


Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation


Mara Breen, Laura C. Dilley, J. Devin McAuley & Lisa D. Sanders

To cite this article: Mara Breen, Laura C. Dilley, J. Devin McAuley & Lisa D. Sanders (2014) Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation, *Language, Cognition and Neuroscience*, 29:9, 1132-1146, DOI: [10.1080/23273798.2014.894642](https://doi.org/10.1080/23273798.2014.894642)

To link to this article: <http://dx.doi.org/10.1080/23273798.2014.894642>

 Published online: 04 Mar 2014.

 Submit your article to this journal [↗](#)

 Article views: 128

 View related articles [↗](#)

 View Crossmark data [↗](#)

Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation

Mara Breen^{a,b,*}, Laura C. Dilley^{c,d,e}, J. Devin McAuley^d and Lisa D. Sanders^b

^aDepartment of Psychology and Education, Mount Holyoke College, South Hadley, MA 01075, USA; ^bDepartment of Psychology, University of Massachusetts, Tobin Hall, Amherst, MA 01003, USA; ^cDepartment of Communicative Sciences and Disorders, Michigan State University, East Lansing, MI 48824, USA; ^dDepartment of Psychology, Michigan State University, Psychology Building, East Lansing, MI 48824, USA; ^eDepartment of Linguistics and Germanic, Slavic, Asian, and African Languages, Michigan State University, East Lansing, MI 48824, USA

(Received 3 July 2013; accepted 5 February 2014)

Prosodic context several syllables prior (i.e., distal) to an ambiguous word boundary influences speech segmentation. To assess whether distal prosody influences early perceptual processing or later lexical competition, EEG was recorded while subjects listened to eight-syllable sequences with ambiguous word boundaries for the last four syllables (e.g., *tie murder bee* vs. *timer derby*). Pitch and duration of the first five syllables were manipulated to induce sequence segmentation with either a monosyllabic or disyllabic final word. Behavioural results confirmed a successful manipulation. Moreover, penultimate syllables (e.g., *der*) elicited a larger anterior positivity 200–500 ms after the onset for prosodic contexts predicted to induce word-initial perception of these syllables. Final syllables (e.g. *bee*) elicited a similar anterior positivity in the context predicted to induce word-initial perception of these syllables. Additionally, these final syllables elicited a larger positive-to-negative deflection (P1-N1) 60–120 ms after onset, and a larger N400. The finding that prosodic characteristics of speech several syllables prior to ambiguous word boundaries modulate both early and late event-related potentials (ERPs) elicited by subsequent syllable onsets provides evidence that distal prosody influences early perceptual processing and later lexical competition.

Keywords: prosody; speech segmentation; event-related potentials; temporal attention

Unlike in most written languages, where words are separated by spaces, spoken language contains no such consistent cues to word boundaries. For example, silences in speech frequently occur within words rather than only occurring at word boundaries (Cole & Jakimik, 1980; Lehiste, 1972; Nakatani & Dukes, 1977). Therefore, speech comprehension necessarily entails a process of segmenting continuous speech into words. The current study employed the precise temporal resolution of event-related potentials (ERPs) to investigate the time course of segmentation based on prosodic cues not immediately adjacent to ambiguous word boundaries.

Prior work demonstrates that listeners can use a wide variety of cues to segment speech, including statistical regularities (Saffran, Aslin, & Newport, 1996), phonotactics (Mattys, Jusczyk, Luce, & Morgan, 1999) and constraints on possible words (Norris, McQueen, Cutler, & Butterfield, 1997). In addition, listeners interpret prosodic cues occurring on or around a potential word boundary as signalling that boundary (Cutler & Butterfield, 1992; Salverda, Dahan, & McQueen, 2003; Salverda et al., 2007).

Two widely studied prosodic cues to word boundaries are metrical stress and duration. Stressed syllables in English are produced with longer durations and greater

intensity than unstressed syllables (Beckman, 1986; Fry, 1955). Postulating a word boundary before a stressed syllable is an advantageous segmentation strategy; for example, 81% of words in the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995) are stress-initial (Vroomen & de Gelder, 1995). Moreover, Cutler and Norris (1988) estimated that 85–90% of content words in everyday English speech have initial stress. Evidence shows that native English speakers use this consistency and interpret stressed syllables as word onsets (cf. Cutler & Butterfield, 1992; Cutler & Norris, 1988).

Durational cues are exemplified by the fact that syllables are lengthened in certain positions. For example, the syllable ‘ham’, produced as a monosyllabic word, is typically longer than the syllable ‘ham-’ produced as the initial syllable of ‘hamster’. Salverda et al. (2003) demonstrated, using eye-tracking, that listeners use syllable duration to determine whether those sequences are monosyllabic (as opposed to disyllabic) words. When listeners heard tokens of ‘hamster’ with the syllable ‘ham-’ spliced from the word ‘ham’, they made more early looks to a picture of a ham than a picture of a hamster, indicating that the longer version of ‘ham’ was interpreted as preceding a word boundary. In follow-up

*Corresponding author. Email: mbreen@mtholyoke.edu

work, Salverda et al. (2007) demonstrated that listeners also use segment length to determine the location of prosodic phrase boundaries. Listeners viewed a display with pictures of a cat, a cap and a captain, while listening to sentences in which the target word ‘cap’ was phrase-medial or phrase-final. Eye-tracking results demonstrated that when ‘cap’ was in phrase-final position, the monosyllabic word ‘cat’ was a stronger competitor for the target ‘cap’ than when it was in phrase-medial position. Conversely, the disyllabic word ‘captain’ was a stronger competitor for ‘cap’ when ‘cap’ was in phrase-medial as opposed to phrase-final position, demonstrating that listeners are aware that ‘cap’ is longer in phrase-final position than in phrase-medial position and can use this information for segmentation.

The findings described earlier demonstrate segmentation effects of prosodic cues occurring at the location of a word boundary. Dilley and colleagues (Brown, Salverda, Dilley, & Tanenhaus, 2011; Dilley, Mattys, & Vinke, 2010; Dilley & McAuley, 2008; Dilley & Pitt, 2010) have demonstrated that speech segmentation is also influenced by distal prosodic cues (i.e., cues not directly adjacent to the location of the potential word boundary). In particular, Dilley and colleagues reported support for a perceptual grouping hypothesis, whereby repeating prosodic patterns of pitch and/or duration distal from the to-be-segmented acoustic material influence how syllables were grouped into words in a manner consistent with general principles of auditory perceptual organisation (cf. Handel, 1989).

In one series of experiments (Dilley et al., 2010; Dilley & McAuley, 2008), participants listened to sequences of eight syllables where the last four syllables could be segmented in one of two ways. For example, in the sequence *banker helpful* /tɑɪ mæ dæ bi/, the final four syllables could be interpreted as *timer derby* or as *tie murder bee*. The pitch and duration of the first five syllables were manipulated via resynthesis to create two conditions. In the monosyllabic condition, the first five syllables were manipulated to induce listeners to segment the last four syllables with a monosyllabic final word (e.g. *tie murder bee*). In the disyllabic condition, the first five syllables were manipulated to induce listeners to segment the last four syllables with a disyllabic final word (e.g. *timer derby*; a similar manipulation is shown in Figure 1). Critically, the final three syllables were acoustically identical across conditions, but the listener’s interpretation of the segmentation of these syllables varied depending on the distal prosody of the first five syllables.

In the disyllabic condition (Figure 1, top), the first two words *banker* and *helpful* were produced with rising pitch, with a low tonal target on the first syllable, and a high target on the second syllable. Listeners’ knowledge of the strong–weak stress pattern of these words was predicted to induce them to interpret the low tonal target as a stressed syllable, and the high target as an unstressed syllable. The perceptual grouping hypothesis predicted that listeners would perseverate in their interpretation of the low target as a stressed syllable, thereby interpreting /tɑɪ/ and /dæ/ as

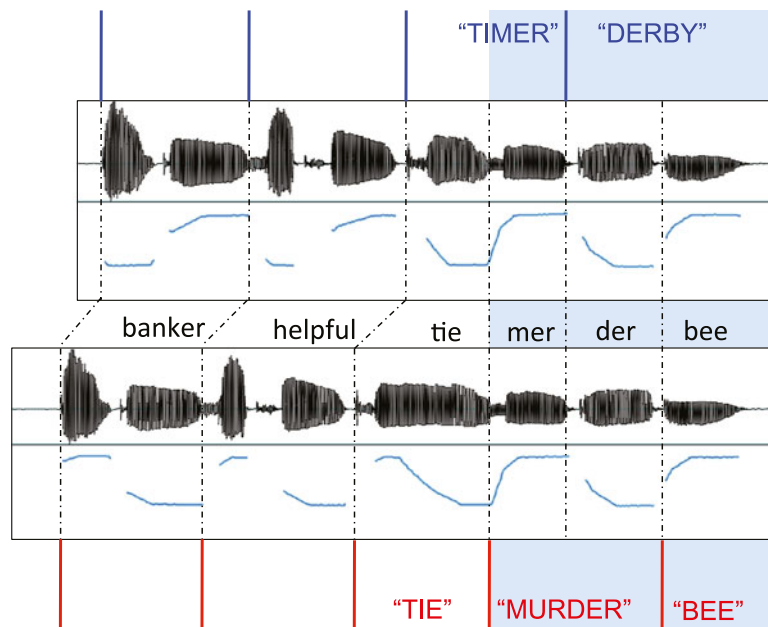


Figure 1. Explanation of the acoustic manipulation applied to the target syllable sequences. The two waveforms and associated pitch tracks show how the Disyllabic (top) and Monosyllabic (bottom) contexts were created by varying the fundamental frequency (F0) of the first five syllables, and the duration of the fifth syllable; the acoustic characteristics of the final three syllables were held constant. The words in all caps indicate the predicted segmentation of the final four ambiguous syllables. The highlighted region indicates the sections of the waveforms which are acoustically identical across conditions. See the text for more information.

word onsets, segmenting the final four syllables as *timer derby*. An additional cue to this segmentation pattern is the fact that the stressed syllables /taɪ/ and /dɛə/ occur at regular temporal (i.e., perceptually isochronous) intervals.

