Published in final edited form as:
Neuropsychologia. 2011 October ; 49(12): 3512–3516. doi:10.1016/j.neuropsychologia.2011.08.014.

Predictability affects early perceptual processing of word onsets in continuous speech

Lori B. Astheimer and **Lisa D. Sanders**
Department of Psychology and Neuroscience and Behavior Program, Tobin Hall, University of Massachusetts, Amherst, MA, USA 01003

Abstract

Event-related potential (ERP) evidence indicates that listeners selectively attend to word onsets in continuous speech, but the reason for this preferential processing is unknown. The current study measured ERPs elicited by syllable onsets in an artificial language to test the hypothesis that listeners direct attention to word onsets because their identity is unpredictable. Both before and after recognition training, participants listened to a continuous stream of six nonsense words arranged in pairs, such that the second word in each pair was completely predictable. After training, first words in pairs elicited a larger negativity beginning around 100 ms after onset. This effect was not evident for the completely predictable second words in pairs. These results suggest that listeners are most likely to attend to the segments in speech that they are least able to predict.

Keywords

speech perception; predictability; selective attention; auditory; ERP; N1

1. Introduction

Speech signals provide an abundance of information to the listener, not all of which is essential for comprehension. Listeners must therefore determine which portions of the speech signal should be processed in detail. For example, behavioral evidence demonstrates that word-initial segments are important for auditory word recognition (Connine, Blasko, & Titone, 1993; Marslen-Wilson & Zwitserlood, 1989; Salasoo & Pisoni, 1985), and evidence from event-related potentials (ERPs) suggests that listeners preferentially process word onsets at an early perceptual stage. In natural speech, word-initial syllables elicit a larger first negative peak (N1) compared to acoustically matched word-medial syllables (Sanders & Neville, 2003). This so-called "word-onset negativity" is evident when listeners learn to recognize sequences in several types of continuous acoustic streams, including nonsense speech (Sanders, Newport, & Neville, 2002), musical tones (Abla, Kentaro, & Okanoya, 2008) and abstract noises (Sanders, Ameral, & Sayles, 2009), suggesting that it is not simply a reflection of speech segmentation.

Corresponding Author: Lori B. Astheimer, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003, fax: 413-545-0996, lastheim@cns.umass.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The observation of an N1 enhancement in response to word- and sequence- initial segments regardless of the segmentation cues available implicates a more general processing difference such as attention directed to times that contain word-initial segments. In support of this, the latency, amplitude, and distribution of this N1 effect closely resemble ERP responses to sounds presented at attended times in temporally selective attention paradigms (Lange, Rösler, & Röder, 2003; Sanders & Astheimer, 2008). One study directly examined the use of temporally selective attention during speech perception by varying the time of auditory probe presentation relative to word onsets in a narrative. Speech-like probes presented within the first 100 ms of a word onset elicited a larger-amplitude N1 compared to probes played before word onsets or at random control times, demonstrating that listeners direct attention to moments that contain word onsets during speech perception (Astheimer & Sanders, 2009).

To date, the reason listeners attend to word onsets in speech remains largely unexplored. Potentially, all word onsets are attended because these segments are important for auditory word recognition. Alternatively, to the extent that selective attention is a tool for accommodating overwhelming amounts of information, listeners may attend to the most informative segments in speech. Based on transitional probabilities, word onsets are relatively unpredictable (Aslin, Saffran, & Newport, 1999) and therefore highly informative, which raises the hypothesis that listeners direct attention to unpredictable moments in speech.

Of course, the notion that predictability affects language processing has been studied extensively in both visual and auditory domains. Eye tracking studies demonstrate that predictable words are read more quickly and skipped more often than unpredictable words (Ehrlich & Rayner, 1981; Frisson, Rayner, & Pickering, 2005). An entire field of ERP research has grown from the observation that contextually coherent but unpredictable words elicit a larger N400 compared to predictable words (Kutas & Hillyard, 1984). Although the long latency of this effect suggests that it reflects post-perceptual, semantic processing, evidence from visual world paradigm and ERP studies suggests that listeners anticipate upcoming words even before they occur (Altmann & Kamide, 1999; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005), which may allow them to direct attention to unpredictable moments in speech. While a growing body of evidence supports a relationship between expectancy and attention outside the domain of language (Lange, 2009; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), to date, it is unclear whether predictability affects early perceptual processing of word onsets in speech.

