, troobies, flipsies) may have contributed to the infants searching for an appropriate referent for that item resulting in decreased looking time overall to the target referent.

Taken at face value, our data allow for resolution of the question of parallel versus integrated processing of the top-down and bottom-up cues. Specifically the fact that there was no significant difference between the latency in the two cue and prosodic cue alone conditions suggests that segmentation was driven by the prosodic cue and the presence of a familiar item preceding the target word had no additive effect on the speed with which infants recognized the target words. This suggests that the cues are being processed in parallel rather than being integrated by a common segmentation mechanism. It appears that the cues are in fact racing one another and in the race for segmentation in this case, prosody, the bottom up cue wins.

Though infants were clearly attending to the presence of the familiar item before the target word, it is interesting that this top-down cue did not show effects on the latency measure. There are a number of caveats that must be considered in interpreting these results. First of all, it is possible that the prosodic cue was simply more prominent and thus easier to use as a cue to segmentation. We know of no a priori way to balance the salience of prosody and word familiarity. If prosody was easier to detect than presence of a familiar word in our stimuli, that may have led to the predominance of the bottom-up cue in initial segmentation. Though we know from Bortfeld et al. that infants are able at 6 months to make use of top-down information in segmentation, in the present case it seems not to have speeded 18-month-olds word-form recognition. The Bortfeld study was a headturn preference study however, in which the dependent measure concerned overall preference for blocks of stimuli – note that we did find an effect of familiar words on overall looking time, a comparable measure.

Another factor that may have contributed to the results we obtained was that infants heard the label for the target word in the familiarization phase of the trial. This may have acted as a prime for the infants’ recognition of that word in fluent speech during test. It is well documented that exposure to a word speeds recog-
nition of that word in subsequent presentation (Scarborough, Cortese, & Scarborough, 1977). Such priming may have resulted in latencies of looks to the target picture that were already at floor in the prosodic cue condition, not allowing for any additional reduction in the two cue condition.

A third potential confound was the presence of a possibly salient phonotactic cue in all of the stimuli. Because this study was initially designed for use with very young infants, we were limited to the highly familiar items mommies, daddies and babies. To avoid confound, all of the unfamiliar words also ended with –ies (topsies, flipbies, dobbies). Work by Friederici and Wessels (1993) indicates that infants are highly sensitive to the phonotactic regularities within their language and would therefore likely recognize ies as a probable point for segmentation. This may have washed out potential effects of the top-down cues we provided.

Despite these caveats, as a result of this examination of the interaction between top-down and bottom-up cues in infant segmentation, we can draw certain conclusions. First, even with a growing lexicon, infants at 18-months are still relying substantially on bottom-up cues for initial segmentation. That is, contrary to the claims made by Dahan and Brent about the efficacy of top-down processing, infants clearly do need bottom-up cues to begin segmentation and infants can segment without top-down cues at all.

Furthermore, these results help illuminate the steps leading from cue recognition to segmentation. Bottom-up cues are drawn from statistical analysis of the perceived properties of the speech stream whereas top-down cues are based on lexical recognition. Because the types of computation involved are quite different, there must be independent systems for processing these cues. Eventually, however, they lead to the same end result. There are a variety of ways that this could happen. Given two potential models of segmentation, one in which information from top-down cues and bottom-up cues are integrated and the other in which the information is processed independently, our data support the independent processing mechanisms by which decisions are based on the fastest cue to provide the information necessary for segmentation.

Finally, these data provide substantial evidence that 18-month-old infants are capable of segmenting words from sentence medial positions given appropriate cues. Fernald and her colleagues (1998) had previously established that infants failed to segment words from the acoustically less salient sentence medial position at 15 months but could reliably do so at 19-months. These results suggest that by 18-months, infants are capable of segmenting words in sentence medial position, but that the ability is fragile enough to dissipate when neither a prosodic cue nor a familiar lexical item preceding the target word is present.

References