The opposite pattern held in the monosyllabic condition (Figure 1, bottom). Falling pitch was imposed on the first two words, *banker* and *helpful*, such that there was a high tonal target on the first syllable and a low tonal target on the second syllable. Listeners' knowledge of these words was expected to lead them to interpret the high tonal target as stressed and the low target as unstressed. Once again, the perceptual grouping hypothesis predicted that listeners would perseverate in how the syllables are grouped into words, such that they would be more likely to interpret /taɪ/, /mɛə/ and /bi/ as word onsets, thereby segmenting the final four syllables as *tie murder bee*. In addition, the fifth syllable /taɪ/ was lengthened such that, as in the disyllabic condition, word onsets occurred at regular temporal intervals

Supporting the perceptual grouping hypothesis, participants reported hearing a disyllabic final word more often in the disyllabic condition than in the monosyllabic condition (93% vs. 35%; Dilley et al., 2010; see also Dilley & McAuley, 2008). In sum, the results from this and other studies show that distal prosody can influence segmentation decisions (Brown et al., 2011; Dilley & Pitt, 2010; Reinisch, Jesse & McQueen, 2011a, 2011b).

This line of research raises the question of *when* distal prosody affects speech segmentation. One possibility is that distal prosody allows listeners to make predictions about where word boundaries are likely to occur that then influence early perceptual processing. A second possibility is that distal prosody affects lexical access, influencing the competition between possible lexical items in a manner that affects later post-perceptual processing. Salverda et al. (2003) suggests that some prosodic cues are used early in processing, demonstrating that local duration cues influence eye movements within the first 100 ms after word onset.

Most relevant here is a recent study by Brown et al. (2011) who demonstrated an online effect of distal prosody using an eye-tracking paradigm. Subjects heard sentences like 'Heidi sometimes saw that panda in the city zoo'. The acoustic characteristics of the final portions of sentences (e.g., 'that panda in the city zoo') were constant across conditions; however, the initial five syllables (e.g., 'Heidi sometimes saw') were manipulated to induce listeners to perceive prosodic phrase boundaries at different places in subsequent material. In one condition, these syllables were resynthesised with a repeating low-high tonal pattern such that low targets were aligned with 'Hei', 'some' and 'saw', and high targets were aligned with 'di' and 'times'. In the other condition, these syllables were resynthesised with a repeating high-low tonal pattern such that high targets were aligned with 'Hei' and 'some', and

low targets were aligned with 'di' and 'times'. Moreover, the syllable 'saw' was lengthened and resynthesised with a falling high-low tonal pattern. These two patterns are very similar to those presented in Figure 1, with the first resynthesis reflected in the top panel, and the second reflected in the bottom panel. In the former case, listeners were expected to perceive phrase boundaries at the lexical boundaries of the critical word (e.g. |panda|, where '|' indicates a prosodic phrase boundary); in the latter case, listeners were expected to perceive a boundary in the middle of the critical word (e.g. pan|da). Consistent with predictions, analyses of fixation patterns from 200 to 566 ms after the onset of the target word (i.e., from the earliest point at which signal-driven fixations were expected, until 200 ms after the mean offset of the embedded word) revealed more looks to a picture of the competitor ('pan') in the latter condition than in the former condition. These results indicate that when the distal prosody induced listeners to perceive a prosodic phrase boundary after 'pan-', listeners also postulated a word boundary at that point.

Additional evidence for the early use of distal prosodic cues in segmentation comes from Dilley et al.'s (2010) study. In a cross-modal identity priming task, listeners heard a string of lexical items ending in syllables that could be segmented as a disyllabic or monosyllabic final word (e.g., 'turnip' vs. 'nip'). Following the ambiguous auditory string, listeners performed lexical decision on a visually presented target. Results demonstrated that the distal prosodic manipulation influenced lexical decision times within the first 1000 ms after visual presentation; reaction times were faster for the visual target that matched the auditory word supported by the distal prosody.

These studies provide evidence that prosody has an online effect on speech segmentations. However, there are at least two reasons why it is unclear whether the observed effects are the result of listeners making predictions about word boundary locations *before* or *after* lexical access. First, effects in eye movement studies are quantified with respect to how likely participants are to look at a target object compared to a competitor object in a visual display. For example, Brown et al. (2011) measured how likely participants were to look at a target picture of a panda compared to a competitor picture of a pan. In order to decide whether to look at either of these objects, participants have to hear at least part of the target syllable 'pan-', meaning that the earliest effects are observable only after the listener has engaged in some level of lexical processing. Second, any effect of context on eye movements is constrained by the time necessary to programme an eye movement, estimated to take between 100 ms (Altmann, 2011) and 200 ms (Allopenna, Magnuson, & Tanenhaus, 1998; Matin, Shao, & Boff, 1993). Therefore, the most conservative estimate of the time course of the

effects observed in Brown et al.'s (2011) study is that distal prosody could be affecting segmentation decisions anywhere from immediately at the word onset through the end of the embedded competitor word (i.e., from 0 ms after the onset of the embedded word in Brown et al.'s stimuli until an average of 366 ms after this onset). The cross-modal identity priming study performed by Dilley et al. (2010) raises similar questions; some lexical material must have been heard before an effect could be observed, and the precise time-course of segmentation of speech into words cannot be determined in this paradigm.

To address these questions related to the time-course of effects of distal prosody on speech segmentation, the current study used an event-related potential (ERP) paradigm, which allows for greater temporal resolution of segmentation effects than eye-tracking or lexical decision tasks. Specifically, ERPs have the potential to show differences in the perceptual processing of target syllables before any lexical access has taken place.

Spoken words (and sounds, more generally) elicit a sequence of peaks in the ERP waveform, which can be classified as early (i.e. perceptual) or late (i.e., post-perceptual). In the first 200 ms after stimulus onset, sounds typically elicit a first positive peak (P1) between 50 and 90 ms, a first negative peak (N1) between 100 and 150 ms, and a second positive peak (P2) between 150 and 200 ms. The amplitude and latency of these peaks depend on several factors, including the abruptness and intensity of the sound onset, and the density of the sound environment (Näätänen & Picton, 1987).

Auditory evoked potentials in the first 200 ms after onset also vary in amplitude depending on a listener's state. For example, the amplitude of the N1 deflection relative to a pre-stimulus baseline or the amplitude of the P1-to-N1 deflection (i.e., the difference in amplitude between the P1 and the N1) is typically larger for sounds presented from attended compared to unattended locations (Hansen, Dickstein, Berka, & Hillyard, 1983; Hillyard, Hink, Schwent, & Picton, 1973). Further, when a rapidly changing stream of sound, such as speech, is presented from a single location, sequence onsets such as the initial segments of words elicit a larger amplitude P1-to-N1 deflection relative to similar event onsets that do not begin a new sequence within the larger continuous stream (Abla, Katahira, & Okanoya, 2008; Sanders, Améral, & Sayles, 2009; Sanders & Neville, 2003; Sanders, Newport, & Neville, 2002). In many of these studies, participants heard sequences of syllables, tones or non-verbal sounds with no acoustic markers (e.g., silence) to indicate sequence boundaries. Therefore, participants could only segment the stream after learning the items based on distributional cues (e.g., Abla et al., 2008) or explicit training (e.g., Sanders et al., 2002, 2009). In a study of tone segmentation, Abla et al. (2008) observed larger N1s to the initial tone in three-tone sequences than to the

medial or final tones in listeners who demonstrated learning most of the sequences on behavioural tests. Sanders et al. (2009) found a similar effect for the initial segments of non-verbal sound sequences in listeners explicitly taught to recognise the sequences. Studies of syllable segmentation have revealed similar results. For example, Sanders et al. (2002) demonstrated a larger P1-to-N1 deflection in response to the onsets of sequences of nonsense syllables after training as compared to before. In addition, Sanders and Neville (2003) observed larger amplitude P1-to-N1 deflections to word-initial syllables (*decisive*) compared to non-initial syllables (*pedestrians*) in natural speech.

