Online methods for the investigation of prosody

1. Introduction

Investigating prosody in natural language poses special challenges. In English, phrasing, stress, and the placement and nature of pitch accents interact in complex ways to produce effects that are hard to deny but often difficult to describe in formal or even informal terms. The challenges extend to the laboratory: the type of information prosody conveys is as difficult to manipulate experimentally as it is to define formally. In an episode of the sitcom Seinfeld, the difficulty of making meta-linguistic judgments about accenting is the vehicle for a joke. The main character Jerry wants to know if he has been invited to a party, so he asks his friend Elaine to ask the host “Should Jerry bring anything?” in the hope that the answer will clarify the hosts’ intentions:

(1)
ELAINE: Well, I talked to Tim Whatley...
JERRY: Yeah...
ELAINE: And I asked him, “Should Jerry bring anything?”
JERRY: So...?
ELAINE: Mmmm...and he said, "Why would Jerry bring anything?"
JERRY: Alright, but let me ask you this question.
ELAINE: What?
JERRY: Which word did he emphasize? Did he say, "Why would JERRY bring anything?" or, "Why would Jerry BRING anything?" Did he emphasize "Jerry" or "bring."
ELAINE (confused): I think he emphasized, "would."

The fact that Jerry takes Tim Whatley’s prosody to be a clue to his expectations is consistent with the widely-held view that phrasing, stress, and accenting patterns reflect aspects of discourse structure (givenness and novelty, e.g.; there is less agreement on particulars) and participant beliefs about the status of uttered content in the discourse. Elaine’s response illustrates both the difficulty of obtaining

---

1 This project was supported by NIH grants HD27206 and DC-005071. The first author was supported by NSF grant SES-0208484.
judgments about prosodic categories directly and the fact that multiple patterns may be appropriate in a given context. Given the difficulty of pinning down prosodic categories, one would ideally like to avoid relying on judgments of appropriateness/inappropriateness or similar measures and instead design experimental tasks that can provide an implicit measure related to interpretation. An ideal measure would be usable in both comprehension and production tasks. Another requirement, given the integral role of context in any characterization of prosody, is that the methodology be both specific enough to operationalize a specific kind of context (e.g., implementing a systematic distinction between ‘given’ and ‘new’ discourse entities so that references to each can be examined), and flexible enough so that the context can be adjusted to evaluate alternative hypotheses. In this paper we suggest the eye-tracking paradigm has the potential to meet these criteria. Within this paradigm, saccadic eye-movements are measured as people generate or listen to spoken utterances and perform simple tasks in a constrained visual display. We give an overview of eye-tracking methodology in Section 2 and applications to the study of pitch accents in particular in Section 3. In Section 4, we illustrate how the method can be applied to timing issues in use of prosody in on-line comprehension.

In addition to methodological issues, there are challenges to experimental work in prosody that stem from uncertainty at a fundamental level about the nature of mapping between phonological categories and semantic/pragmatic ones. To illustrate these challenges and to serve as a lead-in to work discussed in Section 3, we will focus the discussion on particular set of problems involving pitch accent types in English, with a brief review of previous work.

Consider the sentences in (2) and (3) and their corresponding pitch tracks in Figure 1 and 2 (all caps denote an accent).

(2)  
  a. What does Mike enjoy?  
  b. Mike enjoys SCOTCH.

(3)  
  a. Does Mike enjoy beer?  
  b. Mike enjoys SCOTCH.
Figure 1: The F0 track for the word “scotch” with a presentational accent.

Figure 2: The F0 track for the word “scotch” with a contrastive accent.
Intuitively, there seems to be a difference in the accents on SCOTCH in the two examples. A look at the fundamental frequency reveals that the second example has a higher pitch range, a steeper rise in the F0, and a small dip before the rise. In addition, the accent in (3) evokes a sense of contrast that is missing from (2). One gets the impression that Mike enjoys scotch, and does not enjoy beer. For the purposes of this chapter, we will refer to the accent used in (2) as the “presentational” accent and the accent used in (3) as the “contrastive” accent without making a commitment as to whether these are in fact the appropriate linguistic categories.