The current study examines the relationship between predictability and attention during speech segmentation using an artificial language training paradigm that allows for comparison of ERP responses to physically identical syllables before and after they are recognized as word onsets. Participants learned to recognize all words within the language, so any recognition effect should be evident for every word after training. To test the hypothesis that the N1 attention effect is modulated by predictability, words in the language were arranged into pairs so that the second word onset in each pair was completely predictable given the first. If attention is directed to unpredictable moments in speech, then we would expect the first (unpredictable) word onset in each pair to elicit a larger N1 after training, while the second (predictable) onset in each pair would not.

2. Method

Twenty right-handed adults (6 female) ages 18–31 years (*M* = 22) contributed data to the analysis. All were Native English speakers and reported no neurological issues and no use of psychoactive medications. An additional seven participants completed the study but were excluded from the analysis due to EEG artifacts (*N* = 5) or poor behavioral performance (*N* = 2). All participants provided written informed consent and were compensated $10/hr for their time.

Eleven stop consonant-vowel (CV) syllables with durations ranging from 190–310 ms (*M* = 242) were created using text-to-speech software. These syllables were combined into six 3-syllable words (*piputu, bubapu, bidupu, dipida, putabu, tapabi*) with durations ranging from 840 to 930 ms (*M* = 874). In order to mimic transitional probabilities in natural language, some syllables were repeated in multiple words, so transitional probabilities ranged from 0.25 to 1.0 within words. The six words were arranged into three pairs such that the transitional probability between words in the same pair was 1.0 and between words in different pairs was 0.33. 100 repetitions of the pairs were arranged in a continuous 8 minute random stream (with the exception that no pair could follow itself more than once consecutively) with no acoustic boundaries between words in the stream. Three additional randomizations of the stream were created and stored as monaural WAV files with an 11.025 kHz sampling rate. To control for acoustic differences between words in different pair positions, word pair order was reversed in another set of streams, and the two versions were balanced across subjects.

For the behavioral tests, six partwords were created by combining the last syllable of one word with the first two syllables of another word. Individual words and partwords were presented aurally and participants indicated their response to each using a 4-point scale. Responses of one and two were considered “nonwords” and three and four were considered “words,” and results were transformed into a perceptual sensitivity (d′) measure for each test.

As shown in Figure 1, participants began by indicating how much they liked each item to assess any potential biases for words or partwords. Following the pretest, EEG was recorded while subjects listened passively to two of the syllable streams (16 min total). After this exposure, participants were given a second behavioral test to assess statistical learning. Next, they learned to recognize the six words from the stream with a computerized training procedure. Participants used a mouse to click on a written version of each word in order to hear the corresponding sound file, and they could train as long as necessary to memorize the words. Once they felt confident, they completed another behavioral test to assess explicit recognition of each of the words from training. This training/testing procedure was repeated until the participant reached a criterion d′ score of 1.80. Next, EEG was recorded while participants listened passively to the two remaining syllable streams. A final recognition test was then administered to assess word retention. Lastly, participants were given a predictability test that presented pairs from the stream and nonpairs, which combined the first word of one pair and the second of another. Participants indicated on a four-point scale how familiar each combination sounded. Responses of one and two were considered “nonpairs” and three and four were considered “pairs,” and a d′ score was calculated to assess awareness of word order within the stream.

EEG was recorded with a 250 Hz sampling rate from a 128-channel Geodesic Sensor Net (Electrical Geodesics, Inc., Eugene, OR) at a bandwidth of .01–80 Hz. A potassium-chloride solution was applied before each speech stream exposure to maintain impedances below 50 kΩ throughout the experiment. A 60 Hz notch filter was applied offline, and EEG was segmented into 600 ms epochs beginning 100 ms before each syllable onset. Trials were excluded if they contained eye blinks and eye movements, as determined by individual maximum amplitude criteria, or a voltage difference that exceeded 100 μV at any electrode. Only data from participants with at least 100 artifact-free trials in each condition were included in the final analysis. Averaged waveforms were re-referenced to the average mastoid and baseline corrected using the 100 ms before syllable onset.