Importantly, previous ERP studies of speech segmentation have demonstrated larger N1 or P1-to-N1 amplitudes in response to word-initial syllable onsets compared to word-medial syllable onsets regardless of the segmentation cues that are available. The ERP effects were similar in timing, distribution and amplitude when listeners were processing sentences that sounded like their native language but included only non-words (Sanders & Neville, 2003) and when listeners were processing a newly learned, six-word artificial language with no acoustic cues associated with word boundaries (Sanders et al., 2002). That is, the effects of segmentation on early perceptual processing of word onsets were identical when only acoustic segmentation cues were available and when only lexical segmentation cues were available. If this ERP effect were specific to segmentation itself, we would expect it to differ depending on which cues are available to the listener.

The fact that the N1/P1-to-N1 effect observed in these studies does not differ across cues suggests that it reflects a more general cognitive process. Indeed, Astheimer and Sanders (2009) argue that this general cognitive processing difference for word and syllable onsets is selective attention. They demonstrated that auditory probes presented concurrently with a speech stream elicit larger amplitude N1s when played during the first 150 ms of a content word compared to either the 150 ms preceding a content word or during random control times. From this result, they concluded that listeners direct attention to the initial portions of words in continuous speech, and this increased attention results in a larger amplitude response to the sounds played at those critical times. Directing attention to the initial portions of words in continuous speech is an effective processing mechanism because onset segments are typically less predictable from the context than segments in the middle of a word (Connine, Blasko, & Titone, 1993; Marslen-Wilson & Zwitserlood, 1989).

The amplitude of the P1-to-N1 deflection provides a tool for investigating the time course of distal prosodic influence on speech segmentation. If this effect is indeed indexing temporal attention then we expect differential effects for the final syllable of the stream contingent upon

whether the prior syllable had been perceived as the end of a word. Specifically, if listeners are using distal prosody to make predictions about upcoming word boundaries, syllables perceived as word onsets should elicit larger P1-to-N1 deflections than the same syllables perceived as word-final. As Dilley et al. (2010) demonstrated, listeners were more likely to report hearing the disyllabic word *derby* following the manipulation in the top of Figure 1, and the monosyllabic word *bee* after the manipulation in the bottom. We expect that these perceptual effects will translate into differences in P1-to-N1 amplitude. For example, listeners who have heard the words *tie* and *murder* have no basis for predicting which segments they will hear next, and so would be more likely to direct attention to the upcoming syllable, resulting in a larger P1-to-N1 deflection. On the other hand, listeners who have heard the word *timer* followed by the syllable *der* could predict that the next syllable is likely to be *by* to complete the word *derby*, and would not have to direct as much attention to the sounds of the final syllable. Indeed, recent evidence shows that the N1 word-onset effect is eliminated when listeners can predict which word is going to be heard next based on the context (Astheimer & Sanders, 2011).

On the other hand, if distal prosody is only affecting later processing, reflecting the output of, rather than the input to lexical competition, then differences based on distal prosody will be evident only in later portions of the waveform. Previous studies have also observed segmentation effects on the N400, a negative-going deflection between 300 and 500 ms after word onset. Unlike the early P1-to-N1 deflection, the N400 is thought to index post-perceptual processing. Specifically, the N400 has been shown to index the difficulty of lexical access, as it is larger when the target word is semantically anomalous given prior context (Kutas & Federmeier, 2011). Many of the studies reporting larger P1-to-N1 deflections to sequence onsets compared to sequence-medial segments have also reported larger N400s to onsets, which has been proposed to index the amount of learning and the ease with which sequences were matched onto stored representations (Abla et al., 2008; Cunillera, Toro, Sebastián-Gallés, & Rodríguez-Fornells, 2006; Cunillera et al., 2009; de Diego Balaguer, Toro, Rodríguez-Fornells, Bachoud-Lévi, & Marcus, 2007; Sanders et al., 2002). Since lexical access is more likely to be time-locked to word onsets than to syllable onsets, we predict larger N400s to syllables that, on the basis of distal prosody, are predicted to be perceived as word-initial.

Method

Participants

Thirty-two participants completed the experiment; 28 participants (15 female; average age: 21.3 years)

contributed data to the reported analyses. All were self-reported right-handed native speakers of English, with normal hearing and normal or corrected-to-normal vision, who were not taking psychoactive medications and were without known neurological deficits. Of the four participants excluded, one was due to a coding error, while the other three were excluded due to the presence of high frequency noise in their raw electroencephalogram (EEG) recording. All participants provided informed consent and received \$10/hour for their participation.

Materials

One hundred and four eight-syllable experimental item sequences were constructed. The first four syllables were always two primary stress-initial words with unambiguous lexical structure (e.g., *banker helpful*). The final four syllables had ambiguous lexical organisation such that they could form either a sequence of three words ending in a monosyllabic word (e.g., *tie murder bee*), or two disyllabic words (e.g., *timer derby*). The syllables were selected so that all possible disyllabic words formed by the eight syllables would have stress on the first syllable. A full list of items can be found in Appendix 1.

All items and a set of fillers were recorded as connected monotone speech by author M.B. Stimuli were recorded onto SONY MiniDisc in a sound-attenuated chamber using a Shure SM10 head-mounted microphone and a Rolls MP13 Mini-Mic preamplifier at a rate of 22 kHz with 24-bit resolution. From these recordings, resynthesised stimuli were created using the pitch-synchronous overlap-and-add (PSOLA) algorithm (Moulines & Charpentier, 1990) as implemented in Praat (Boersma & Weenink, 2002). For filler items, only the pitch was manipulated for each sequence following the method described later. For experimental items, the duration and pitch of the first five syllables were manipulated to induce the percept of either a monosyllabic or disyllabic final word following the method described in Dilley et al.'s (2010; see Figure 1) study.

The disyllabic conditions were created first. The intonation of the first four syllables (i.e., first two words) was manipulated so that each word had a low target pitch on the first syllable and a high target pitch on the second (i.e., rising pitch across the entire word). The fifth syllable was given a flat low pitch. High targets were set at 275 Hz; low targets were set at 175 Hz. High and low points or regions corresponding to targets were connected with straight line interpolations. The final three syllables were resynthesised with high, low and high targets, respectively. The average length of disyllabic stimuli was 4070 ms ($SD = 396$ ms).

The monosyllabic stimuli were resynthesised from the disyllabic manipulations: the acoustic pattern of the final three syllables was identical across conditions. The first

four syllables (i.e., first two words) were manipulated so that each word had a high target pitch on the first syllable and a low target pitch on the second (i.e., a falling pitch across the entire word). In addition, the fifth syllable was synthesised so that the first part of the vowel had a high target pitch and the second part had a low target pitch (i.e., a falling pitch across the syllable). The fifth syllable was also lengthened so that the entire syllable was comparable in length to the average duration of the first two words of the sequence; this was intended to create a perceptual impression of isochrony of the first three words, following the method of Dilley and McAuley (2008). The final three syllables were unchanged, and therefore acoustically identical to those of the disyllabic condition. The average length of the monosyllabic condition was 4304 ms ($SD = 385$ ms). The average amount of lengthening of the fifth syllable was 234 ms ($SD = 131$ ms).

In order to keep participants from generating expectations about the presence or prosodic form of the experimental items, 200 fillers were constructed, consisting of sequences of alternating mono- and disyllabic words with unambiguous lexical structure. There were 50 fillers of 6, 7, 9 and 10 syllables each. For words which were non-final in each filler item, each word of the item incurred a pitch change, regardless of whether it was a monosyllabic or disyllabic word. There were four possible tonal patterns on the final word of each filler item: a sustained high target, a sustained low target, a falling pattern or a rising pattern. Each of the 50 filler items was paired with each of the four possible tonal patterns on the final word, for a total of 200 filler stimuli. The average length of the fillers was 4193 ms ($SD = 825$ ms).

For the purposes of time-locking the ERP waveforms to the onset of the critical words and syllables, two annotators identified the onset of every word of the fillers and items, and the final four syllables of the items. If these annotators did not agree on the location of an onset within 10 ms, author M.B. independently identified the onset. Out of 3243 total onsets, there were 69 onsets (approximately 2%) for which M.B.'s annotations failed to agree with any other annotator's location within 10 ms. Author L.D. annotated these onsets, and, if authors M.B. and L.D. did not agree within 10 ms, the onset was defined as the average of the two authors' annotations. There were 19 such cases (0.6% of onsets).

All experimental items and fillers were normalised to the same maximal intensity, and assigned to two experimental lists. The 104 experimental items were randomly divided in half. The disyllabic conditions from one half were assigned to List 1 and the monosyllabic conditions to List 2. For the second half of items, the monosyllabic conditions were assigned to List 1 and the disyllabic conditions to List 2. In this way, each of the two lists contained 52 items from each of the monosyllabic and disyllabic conditions and all 200 fillers. Participants were

randomly assigned to one list, and lists were presented in a different random order for each participant.