The differences between the accents in (2) and (3) have been represented in terms of distinct phonological categories, H* and L+H*, by Pierrehumbert & Hirschberg (1990) and the subsequent instantiation of her system in ToBI (Silverman et al., 1992). In addition, Pierrehumbert and Hirschberg (1990) propose that there is a meaning difference that corresponds with the phonological difference. They suggest that H* is used to instantiate new information into the discourse model, whereas L+H* is used to instantiate information in the discourse model over some salient alternative, an idea at the core of the notion of contrastiveness. We should note that there are complexities associated with both assigning interpretations to these pitch accents and to assuming they correspond to distinct phonological categories. On the semantic side, it is not clear how the notion of contrastiveness is to be differentiated from the more general notion of focus as invoking a set of salient alternatives (Rooth, 1985). Moreover, any account of pragmatic reasoning about discourse choices involves evaluating choices against sets of plausible alternatives. On the phonological side, even among those who posit a contrastive category, there is controversy over exactly how to define its acoustic correlates.

Efforts have been made to address these issues experimentally. A common approach for perceptual studies of intonational meaning is to posit the existence of two or more intonational categories, corresponding to a hypothesized meaning/function difference, and present subjects with the stimuli together with a description of discourse contexts for the target utterances. Subjects are then asked directly to make a judgment about the interpretation and/or suitability of the utterance in the context. For example, Bartels and Kingston (1994), investigating acoustic cues to contrastive focus, varied both properties of the pitch contour (peak height, alignment, etc.) and the discourse context in which sentences were presented; listeners were asked for categorical ‘contrastive’ vs. ‘non-contrastive’ judgments. The results provided little evidence for the categorical distinction between L+H* and H* posited by Pierrehumbert and Hirschberg. In a series of perceptual studies on Dutch intonational patterns, Caspers and colleagues have
obtained acceptability ratings for a variety of patterns in described discourse contexts, using hypotheses about meaning from the linguistics literature to construct examples and contexts (Caspers, Van Heuven, & Zwol, 1998; Caspers, 1998; Caspers, 2000). Krahmer and Swerts (2001) asked naïve listeners to evaluate the relative prominence of descriptions in a card-sorting task to investigate differences between contrastive and presentational accents. Accents on words that were produced in contexts that had the potential for contrast were judged to be more prominent. Similarly, in a categorical perception task, Ladd and Morton (1997) presented subjects with words that varied in levels of pitch excursion, and they were asked to perform an identification task (normal vs. emphatic) and a discrimination task. The results from the identification task were consistent with categorical perception, but there were no peaks in the discrimination tasks, suggesting that the interpretation of these accents may be categorical but their perception may be continuous. This study highlights an intrinsic limitation to focusing exclusively on acoustic/phonetic intuitions. While perceptual judgments provide useful information about acoustic-phonetic processing, they do not tell us how these judgments map on to interpretation. In general these perceptual experiments show that asking subjects directly for judgments about intonational patterns can, with careful design, yield useful results. Nonetheless, such approaches by their nature introduce a layer between interpretation and response by drawing attention to the phenomenon under investigation and requiring the subject to process descriptions of the context and the interpretive choices available.

A second strategy has been to study the presentational and contrastive accents in language production (Krahmer & Swerts, 2001; e.g. Selkirk, 2002). Participants are asked to describe a scene or read a script where a critical word is likely to be contrasted with another word in the discourse. The goal is to see how accents produced in the contrast context differ from those produced in the presentational context. This work has been useful in exploring the conditions under which accents are produced, and in general, these studies have demonstrated that speakers are more likely to produce an accent that sounds contrastive when the experimenter creates a context that highlights a potential contrast. However, a consistent result is that pitch accents are not in complementary distribution. For example, even the most contrastive contexts will elicit a high proportion of accents that look like Figure 1. The researcher is forced to decide a priori whether to group all accents that are produced in this context together into a single functionally defined category or to categorize the accents into contrastive and non-contrastive categories based on some acoustic or perceptual criteria. These theoretical
commitments are problematic for investigating whether the accents are in fact different and whether they belong to distinct categories.

There are no simple solutions to these difficulties, and no single methodology can be ideal for investigating all aspects of prosodic interpretation. We suggest, however, that the eye-tracking paradigm described in the next section should be included in the arsenal at our disposal for investigating prosody, and we argue that it has certain clear advantages.