Mean amplitude of ERPs elicited by each syllable in N1 (115–200 ms) and N400 (200–500 ms) time windows was measured at 50 electrodes arranged in pairs in a 5 × 5 grid over the scalp (Figure 2). Mean amplitudes of each electrode pair were entered into a 3 (syllable: initial, medial, & final) × 2 (training phase: before or after) × 2 (word position in pair: 1^{st} or 2^{nd}) × 5 (Left/Right position, or LR) × 5 (Anterior/Posterior position, or AP) repeated measures ANOVA (Greenhouse-Geisser adjusted). Follow-up analyses were conducted for all significant ($p < .05$) main effects and interactions.

3. Results

3.1 Behavior

Participants' behavioral responses were transformed into d′ scores for each test, and are shown in Figure 3. One-sample t-tests revealed that recognition performance did not differ significantly from chance during the pretest or after the initial speech stream exposure (*p*'s >.25). Recognition training took between five and 18 min (*M* = 9) and participants took between one and three tests (*M* = 1.7) to reach criterion. For the test on which they reached the criterion score of d′ = 1.80, performance was, of course, significantly better than chance (*M* = 2.53; *t*(19) = 20.623, $p < .001$). Despite a slight decline in performance from the criterion test to the final test (*M* = 2.21; *t*(19) = 2.189, $p < .05$), scores remained significantly above chance (*t*(19) = 11.499, $p < .001$) following the final speech stream exposure.

Results of the predictability post-test indicated that participants were able to discriminate between word pairs presented during the stream and nonpair foils (d′ *M* = 0.36, *SD* = 0.69). A single sample t-test indicated that performance was slightly but significantly above chance (*t*(19) = 2.35, $p < .05$)).

3.2 Event-Related Potentials

As shown in Figure 4, syllable onsets elicited a series of auditory evoked potentials that were broadly distributed over medial central regions. Although using continuous speech streams resulted in low amplitude ERPs, the first negative peak (N1) had a latency of 120 ms and was largest over central/medial electrodes (LR × AP *F*(16,304) = 3.087, $p < .01$), similar to what is observed in response to abrupt acoustic onsets. Across all electrode sites, there was a main effect of syllable position on N1 amplitude (*F*(2,38) = 20.947, $p < .001$), with word-initial syllables eliciting a larger N1 than medial or final syllables. Although these differences across syllable position were evident before training (*F*(2,38) = 15.383, *p* < 001), this difference was not modulated by pair position ($p >.4$). These pre-training differences likely reflect variability in the acoustic properties of syllables in different word positions. Importantly, the main effect of syllable position was qualified by a Syllable × Pair Position × Training interaction (*F*(2,38) = 3.923, $p < .04$). Planned comparisons revealed that, as shown in Figure 4, the initial syllable from the *unpredictable* word in each pair elicited a larger N1 after training compared to before (*F*(1,19) = 5.185, $p < .04$). Although the response to the *predictable* word-initial syllable appeared to be smaller after training, this effect did not approach significance (*F*(1,19) = 2.574, p=.125).

Later differences in the waveforms were also modulated by syllable position. A broad negativity that began around 300 ms was time-locked to word onsets rather than other syllables (*F*(2,38) = 9.043, $p < .001$), and had a right-lateralized scalp distribution (Syllable × LR *F*(8,152) = 3.916, $p < .01$). This later difference was not modulated by pair position ($p > .15$) or training ($p > .8$).

4. Discussion

The current study tested the hypothesis that attention is directed to times in speech that provide unpredictable information by manipulating the predictability of word onsets in an artificial language paradigm. As found previously (Sanders et al., 2002), unpredictable word onsets elicited a larger N1 after recognition training. Importantly, this enhancement was absent for the completely predictable second word onset in each pair. These results indicate that listeners selectively attend to word onsets that cannot be predicted from the context, enhancing early perceptual processing of information presented at these times.

Attending to word onsets is an effective listening strategy insofar as word onsets are relatively unpredictable and therefore highly informative. However, when a word onset is highly predictable, it no longer offers novel information, and so the listener must only attend enough to confirm that the incoming speech signal matches his or her prediction. In the current study, participants showed no early attention effect in response to completely predictable word onsets. This observation clarifies the nature of the previously reported “word onset negativity” (Sanders & Neville, 2003) by demonstrating that in some cases, auditory word recognition can proceed without an increase in attention. Therefore, listeners may not be selectively attending to word onsets, but rather to unpredictable moments in speech. Although completely predictable word onsets like those in the current study may not exist in natural speech, there is ample evidence that listeners are sensitive to even subtle differences in contextual (Van Berkum, et al., 2005) and syntactic (Mattys, Melhorn, & White, 2007) constraint during speech perception, so the allocation of attention may also be sensitive to these differences.