Procedure

Participants were seated in a comfortable chair 150 cm from the computer monitor. A fixation cross appeared at the start of each trial and remained onscreen until the response cue appeared. After 1000 ms, the syllable sequence was played from two loudspeakers on either side of the monitor at an average intensity of 50–60 dBA measured at the location of participants. The response cue ('What was the last word you heard?') appeared 1000 ms after the onset of the final syllable of the sequence and remained onscreen until the participant responded. The participant's response was recorded with an Audio Technica ATR20 cardioid microphone positioned near the participant's head and digitised directly at a sampling rate of 22 kHz with 16-bit resolution.

An experimenter, listening through an intercom in the adjoining room, recorded the participant's response by pressing one of two keys on a button box, corresponding to a monosyllabic or disyllabic word response. After the experimenter responded, the participant saw the cue 'Ready?' and pressed a button to proceed to the next trial. Following the session, the participant's verbal responses were confirmed by a second listener who checked that the responses entered by the experimenter during the session matched the recorded responses. Any errors were corrected before the behavioural results were analysed.

Short breaks were offered approximately every 15 minutes. Participants were encouraged to ask for longer breaks as needed. The 104 items and 200 fillers were presented in random order in six blocks of 50–51 trials each. The entire experimental session took 2–2.5 hours to complete.

Continuous EEG was sampled at 250 Hz and a bandwidth of .01–100 Hz throughout the duration of the experiment from 128-channel HydroCel Geodesic Sensor Nets (Electrical Geodesics Inc., Eugene, OR). Impedance was brought below 50 k Ω at every electrode at the beginning of the experiment and maintained below 100 k Ω for the duration. The continuous EEG was divided into 900 ms epochs, from 100 ms before to 800 ms after the onset of a target.

EEG from individual trials was visually inspected and excluded from analysis if it contained eye blinks, eye movements, or other identifiable artefacts. Data from the remaining trials, regardless of the subjects' behavioural response, were averaged by subject and condition, re-referenced to the average of the two mastoid electrodes and corrected to a 100 ms pre-target baseline. Subjects included in the analysis contributed data from at least 25 out of 52 trials ($M = 40.4$; $SD = 7$) in each condition.

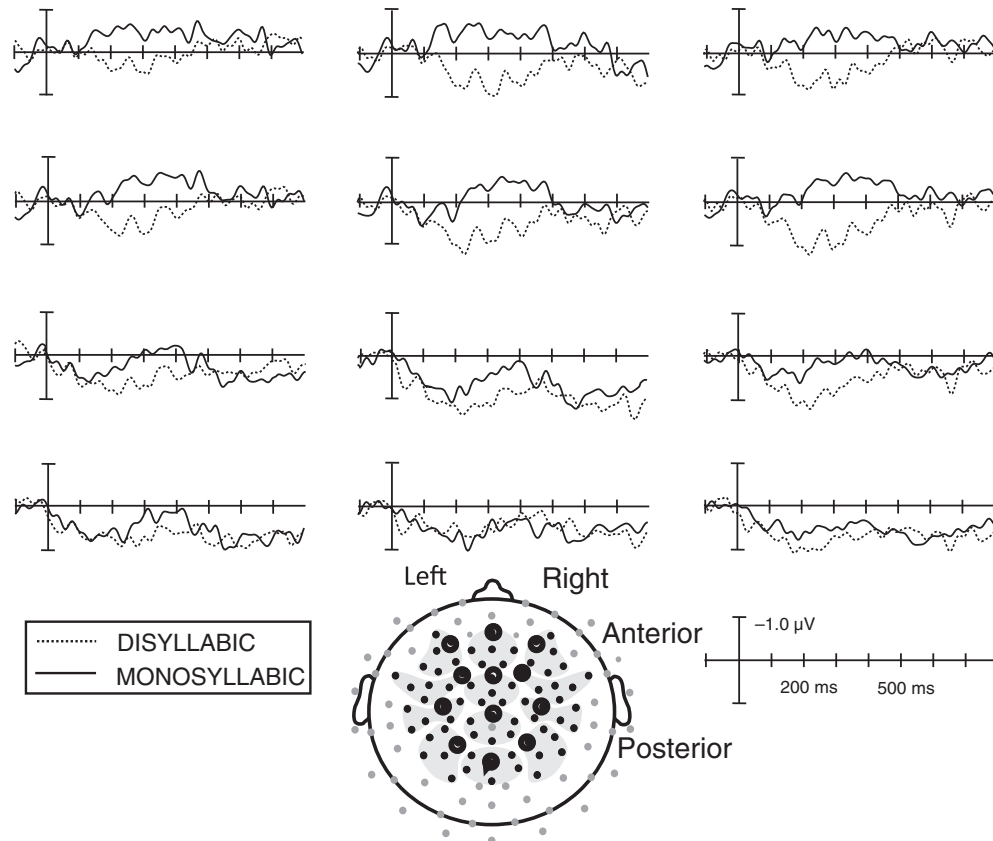


Figure 2. Grand average event-related potentials time-locked to the onset of the penultimate syllables following the monosyllabic (solid line) and disyllabic (dotted line) contexts. Waveforms are shown for the twelve recording sites depicted on the electrode map. They have been low-pass filtered at 30 Hz for presentation purposes.

Analysis

Data from 72 electrodes were divided into 12 groups of 6 electrodes each based on scalp location. Data within each group of six electrodes were averaged and scalp position was treated as two factors in repeated-measures analysis of variances (ANOVAs): Anterior/Posterior position (AP) with four levels (anterior, anterior-central, central and posterior) and Lateral/Medial position (LM) with three levels (left, medial and right). Approximate locations of electrodes are included in the ERP waveform Figures (2–4). Locations of electrode-position factors in the ANOVAs are included in Figures 2 and 4. Mean amplitude measurements were taken at each group of electrodes in four time intervals to test for the hypothesised effects: 60–90 ms (P1), 120–150 ms (N1), 200–500 ms (N400) and 500–800 ms. To avoid effects of differences in pre-stimulus baseline amplitude on the high-frequency, low-amplitude ERP components used to index perceptual processing, mean amplitude in the second time window (120–150 ms) was subtracted from that in the first time window (60–90 ms), resulting in the P1-N1 deflection, similar to measurements taken in

previous ERP segmentation studies (Sanders & Neville, 2003). Average amplitude in each time range was entered into a 2 (Prosodic Context: Monosyllabic vs. Disyllabic) \times 2 (Experimental List) \times 4 (AP position) \times 3 (LM position) repeated-measures ANOVA. Greenhouse-Geisser corrections were applied for comparisons that included factors with more than two levels. All significant main effects and interactions ($p < .05$) were further investigated with post hoc analyses. Only effects and interactions which involve the factor of Prosodic Context will be discussed. Experimental List never interacted with Prosodic Context, so will not be mentioned in the results.

Results

Behavioural

Participants produced the correct word to 97.8% ($SD = 2\%$) of the filler items, demonstrating that they were engaged in the experimental task. The vast majority of incorrect responses were the same number of syllables as the correct word; however, on five occasions, participants produced a three-syllable word which was treated as incorrect. Participants responded that they heard a