2. Eye-tracking and spoken language

2.1 Saccadic eye movements

During everyday tasks involving vision, such as reading a newspaper, looking for the car keys, making a cup of coffee and conversing about objects in the immediate environment, people rapidly shift their gaze to bring task-relevant regions of the visual field into the central area of the fovea (Hayhoe & Ballard, 2005; Hayhoe, 2004; Land, 2003; for reviews see E. Kowler, 1995). Eye movements are necessary because visual sensitivity differs across the retina; acuity is greatest in the central portion of the fovea, and then markedly declines. The organization of the retina can be viewed as a compromise between the need to maintain sensitivity to visual stimuli across a broad range of the visual field, while also allowing detailed spatial resolution for task-relevant aspects of the visual field. This division of labor also helps restrict most processing to a relevant subset of the visual field, thus reducing the amount of information from the visual environment, in service of selection for action (Allport, 1989). Gaze shifts that bring new regions of the field into the fovea, where visual acuity is greatest, are accomplished by saccadic eye movements (Hayhoe, 2000; Hayhoe, 2004; Kowler, 1999; Kowler, 1995; Liversedge & Findlay, 2000).

Saccades are rapid ballistic eye movements. During a saccade, the eye is in motion for 20 to 60 ms, with the duration of the saccade related to the distance that the eye travels. During a saccade, sensitivity to visual information is dramatically reduced. Suppression of visual information occurs in part because of masking and, in part, because of central inhibition (Kowler, 1995; Liversedge & Findlay, 2000; see Rayner, 1998 and references, therein). A saccade is followed by a fixation that typically lasts for 200 ms or more depending upon the task. The minimal latency for planning and executing a saccade is approximately 150 ms when there is no uncertainty about target location. In reading, visual search, and in other tasks in which there are multiple target locations, saccade latencies are somewhat slower, typically about 200 to 300 ms. The pattern and timing of saccades, and the
resulting fixations, are one of the most widely used response measures in the cognitive sciences, providing important insights into the mechanisms underlying attention, visual perception, reading and memory (Rayner, 1998). Recent overviews of eye movements in scene perception are provided by Henderson and Hollingsworth (2003) and Henderson and Ferreira (2004b).

Recently, the development of accurate, relatively inexpensive head-mounted and remote eye-tracking systems has made it possible to monitor eye movements as people perform natural tasks. Eye movements naturally occur in tasks involving vision, they occur rapidly in response to even low-threshold signals, and because they are ballistic, there is little uncertainty about when a saccade has been initiated and what part of the visual field is being fixated. Crucially, they are closely linked to attention. Although, attention can be directed to regions of space that are not currently being fixated, or about to be fixated, a growing body of behavioral and neurophysiological research supports a close link between fixation and spatial attention (Gilchrist, Heywood, & Findlay, 2003; Kowler, 1999; Liversedge & Findlay, 2000). Thus, to the extent that attention and shifts in attention are closely time-locked to the processes that underlie comprehension and production, eye movements to taskrelevant objects should be informative about real-time language processing.

In a seminal study, Cooper (1974) demonstrated that participants’ eye movements to pictures displayed on a screen were closely time-locked to relevant information in a spoken story. The recent surge of interest in using head-mounted eye-tracking to study language processing began with Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy (1995), who examined syntactic ambiguity resolution using a task in which participants followed spoken instructions to manipulate objects in a visual workspace.

A rapidly expanding community of psycholinguists is now using head-mounted eye-trackers to study spoken language comprehension, and, more recently, language production (For some representative examples, see the chapters in Henderson & Ferreira, 2004a and Trueswell & Tanenhaus, 2005). In the action-based variant of the “visual world” paradigm, introduced by Tanenhaus et al. (1995), participants follow instructions to look at, pick up, or move, objects presented in a well-defined visual workspace or on a computer display. The timing and pattern of fixations to potential referents in the visual display are used to draw inferences about comprehension. Other studies have followed variations of the procedure introduced by Cooper (1974) often focusing on eye movements made in anticipation that a picture in a display will become task relevant (Altmann & Kamide, 1999). In production studies, eye movements are measured as the speaker
names objects, describes depicted events or generates utterances in interactive language game tasks (Griffin & Bock, 2000; Meyer & Bock, 1999).