An obvious concern is the possibility that when presented with 3-syllable words arranged in pairs, listeners segmented streams into 6-syllable words on the basis of statistical cues alone. Participants’ performance was at chance on the recognition test following the first speech stream exposure, suggesting that the statistical regularities in this particular artificial language were not sufficient for listeners to learn the 3-syllable words. Although this observation runs counter to previous artificial language studies that report various degrees of statistical learning (Cunillera et al., 2009; Cunillera, Toro, Sebastián-Gallés, & Rodríguez-Fornells, 2006; Saffran, Newport, & Aslin, 1996), the high transitional probability between pairs of words may have disrupted statistical learning. Subjects’ performance on the predictability test that assessed knowledge of word pair order also indicates that listeners did not reliably group the stream into 6-syllable strings even by the end of the experiment. It is possible that listeners used transitional probability to segment the streams into 6-syllable groups before training, relearned them as 3-syllable words during training, and forgot the 6-syllable sequences by the final behavioral test. However, this pattern of learning would have resulted in a lack of training effects on ERPs for the unpredictable words that were heard as onsets both before and after training and differences in ERPs for the predictable words that were heard as onsets only after training. The opposite pattern of ERP effects was observed.

The mediocre performance on the predictability test makes the modulation of attention by differences in predictability somewhat surprising. Participants were not instructed to anticipate upcoming words in the stream, and their behavioral performance indicates that they were only vaguely aware of word pair order, and yet they allocated attention accordingly. This suggests that listeners make predictions about upcoming stimuli without conscious awareness. This type of anticipation has been reported across the domains of language processing (Altmann & Kamide, 1999; Van Berkum, et al., 2005) and visual perception (Bar, 2007; Zacks et al., 2007). Bar (2007) observes that the brain areas involved in generating predictions share a striking overlap with the “default network” observed in many neuroimaging studies, suggesting that the brain automatically generates associative predictions.

Predictability is only one of many potential cues that listeners employ to allocate attention to word onsets in speech. Studies of speech segmentation demonstrate that listeners use knowledge of a language's phonotactic, lexical, syntactic and prosodic structure to find word boundaries. The current study demonstrates the need to consider the association between these cues and predictability rather than just word boundaries, because they may serve as attention cues rather than segmentation cues, thus allowing listeners to preferentially process the most informative portions of the speech signal.

Acknowledgments

This research was supported by NIH, NIDCD training fellowship F31DC009965 to LBA. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH or NIDCD. We thank Charles Clifton and two anonymous reviewers for comments on earlier versions of this manuscript

References

Abla D, Kentaro K, Okanoya K. On-line assessment of statistical learning by event-related potentials. J Cogn Neurosci. 2008; 20(6):952–964. [PubMed: 18211232]

Altmann GTM, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition. 1999; 73(3):247–264. [PubMed: 10585516]

Aslin, R.; Saffran, J.; Newport, E. Statistical learning in linguistic and nonlinguistic domains. In: MacWhinney, B., editor. The emergence of language. Mahwah: Lawrence Erlbaum Associates Publishers; 1999. p. 359-380.

Astheimer L, Sanders L. Listeners modulate temporally selective attention during natural speech processing. Biol Psychol. 2009; 80(1):23–34. [PubMed: 18395316]

Bar M. The proactive brain: using analogies and associations to generate predictions. Trends Cogn Sci. 2007; 11(7):280–289. [PubMed: 17548232]

Connine C, Blasko D, Titone D. Do the beginnings of spoken words have a special status in auditory word recognition? J Mem Lang. 1993; 32(2):193–210.