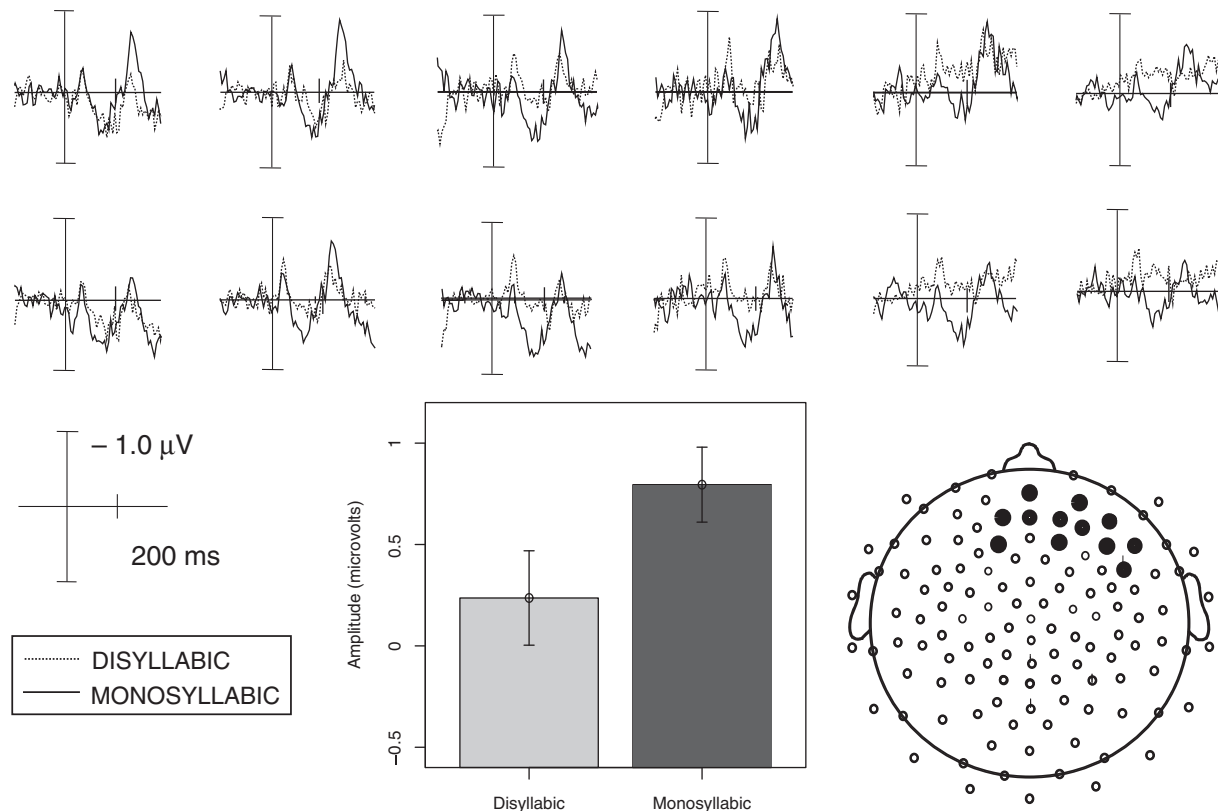


Figure 3. Grand average event-related potentials time-locked to the onset of the final syllables following the monosyllabic (solid line) and disyllabic (dotted line) contexts. Waveforms are shown for the twelve recording sites depicted on the electrode map. The bar graph depicts mean amplitude 60–90 ms after onset (P1) – mean amplitude 120–150 ms after onset (N1) measured over the depicted electrodes. Unlike the previous waveform image, data depicted here were not filtered after analysis so that the smaller amplitude effects over the smaller time window are evident.

disyllabic item in the disyllabic condition 91.7% of the time ($SD = 6\%$); they reported hearing a disyllabic item in the monosyllabic condition 69.0% of the time ($SD = 15\%$), $t_{\text{subjects}}(27) = 2.31$, $p < .05$, $t_{\text{items}}(103) = 8.21$, $p < .001$. Signal detection measures d' and c were used to separate participant's sensitivity to the distal prosody manipulation and any general tendency to make a disyllabic or monosyllabic response, respectively (MacMillan & Creelman, 1991). For this analysis, disyllabic responses following the disyllabic context were hits, while disyllabic responses following the monosyllabic context were false alarms. Values of d' across participants were reliably different from zero ($M = .97$, 95% CI = .77–1.17), confirming that participants were sensitive to the distal prosodic manipulation.

With respect to response criterion, c , participants were found to be more likely to give a disyllabic response than a monosyllabic response ($M = -1.03$, 95% CI = -0.88 to -1.17). The negative value of the criterion shows that, separate from the distal prosody effect, listeners had a general tendency to report a disyllabic final word. Participants in Experiment 1a in Dilley et al.'s (2010) study also showed a disyllabic bias reporting a disyllabic

word 93% of the time when it appeared in a disyllabic context, but also reporting a disyllabic word 35% of the time in a monosyllabic context. However, we observed an even stronger bias in the current experiment, such that listeners reported a disyllabic word 92% of the time in a disyllabic context vs. 69% of the time in a monosyllabic context. This result may be due to the fact that utterances were initially read with a disyllabic interpretation, resulting in an underemphasised final syllable which was less conducive to a monosyllabic parse than materials used in prior work. Moreover, lexical statistics obtained for many of the items demonstrate that the orthographic and phonological neighbourhoods are smaller for words in the disyllabic condition than those of the monosyllabic condition, $t_1(464) = -13.89$, $p < .001$; $t_2(464) = -15.56$, $p < .0001$.¹

Event-related potentials

Penultimate syllable

The predictions of the differential processing accounts can be tested by time-locking ERP waveforms to the penultimate syllable in all of the strings. Recall that the

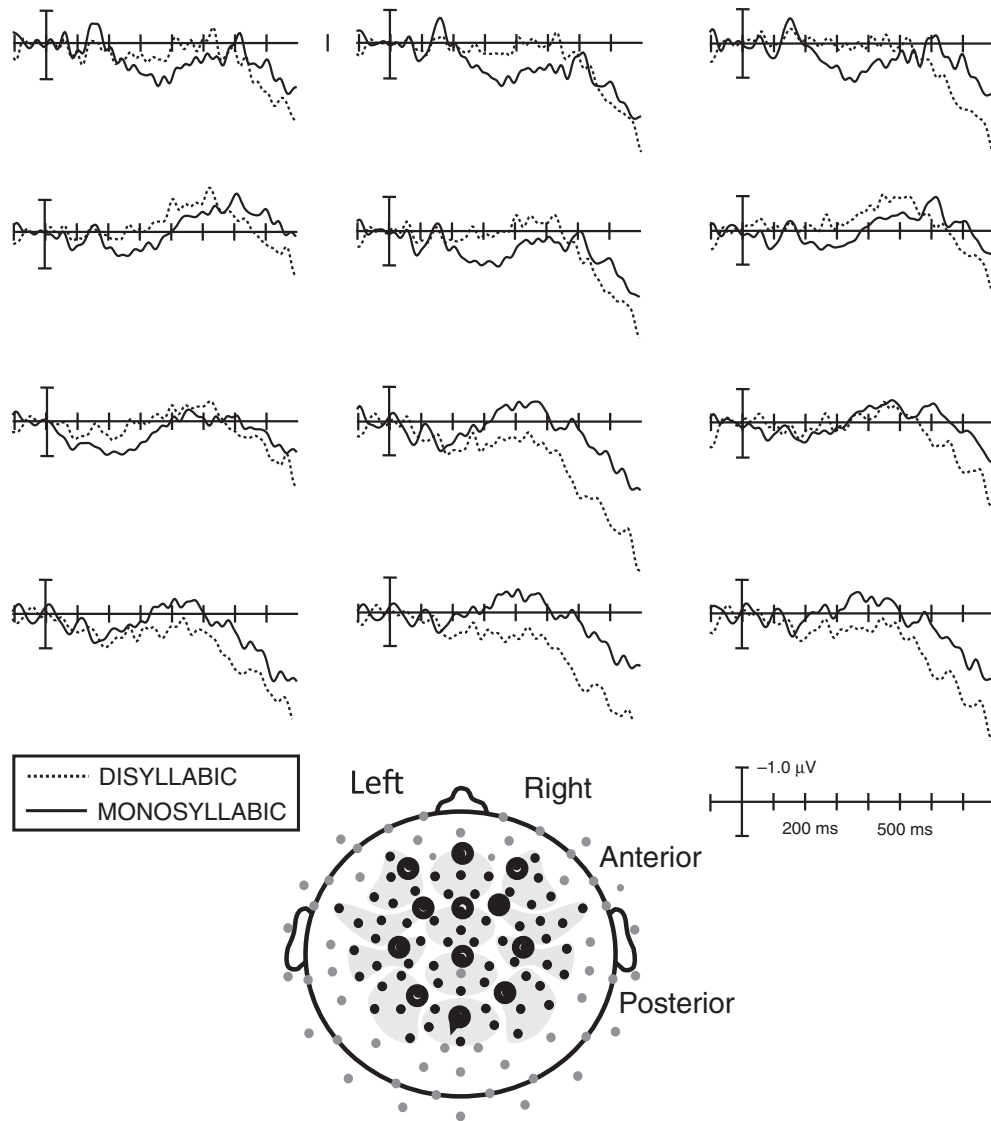


Figure 4. Grand average event-related potentials time-locked to the onset of the final syllables following the monosyllabic (solid line) and disyllabic (dotted line) contexts. Waveforms are shown for the twelve recording sites depicted on the electrode map. They have been low-pass filtered at 30 Hz for presentation purposes.

penultimate syllable (e.g., /dɜː/ in *banker helpful* /tɑɪ mɔː dɜː bi/) was predicted to be perceived as word-initial in the disyllabic condition (e.g., 'der' in *derby*) and word-final in the monosyllabic condition (e.g., 'der' in *murder*). If this pattern of segmenting continuous speech into words affects early perceptual processing, penultimate syllables following the disyllabic context should elicit a larger P1-to-N1 deflection in addition to any later effects reflecting differences in post-perceptual processing.

Mean amplitude P1-N1. Contrary to our prediction that the P1-N1 deflection would be larger when the penultimate syllable was predicted to be perceived as word-initial (e.g., 'der' in *derby*) than word-final (e.g., 'der' in *murder*), we observed very little difference between these

conditions in this early time window, as shown in Figure 2. In fact, the onset of the second positivity in response to penultimate syllables in the disyllabic context had a short enough latency that it overlapped with N1 amplitude resulting in a measured P1-N1 deflection that was smaller than that for the same syllables in the monosyllabic condition. When data from all electrodes were included in analysis, Prosodic Context interacted with AP position, $F(3, 81) = 9.70, p < .005$, such that differences due to distal prosody were larger over anterior regions. At anterior electrodes only, there was a main effect of Prosodic Context, $F(1, 27) = 10.33, p < .005$, such that syllables predicted to be word-final elicited a larger P1-N1 deflection than syllables predicted to be word-initial. Neither the P1 amplitude nor the N1 amplitude computed relative to

the pre-stimulus baseline showed main effects of Prosodic Context, suggesting that the difference between these two mean amplitude measures accurately reflects the effects of distal prosody on early perceptual processing.