Interest in the visual world paradigm has grown for several reasons. Eye movements provide a continuous measure of spoken language processing in which the response is closely time-locked to the input without interrupting the speech stream. The use of a visual world allows researchers to explore questions about higher-level processes, such as reference resolution (Altmann & Kamide, 1999; Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; J. C. Trueswell, Sekerina, Hill, & Logrip, 1999), with a measure that provides sufficient temporal resolution to measure the effects of fine-grained phonetic variation on lexical access (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; McMurray, Tanenhaus, & Aslin, 2002; Salverda, Dahan, & McQueen, 2003). The visual world paradigm can be used with natural tasks that do not require meta-linguistic judgments. Thus it is well suited for studies with young children (Trueswell et al., 1999) and with brain-damaged populations (Yee, Blumstein, & Sedivy, 2000). The presence of a visual world also makes it possible to ask questions about real-time interpretation, that would be difficult, perhaps intractable, if one were limited to measures of processing complexity for written sentences or spoken utterances (c.f. Sedivy et al., 1999). Finally, the paradigm allows one to study real-time language production and comprehension in natural tasks involving conversational interaction (Brown-Schmidt, 2005; Brown-Schmidt, Campana, & Tanenhaus, 2005).
Before we turn our focus to how the visual world paradigm can be used to study prosody, we discuss a study by Allopenna, Magnuson and Tanenhaus (1998) in order to illustrate how eye movement data are analyzed.

2.2 Tracking lexical access in continuous speech

Allopenna, et al. (1998) evaluated the time course of activation for lexical competitors that shared initial phonemes with the target word (e.g., beaker and beetle) or that rhymed with the target word (e.g., beaker and speaker). In the Allopenna et al. studies, participants were instructed to fixate a central cross and then followed a spoken instruction to move one of four objects displayed on a
computer screen with the computer mouse (e.g., “Look at the cross. Pick up the beaker. Now put it above the square”).

A schematic of a sample display of pictures is presented in Figure 3, Panel A. The pictures include the target (the beaker), a word sharing the same initial phonemes, which we will refer to as a cohort competitor (the beetle), a picture with a name that rhymes with the target (speaker) and the unrelated picture (the carriage). For purposes of illustrating how eye movement data are analyzed, we will restrict our attention to the target, cohort and unrelated pictures. The particular pictures displayed are used to exemplify types of conditions and are not repeated across trials in a typical experiment. Panel B shows five hypothetical trials. The 0ms point indicates the onset of the spoken word beaker. The dotted line begins at about 200 ms—the earliest point where we would expect see signal-driven fixations (Hallett, 1986).

On trial one, the hypothetical participant initiated a fixation to the target about 200 ms after the onset of the word, and continued to fixate on it (typically until the hand brings the mouse onto the target). On trial two, the fixation to the target begins a bit later. On trial three, the first fixation is to the cohort, followed by a fixation to the target. On trial four, the first fixation is to the unrelated picture. Trial five shows another trial where the initial fixation is to the cohort. Panel C illustrates the proportion of fixations over time for the target, cohort and unrelated pictures, averaged across trials and participants. These fixation proportions are obtained by determining the proportion of looks to the alternative pictures at a given time slice and they show how the pattern of fixations change as the utterance unfolds. The fixations do not sum to 1.0 as the word is initially unfolding because participants are often still looking at the fixation cross.

Researchers often define a window of interest, illustrated by the rectangle in Panel C. For example, one might want to focus on the fixations to the target and cohort in the region from 200 ms after the onset of the spoken word to the point in the speech stream where disambiguating phonetic information arrives. The proportion of fixations to pictures or objects, the time spent fixating on the alternative pictures (essentially the area under the curve, which is a simple transformation of proportion of fixations), and the number and/or proportion of saccades generated to pictures in this region can then be analyzed. These measures are all highly correlated.
Figure 4 shows the actual data from the Allopenna et al. (1998) experiment. The figure plots the proportion of fixations to the target, cohort, rhyme and unrelated picture. Until 200 ms, nearly all of the fixations are on the fixation cross. These fixations are not shown. The first fixations to pictures begin at about 200 ms after the onset of the target word. These fixations are equally distributed between the target and the cohort. These fixations are remarkably time-locked to the utterance: input-driven fixations occurring 200 to 250 ms after the onset of the word are most likely programmed in response to information from the first 50 to 75 ms of the speech signal. At about 400 ms after the onset of the spoken word, the proportion of fixations to the target began to diverge from the proportion of fixations to the cohort. Subsequent research has established that cohorts and targets diverge approximately 200 ms after the first phonetic input, including coarticulatory information in vowels, that provides probabilistic evidence favoring the target (Dahan et al., 2001; Dahan & Tanenhaus, 2004; Dahan, Tanenhaus, & Chambers, 2002).
Figure 4: The fixation proportion over time adapted from Allopena et al. (1998)