Cunillera T, Càmara E, Toro J, Marco-Pallares J, Sebastián-Gallés N, Ortiz H, Pujol J, et al. Time course and functional neuroanatomy of speech segmentation in adults. Neuro Image. 2009; 48(3): 541–553. [PubMed: 19580874]

Cunillera T, Toro J, Sebastián-Gallés N, Rodríguez-Fornells A. The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. Brain Res. 2006; 1123(1):168–178. [PubMed: 17064672]

Ehrlich SF, Rayner K. Contextual effects on word perception and eye movements during reading. J Verb Learn Verb Beh. 1981; 20(6):641–655.

Frisson S, Rayner K, Pickering MJ. Effects of contextual predictability and transitional probability on eye movements during reading. J Exp Psychol Learn Mem Cog. 2005; 31(5):862–877.

Kutas M, Hillyard S. Brain potentials during reading reflect word expectancy and semantic association. Nature. 1984; 307:161–163. [PubMed: 6690995]

Lange K. Brain correlates of early auditory processing are attenuated by expectations for time and pitch. Brain Cogn. 2009; 69(1):127–137. [PubMed: 18644669]

Lange K, Rösler F, Röder B. Early processing stages are modulated when auditory stimuli are presented at an attended moment in time: an event-related potential study. Psychophysiology. 2003; 40(5):806–817. [PubMed: 14696734]

Marslen-Wilson W, Zwitserlood P. Accessing spoken words: the importance of word onsets. J Exp Psychol: Hum Percept Perform. 1989; 15(3):576–585.

Mattys S, Melhorn J, White L. Effects of syntactic expectations on speech segmentation. J Exp Psychol: Hum Percept Perform. 2007; 33(4):960–977. [PubMed: 17683240]

Saffran J, Newport E, Aslin R. Word segmentation: The role of distributional cues. J Mem Lang. 1996; 35(4):606–621.

Salasoo A, Pisoni D. Interaction of knowledge sources in spoken word identification. J Mem Lang. 1985; 24(2):210–231.

Sanders L, Ameral V, Sayles K. Event-related potentials index segmentation of nonsense sounds. Neuropsychologia. 2009; 47(4):1183–1186. [PubMed: 19056408]

Sanders L, Astheimer L. Temporally selective attention modulates early perceptual processing: event-related potential evidence. Percept Psychophys. 2008; 70(4):732–742. [PubMed: 18556935]

Sanders L, Neville H. An ERP study of continuous speech processing I. Segmentation, semantics, and syntax in native speakers. Cogn Brain Res. 2003; 15(3):228–240.

Sanders L, Newport E, Neville H. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. Nat Neurosci. 2002; 5(7):700–703. [PubMed: 12068301]

Van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. J Exp Psychol: Learn Mem Cog. 2005; 31(3):443–467.

Zacks J, Speer N, Swallow K, Braver T, Reynolds J. Event perception: a mind-brain perspective. Psychol Bull. 2007; 133(2):273–293. [PubMed: 17338600]

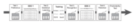

Figure 1.
Experimental paradigm. A single experimental session consisted of eight phases, which alternated between behavioral testing or training and syllable stream exposure while EEG was recorded.

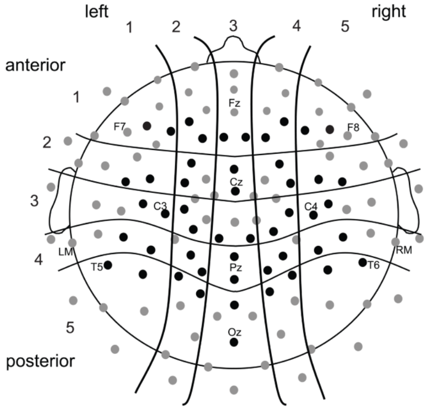

Figure 2.

Approximate scalp location of 128 recording electrodes. Measurements were taken from the 50 electrodes shown in black and averaged into 25 pairs arranged in a 5 (Left/Right position) × 5 (Anterior/Posterior position) grid.