Mean amplitude 200–500 ms. In this middle time window, the penultimate syllables elicited a larger positivity when predicted to be word-initial compared to word-final as shown in Figure 2. When data from all electrodes were included in analysis, Prosodic Context interacted with both AP and LM electrode position, $F(6, 162) = 4.15, p < .01$. Over anterior electrodes only, there was a main effect of Prosodic Context, $F(1, 27) = 9.23, p < .01$, such that the amplitude was more positive for syllables predicted to be word-initial than for syllables predicted to be word-final.

Mean amplitude 500–800 ms. In this late time window, there were no amplitude differences based on Prosodic Context, $F's < 1$.

Final syllable

The predictions of the differential processing accounts can also be tested by time-locking ERP waveforms to the final syllable in all of the strings. Recall that the final syllable (e.g., /bi/ in *banker helpful* /taɪ mæ dʌ bi/) was predicted to be perceived as word-initial in the monosyllabic condition (e.g., 'bee') but as word-final in the disyllabic condition (e.g., 'by' in *derby*). If this pattern of segmenting continuous speech into words affects early perceptual processing, final syllables following the monosyllabic context should elicit a larger P1-to-N1 deflection in addition to any later effects reflecting differences in post-perceptual processing.

Mean amplitude P1-N1. As predicted, the final syllable in streams elicited a larger P1-N1 deflection when predicted to be word-initial compared to word-final, as shown in Figure 3. When data from all electrodes were included in analysis, Prosodic Context interacted with AP position, $F(3, 81) = 5.58, p < .05$ and with LM position, $F(2, 54) = 3.80, p < .05$. For anterior electrodes only, Prosodic Context still interacted with LM position, $F(2, 54) = 7.77, p < .005$, with larger effects of Prosodic Context over right anterior locations. At these electrode locations, the P1-N1 deflection was larger for predicted word-initial syllables than predicted word-final syllables, $F(1, 27) = 4.28, p < .05$. Again, neither P1 nor N1 amplitude alone relative to the pre-stimulus baseline differed by context.

Mean amplitude 200–500 ms. There are two effects of interest in this middle time window. At anterior electrodes, the final syllables elicited a larger positivity when they were predicted to be word-initial than when predicted to be word-final, as shown in Figure 4. Over posterior

electrodes, this effect reversed, such that syllables predicted to be word-initial elicited a larger negativity than syllables predicted to be word-final. When data from all electrodes were included in the analysis, Prosodic Context interacted with AP and LM positions, $F(6, 162) = 5.35, p < .001$, an effect which was driven by competing effects in anterior and posterior electrode positions. Over anterior regions, syllables predicted to be word-initial elicited a larger positivity than syllables predicted to be word-final, $F(1, 27) = 4.46, p < .05$. Over posterior regions, Prosodic Context interacted with LM position, $F(2, 54) = 10.423, p < .0005$, such that the effects of distal prosody were larger over posterior central electrodes, where predicted word onsets elicited a larger negativity than predicted word-final syllables, $F(1, 27) = 9.09, p < .01$.

Mean amplitude 500–800 ms. In this late time window, the final syllables elicited a positivity which varied in amplitude depending on distal prosody over posterior, medial and right electrodes, as shown in Figure 4. When data from all electrodes were included in analysis, Prosodic Context interacted with AP and LM electrode position, $F(6, 162) = 2.95, p < .05$. For posterior electrodes only, Prosodic Context still interacted with LM position, $F(2, 54) = 5.46, p < .01$. These interactions were driven by a main effect of Prosodic Context over posterior, and medial and right electrode positions, $F(1, 27) = 13.19, p < .005$, such that final syllables in the disyllabic condition predicted to be word-final elicited a larger positivity than final syllables in the monosyllabic condition predicted to be word-final. The timing and distribution of this effect is similar to the Closure Positive Shift (CPS) (Steinhauer, 2003), a component observed at perceived phrase boundaries.

Discussion

The current experiment was designed to investigate whether distal prosody influences early perceptual processing of word boundaries in addition to later competition between potential lexical items. We recorded EEG while participants listened to syllable strings with ambiguous word boundaries (e.g., /taɪ mæ dʌ bi/) which were embedded in prosodic contexts designed to induce perception of the final syllable as a word onset (e.g., *bee*) or a word-final syllable (e.g., *derBY*). Behavioural results demonstrate that the manipulation successfully induced listeners to differentially report hearing final monosyllabic or disyllabic words, respectively. Moreover, and critically, ERP results demonstrate early differences in the processing of syllables at the ends of experimental sequences depending on whether they were reported as word-initial vs. word-final.

The behavioural results from the current experiment replicate those of Dille et al. (2010). Participants were

more likely to report hearing a disyllabic word in a disyllabic context than a monosyllabic one. These results demonstrate the validity of experimental manipulation and provide further support for the perceptual grouping hypothesis originally proposed by Dilley and McAuley (2008).

Finding significant differences in the ERP waveforms averaged across behavioural responses potentially provides additional support for the idea that segmenting continuous speech into words affects early perceptual processing. That is, even though the monosyllabic context resulted in monosyllabic verbal responses on only 31% of trials, differences in the ERP waveforms averaged across all trials with the same prosodic context suggest the effects of distal prosody were consistent when investigated with online processing measures. In fact, the ambiguous syllables may initially have been grouped in a manner that was determined by distal prosody. However, before participants gave a verbal response, at least 1000 ms after the final syllable onset, they may have retroactively resegmented the stream based on analysis of the acoustic information provided at the very end, resulting in a disyllabic bias.

In turning to the ERP results, it is important to note that the acoustic information was identical across prosodic conditions only for the final three syllables. Particularly of note is the fact that in the disyllabic condition, the first ambiguous syllable (e.g., /tai/) is shorter than the first ambiguous syllable of the monosyllabic condition (see Figure 1). These durational differences could affect the auditory system's response to subsequent ambiguous syllables as follows: Anytime the auditory system responds to a sound, it must recover from that response before responding with the same intensity again. This recovery time is called a refractory period. Sounds entering the auditory system during the refractory period elicit smaller responses than sounds entering after recovery (Bess and Ruhm, 1972; Budd, Barry, Gordon, Rennie, & Michie, 1998; Coch, Skendzel, & Neville, 2005). In the disyllabic condition, the auditory system's response to the antepenultimate and penultimate syllables (e.g., /mæ/ and /dæ/, respectively) may be smaller as these syllables occur earlier in the refractory period of the first ambiguous syllable than they do in the monosyllabic condition. For this reason, we are not surprised that we observed no effect of distal prosody on the antepenultimate syllable, and that the observed effects were reduced for the penultimate syllable as compared to the final one. Moreover, as we observed only one significant effect of the distal prosody manipulation on the penultimate syllable, which was similar in timing and morphology to an effect on the final syllable, we will discuss those effects together below. Before doing so, however, we first address the earliest effect of distal prosody, which we observed on the

final syllable of the stream, which is least likely to have been influenced by refractory effects.

There were several important effects on the final syllable of the ambiguous streams, occurring in early, middle, and late time windows. First, we observed a larger P1-to-N1 deflection to final syllables (e.g., /bi/) when they were predicted to be perceived as word-initial rather than as word-final. This early ERP difference demonstrates that distal prosody has an online effect on word segmentation such that distal prosody serves to perceptually group ambiguous syllables into words, thereby allowing listeners to predict whether an upcoming syllable is a word onset.

The early ERP effect is consistent with previous studies showing a larger N1 or P1-to-N1 deflection in response to word onsets in continuous speech (Astheimer & Sanders, 2009, 2011; Sanders et al., 2002; Sanders & Neville, 2003). Sanders et al. (2002) observed a larger P1-to-N1 deflection to sequence-initial syllables after listeners had been trained to recognise the words in an artificial language. Relatedly, Sanders and Neville (2003) observed larger N1s to English syllables which were word-initial. Finally, Astheimer and Sanders (2009) observed that auditory probes coinciding with word onsets elicited larger N1s than probes presented either 100 ms before onsets or at random times. In all of these cases, the authors observed larger N1s or P1-to-N1 deflections to syllables immediately following a segmented portion of the stream. The current results demonstrate that by the onset of the final syllable in the monosyllabic condition, listeners had determined that the penultimate syllable was the end of one word, and that the final syllable was beginning of another, thereby demonstrating online segmentation of the ambiguous stream. Critically, whereas previous studies have looked at N1 differences across acoustically non-identical sounds in natural speech (Astheimer & Sanders, 2009; Sanders & Neville, 2003), or acoustically identical sounds in an artificial language (Astheimer & Sanders, 2011; Sanders et al., 2002), these data are the first demonstration of P1-to-N1 differences for acoustically identical natural language stimuli.