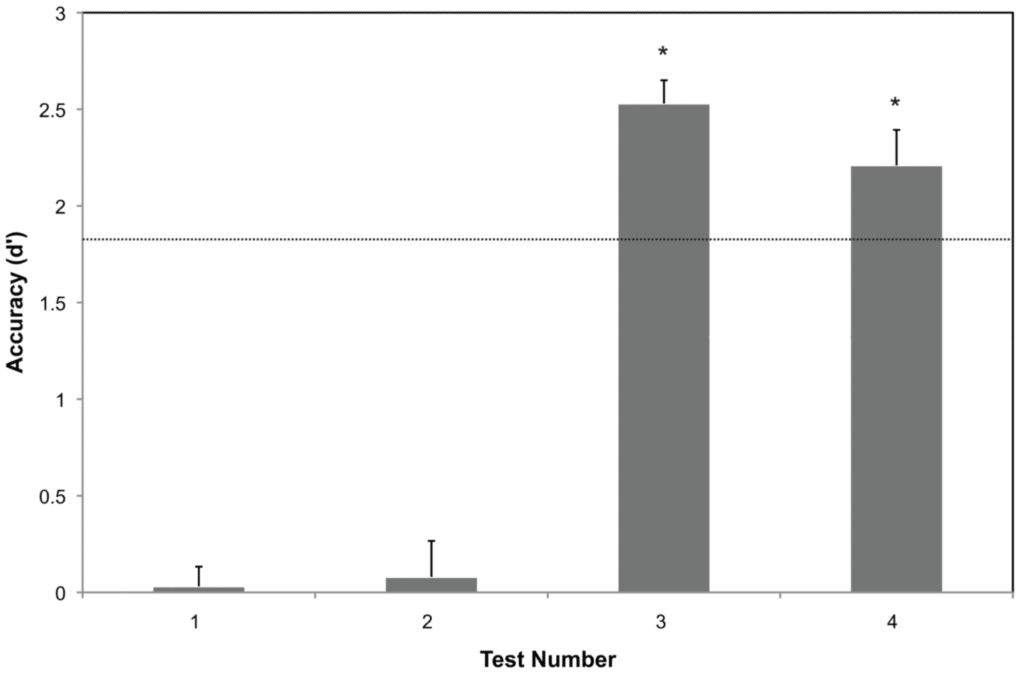

Figure 3.

Accuracy scores (d′) and standard errors on four behavioral tests: (1) before speech stream exposure, (2) after 16 minutes of exposure, (3) after mastering the words during training, and (4) after 16 additional minutes of exposure. Zero indicates chance performance, and the dotted line indicates the training criterion (d′ = 1.8). Performance on tests 3 and 4 was significantly above chance (* indicates p < .001).

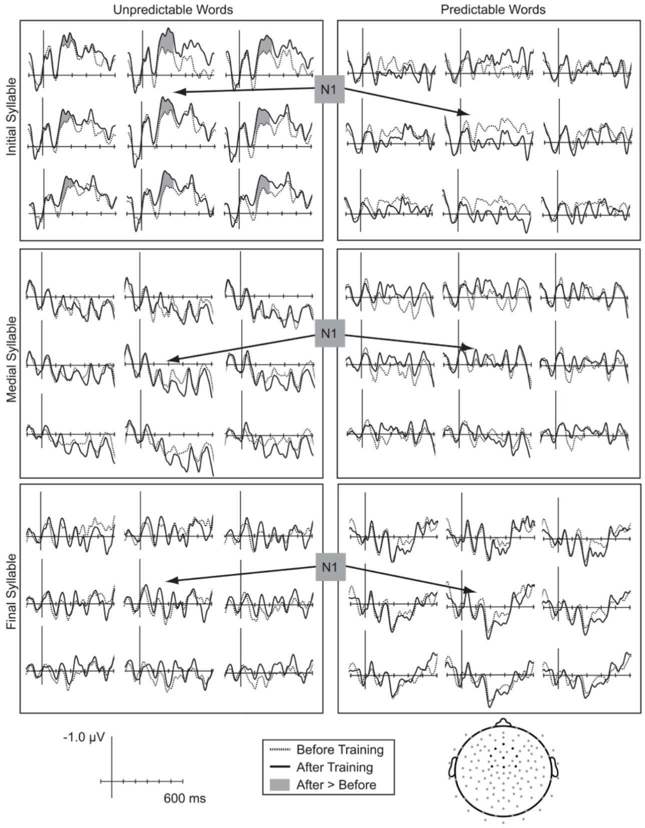

Figure 4.

Event-related potentials elicited by initial, medial, and final syllable onsets in unpredictable and predictable words before and after recognition training. Data are shown from 9 central electrode sites, indicated on the electrode map. After training, word-initial syllables in unpredictable words elicited a larger negativity between 115 and 200 ms.