The P1-to-N1 deflection observed here is not as large or broadly distributed across the scalp as that reported by others, for at least two likely reasons. First, subjects contributed fewer trials to each condition ($M = 40.4$) than in previous studies, thereby decreasing the signal-to-noise ratio. For example, Astheimer and Sanders (2009) report that only subjects who contributed 60 trials or more to each condition were included in analysis. Second, the current stimuli contained a variety of syllable onsets. Ideally, all of the onsets would be stop consonants, which elicit the largest and most temporally consistent onset components. However, the difficulty of generating stimuli which satisfied the experimental constraints meant that it was impossible to limit the syllable onsets to stop consonants. The resulting variation in phonemes used to

drive the auditory onset components lead to increased variability in P1 and N1 amplitudes across trials, and, therefore, an overall decrease in P1-to-N1 amplitude.

The similarity of the results of the current study to those observed previously is consistent with the view that P1-to-N1 amplitude reflects temporally selective attention (Astheimer & Sanders, 2009) that is guided in the present case by temporal expectations induced by distal prosody. N1 effects have been observed in response to a variety of segmentation cues: statistical regularities, lexical stress, recognition of newly learned nonsense words, acoustic cues in unfamiliar speech designed to sound like a native language, and now, distal prosody. The timing with which these different segmentation cues become available is likely to differ depending on the type of cue; however, the timing and distribution of the ERP effects they elicit does not. Specifically, the timing of the earliest ERP waveform differences across studies, including the present one, is similar, beginning around 100 ms after sound onset. Therefore, it is likely that the P1-to-N1 deflection effects are not directly reflecting the process of segmentation itself but of temporally selective attention to word onsets since these segments are particularly important for understanding speech.

In addition to the differences in the P1-to-N1 deflection observed on the final syllable of the stream, we also observed a significant difference in this response for the penultimate syllable. Specifically, in this time window, the penultimate syllable elicited a smaller deflection when it was predicted to be word-initial (e.g., /dæ/ in *derby*) than when it was predicted to be word-final (e.g., /dæ/ in *murder*). However, we do not interpret this result as reflecting P1-to-N1 differences on the penultimate syllable, but rather as the result of temporal overlap with the positivity observed in the subsequent time window for syllables predicted to be word-initial.

The second effect we observed was a larger positivity 200 to 500 ms after syllable onset over anterior regions in response to syllables predicted to be word-initial rather than word-final. This effect was evident for both the penultimate and final ambiguous syllables and may reflect the fact that word onsets in English are perceptually stressed. Cunillera, Gomila, and Rodríguez-Fornells (2008), for example, reported larger amplitudes in the P2 time window (between 120 and 420 ms) for stressed syllables compared to unstressed syllables, consistent with our observation of a larger positivity for word onsets. In our study, the third ambiguous syllable was more likely to be perceived as stressed in the disyllabic condition, when it was a word onset (/dæ/ in *derby*). On the other hand, the final syllable was more likely to be perceived as stressed in the monosyllabic condition, as this was the case where it was a word onset (*bee*).

In the same 200–500 ms time window in which we observed a larger positivity for word-initial syllables at

anterior electrode positions, we observed the opposite effect over posterior central regions, where final syllables elicited a larger negativity when predicted to be word-initial. This effect is similar in latency and distribution to the N400 observed in prior segmentation studies (Abla et al., 2008; Cunillera et al., 2006, 2009; de Diego Balaguer et al., 2007; Sanders et al., 2002), suggesting that lexical access was likely time-locked to the onset of the final syllable when it was a word onset. We therefore interpret this effect as additional evidence that listeners were interpreting the final syllable of the monosyllabic streams as a complete word.

Finally, we observed a larger posterior positivity on the final syllables which were predicted to be perceived as word-final (e.g., /bi/ in *derby*). We interpret this effect as an example of the CPS, a positive-going wave observed at the location of phrase boundaries (Steinhauer, 2003). The late positivity that we observed was not necessarily larger in one condition, but was rather *earlier* in the disyllabic condition than the monosyllabic condition. Dilley and McAuley (2008) argue that distal prosody gives rise to the perception of word boundaries in part because it induces listeners to impose prosodic boundaries on the ambiguous syllable. If listeners in the current study perceived phrase boundaries at word boundaries, than we could expect an earlier CPS to the disyllabic items than to the monosyllabic items. Specifically, in the disyllabic condition, every rising syllable marks a word boundary; therefore, the listener is induced to perceive /bi/ in the disyllabic context as a phrase-final syllable. In contrast, rising syllables are predicted to be perceived as word-initial syllables in the monosyllabic condition. When /bi/ occurs as the final syllable of the monosyllabic condition, it is perceived as a word onset, and, therefore, *not* as phrase-final, at least not initially. However, we would expect to see a CPS *later* for the monosyllabic conditions, when the listener perceives the boundary signalled by the end of the speech stream. The presence of an earlier CPS in the disyllabic than the monosyllabic conditions provides more evidence for Dilley and colleagues' perceptual grouping hypothesis.

Conclusion

The current results are important for understanding the role of prosody in speech segmentation. Specifically, they demonstrate that distal prosody affects early perceptual processing of word boundaries in addition to later lexical processing and interpretations of what was heard when listeners are prompted to give explicit reports. The latency of the early ERP effects suggests that listeners predict word onsets given a supportive prosodic context rather than waiting for all potential segmentation cues. These data also help inform our understanding of what is being measured by the amplitude of the P1-to-N1 deflection, as they are the first demonstration of differences in auditory

evoked potentials for acoustically identical untrained natural language stimuli. As such, the results lend further support to the claim that the amplitude of the P1-to-N1 deflection during speech processing indexes a more general attentional process rather than segmentation itself. Finally, they demonstrate the effectiveness of using ERPs to investigate the stages of processing that are influenced by prosodic manipulations.

Acknowledgements

This material is based upon work supported by the National Science Foundation [grant number BCS-0847653] awarded to L.C.D. and was partially funded by a National Institutes of Health Institutional Training Grant [grant number MH16745] which provided post-doctoral training for M.B., and a John Merck Fellowship in the Biology of Developmental Disabilities awarded to L.D.S. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the NIH, or the John Merck Foundation. We thank Paul Dainton, Ashley Fitzroy, Martin Nolet and Anton Zakashansky for help with stimulus preparation, and Paul Dainton, Brian Keane and Chase Langerap for help with data collection and analysis.

Note

1. Data obtained from the English Lexicon Project Web Site (Balota et al., 2007).

References

- Abla, D., Katahira, K., & Okanoya, K. (2008). Online assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*, 952–964. doi:10.1111/j.1469-8986.1987.tb00324.x
- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439. doi:10.1006/jmla.1997.2558
- Altmann, G. T. M. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, *137*(2), 190–200.
- Astheimer, L. B., & Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology*, *80*(1), 23–34. doi:10.1016/j.biopsycho.2008.01.015
- Astheimer, L. B. & Sanders, L. D. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia*, *49*, 3512–3516. doi:10.1016/j.neuropsychologia.2011.08.014
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459. doi:10.3758/BF03193014
- Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht: Foris.
- Bess, J., & Ruhm, H. (1972). Recovery cycle of the acoustically evoked potential. *Journal of Speech, Language, and Hearing Research*, *15*, 507–517.
- Boersma, P. & Weenink, D. (2002). *Praat, a system for doing phonetics by computer*. Software and manual Retrieved from <http://www.praat.org>.
- Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review*, *18*, 1189–1196. doi:10.3758/s13423-011-0167-9
- Budd, T. W., Barry, R. J., Gordon, E., Rennie, C., & Michie, P. T. (1998). Decrement of the N1 auditory event-related potential with stimulus repetition: Habituation vs. refractoriness. *International Journal of Psychophysiology*, *31*(1), 51–68. doi:10.1016/S0167-8760(98)00040-3
- Coch, D., Skendzel, W., & Neville, H. J. (2005). Auditory and visual refractory period effects in children and adults: An ERP study. *Clinical Neurophysiology*, *116*, 2184–2203. doi:10.1016/j.clinph.2005.06.005
- Cole, R. A., & Jakimik, J. (1980). Segmenting speech into words. *Journal of the Acoustical Society of America*, *67*, 1323–1332. doi:10.1121/1.384185
- Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, *32*, 193–210. doi:10.1006/jmla.1993.1011
- Cunillera, T., Càmara, E., Toro, J. M., Marco-Pallares, J., Sebastián-Galles, N., Ortiz, H., ... Rodríguez-Fornells, A. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage*, *48*, 541–553. doi:10.1016/j.neuroimage.2009.06.069
- Cunillera, T., Gomila, A., & Rodríguez-Fornells, A. (2008). Beneficial effects of word final stress in segmenting a new language: Evidence from ERPs. *BMC Neuroscience*, *9*(1), 23. doi:10.1186/1471-2202-9-23
- Cunillera, T., Toro, J. M., Sebastián-Gallés, N., Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research*, *1123*, 168–178. doi:10.1016/j.brainres.2006.09.046
- Cutler, A. & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*, 218–236. doi:10.1016/0749-596X(92)90012-M
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121. doi:10.1037/0096-1523.14.1.113
- de Diego Balaguer, R., Toro, J. M., Rodríguez-Fornells, A., Bachoud-Lévi, A.-C., & Marcus, G. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE*, *2*, e1175. doi:10.1371/journal.pone.0001175.t004
- Dilley, L. C., Mattys, S. L. & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*, 274–294. doi:10.1016/j.jml.2010.06.003
- Dilley, L. C. & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*, 294–311. doi:10.1016/j.jml.2008.06.006
- Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*, 1664–1670. doi:10.1177/0956797610384743

- Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27, 765–768. doi:10.1121/1.1908022
- Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, MA: MIT Press.
- Hansen, J. C., Dickstein, P. W., Berka, C., Hillyard, S. A. (1983). Event-related potentials during selective attention to speech sounds. *Biological Psychology*, 16, 211–224. doi:10.1016/0301-0511(83)90025-X
- Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, 182, 177–180. doi:10.1126/science.182.4108.177
- Kutas, M., & Federmeier, K. D., (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. doi:10.1146/annurev.psych.093008.131123
- Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, 51, 2018–2024. doi:10.1121/1.1913062
- MacMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
- Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception & Performance*, 15, 576–585. doi:10.1037/0096-1523.15.3.576
- Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception and Psychophysics*, 53, 372–380. doi:10.3758/BF03206780
- Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465–494. doi:10.1006/cogp.1999.0721
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 453–467. doi:10.1016/0167-6393(90)90021-Z
- Näätänen, R. & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, 24, 375–425. doi:10.1111/j.1469-8986.1987.tb00311.x
- Nakatani, L. H. & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *The Journal of the Acoustical Society of America*, 62, 714–719. doi:10.1121/1.381583
- Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34, 191–243. doi:10.1006/cogp.1997.0671
- Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech*, 54, 147–165. doi:10.1177/0023830910397489
- Reinisch, E., Jesse, A., & McQueen, J. M. (2011b). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 978–996. doi:10.1037/a0021923
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928. doi:10.1126/science.274.5294.1926
- Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90(1), 51–89. doi:10.1016/S0010-0277(03)00139-2
- Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, 105, 466–476. doi:10.1016/j.cognition.2006.10.008
- Sanders, L. D., Ameral, V., & Sayles, K. (2009). Event-related potentials index segmentation of nonsense sounds. *Neuropsychologia*, 47, 1183–1186. doi:10.1016/j.neuropsychologia.2008.11.005
- Sanders, L. D., & Neville, H. (2003). An ERP study of continuous speech processing. I. Segmentation, semantics, and syntax in native speakers. *Cognitive Brain Research*, 15, 228–240. doi:10.1016/S0926-6410(02)00195-7
- Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience*, 5, 700–703. doi:10.1038/nn873
- Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language*, 86(1), 142–164. doi:10.1016/S0093-934X(02)00542-4
- Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1), 98–108. doi:10.1037/0096-1523.21.1.98

Appendix 1

Items in italics were adapted from Dilley et al. (2010).

trailing greeted (aero mentor/air roman tour)
feather onion (baby curfew/bay beaker few)
 Venus pollen (backhoe bovine/back hobo vine)
 cryptic taken (Bangor maple/bang gourmet pull)
 kettle heaven (barber oboe/bar burrow bow)
pebble dollar (barley virtue/bar lever chew)
 flatly caution (basic peso/bay sickpay so)
 program perish (blister boasting/bliss turbo sting)
 golfers wanted (bluebell jumbo/blue Belgium bow)
 rises Melbourne (Bombay cyclic/bomb basic click)
lady jacket (brandy sultry/bran diesel tree)
 banner Arthur (cancan deejay/can candy jay)
 loser micron (cargo furlough/car gopher low)
 cabinet swivel (catnap purple/cat napper pull)
hero vacuum (cellar legal/cell early gull)
lumpy danger (cherry gurney/chair eager knee)
mixture pleasure (classy depose/class seedy pose)
 slipper stony (closeout lawful/close outlaw full)
 cannon wedding (coffee murky/cough femur key)
 simplex Steelers (conjure neon/con journey yawn)
 grooming studies (contour sewing/con torso wing)
magnet guilty (crisis turnip/cry sister nip)
 busted inning (daisy roasting/day zero sting)
 module forfeit (dancehall locate/dance hollow Kate)
 unions revel (deadpan zero/dead pansy row)
 ponder anchor (Diane exile/die annex isle)
 budding lobster (diaper doing/die Purdue wing)
 lacy mention (dilate textile/die latex tile)
 xerox prelude (dingy nomad/din genome mad)
 gremlin pending (dinner wintry/din Irwin tree)
 droplet butler (dog-ear rebate/dog eerie bait)
 habit hiring (doorbell freedom/door belfry dumb)
 ration forceful (dovetail spindle/dove tailspin dull)
 ringer typing (downplay boycott/down Playdoh cot)
 tourist robin (drama steeply/draw musty plea)
 ladder acne (duty Baghdad/duo teabag dad)
 jackel local (easel truffle/ease ultra full)
 Pavlov gallons (fairly content/fair recon tent)
 Haiti peasant (fancy solid/fan seesaw lid)
horses kayak (fancy munchies/fan seaman cheese)
 sandwich rosy (fanfare resource/fan fairy source)
rushes statutes (Fargo ferment/far gopher meant)
 nature lazy (flatland filtered/flat landfill turd)
nature lazy (foamy detour/foe meaty tour)
 quicken jaguar (freebie kindred/free beacon dread)
 Stacey lowly (freedom peanut/free dumpy nut)
 values tactful (fungi robot/fun gyro bought)
 traffic yielded (furrow dermis/fur odor miss)
 comment sample (gangster notion/gang Sterno shun)
 mixture campers (gatepost cardinal/gate postcard null)
 tainted copper (glassy gullet/glass eagle let)

nicely equal (gravy toaster/gray veto stir)
 shadows prison (grocer custard/grow circus turd)
 gypsy abbot (hairdo inkling/hair doing cling)
 platter catchy (hammer during/ham murder ring)
 happy northern (hamster number/ham sternum burr)
 fussy conscience (handbag erect/hand baggie wrecked)
 schedule testing (handstand bible/hand standby bull)
lender dentist (harem burlap/hair ember lap)
 Taroh cunning (hearsay burping/hear saber ping)
 pleading packers (highchair research/high cherry search)
 magic notice (hipster lingo/hip sterling go)
 pager nanny (howdy cadence/how decay dense)
 bullet junior (iffy dinky/if feeding key)
 locate slaughter (income phoenix/in comfy nicks)
 quotas nicest (kneehigh jackson/knee highjack son)
angry index (labor defense/lay birdie fence)
 fathom dragon (leanto candor/lean toucan door)
trouble wealthy (limber nursing/limb burner sing)
 plaster clusters (maybe feeding/may beefy ding)
shortly polar (maybe negro/may beanie grow)
 wrinkles mallard (mistress passport/miss trespass port)
 poster laces (mohair retail/mow hairy tail)
 llama busy (moron corpus/more encore puss)
 theorist slalom (mountain Dexter/mount index stir)
 wrapper hammock (mustang girdle/must anger dull)
 Navy ripples (oatbran detail/oat brandy tail)
 worthy Russia (obese trophy/owe bistro fee)
 goofy carry (outplay domain/out Playdoh main)
gossip oyster (pantry decoy/pan treaty coy)
 stomach rubbish (paper Sunday/pay person day)
harmful tickled (pecan termite/pea cantor mite)
herbal belly (pigsty polo/pigs typo low)
 glory lawful (prairie venue/prayer even you)
 Yorkie duplex (rainy thirty/rain ether tee)
 easement cabby (rancid needing/ran Sydney ding)
chapter elbow (ruby virgin/rue beaver gin)
 surely winded (sawmill duo/saw mildew owe)
 pastures Presley (scarecrow borrow/scare crowbar row)
 earthen kindness (schoolbus ulcer/school bustle sir)
 ramming feudal (sinker veto/sin curvy tow)
 blanket mounted (slammer scenic/slam mercy nick)
 hazard vacant (therefore mermaid/there former maid)
 panic nomad (therefore rayon/there foray yawn)
gazing hackers (timber ozone/Tim burrow zone)
banker helpful (timer derby/tie murder bee)
center northern (toucan surplus/two cancer plus)
plenty fluid (traitor decrease/tray dirty crease)
 Lisbon partial (tutu nothing/two tuna thing)
 whither chamber (useless orbit/use lessor bit)
 racking tango (welcome female/well comfy male)
 sanction straighten (wherefore castrate/where forecast rate)
 soaking Susan (willow verses/will over says)
fortune decade (windy perfumes/win deeper fumes)