Shortly after fixations to the target and cohort begin to rise, fixations to rhymes start to increase relative to the proportion of fixations to the unrelated picture. This result differentiates between predictions made by the cohort model of spoken word recognition and its descendants (e.g., Marslen-Wilson, 1990; Marslen-Wilson, 1993; Marslen-Wilson, 1987), which assume that mismatch at the onset of a word is sufficient to strongly inhibit a lexical candidate, and continuous mapping models, such as TRACE (McClelland & Elman, 1986), which predict competition from similar words that mismatch at onset (e.g., rhymes). The results strongly confirmed the predictions of continuous mapping models.

3. Eye-tracking and pitch accents

Dahan, Tanenhaus, and Chambers (2002) hereafter DTC, used cohort competitors to investigate the influence of pitch accenting on reference resolution. The cohort manipulation is ideal for studying pitch accents because one can examine effects that are localized to the vowel that carries the pitch accent. In DTC, the display contained four objects. On critical trials, two of the objects had names that began with the same initial stressed syllable, e.g., candle and candy. (For monosyllabic names, the overlap always included the vowel.) A trial consisted of two consecutive instructions, as exemplified in (4).

(4) a. Put the candle/candy below the triangle.
   b. Now put the CANDLE/candle above the square.

The first instruction mentioned one of the objects (e.g., "Put the candle / candy below the triangle"), making it given information and setting the context for the second, critical instruction. Depending on which object was mentioned in the first instruction, the second instruction could refer either to the same object or to a different object (e.g., "Now put the candle above the square"). Upon just hearing can-, the word is ambiguous because the listener does not yet know whether the candle or the candy is being referred to. DTC took advantage of this ambiguity by manipulating the prosodic event occurring on the vowel of the ambiguous portion of the word. Any fixation preferences seen during this period must be attributed to effects of prosody.
In DTC’s study, the temporal region of interest was the onset of *can-* in (4b). If listeners rapidly interpret accented words as referring to new information, there should be a bias towards fixating on the candle when it is new and on the candy when the candle is given. In the unaccented condition, listeners should be biased towards the candle when it is given and the candy when the candle is new. In fact, this is what DTC found. The fixation data is presented in Figure 5.

**Figure 5.** Fixation proportions over time to the target and competitor in the given condition (A) and the new condition (B). There are more fixations to the given object when the cohort is de-accented and more fixations to the new object when the cohort is accented.
These data suggest that listeners interpret pitch accents rapidly on-line and that accented words are associated with new information and de-accented words are associated with given information. However, DTC also explored an alternative hypothesis. Researchers have argued that subjects and themes occupy privileged status in the discourse, and are more highly activated than other information. An alternative to the given-new hypothesis is one in which information receives an accent when it becomes more salient in the discourse. De-accenting occurs when information is already highly salient. Thus, in DTC’s study, de-accenting created a bias towards looking at the theme because the theme was highly salient. Accenting created a bias towards looking at a new entity because the entity was shifting from low to high salience in the discourse.

To disentangle the predictions of this hypothesis and the given-new hypothesis, DTC conducted a second experiment to investigate whether accenting would create a bias towards a given entity when that entity was not salient in the discourse. They compared conditions in which the critical word referred to either the theme of the previous sentence, which was highly activated, or the goal of the previous sentence, which was not activated. Note that in both instances, the referent is given. The target word “candle” was always accented in the second sentence. Sample materials are presented in (5):

(5)  
a. 1st Sentence: Theme condition
    “Put the candle below the triangle.”
b. 1st Sentence: Goal Condition
    “Put the necklace below the candle.”

c. 2nd Sentence
    “Put the CANDLE below the square.”

DTC found that listeners were quicker to fixate on the candle upon hearing the accented word “candle” in the second instruction when it was preceded by a sentence where the candle was the goal than a sentence where it was the theme, a result that has been replicated by Arnold et al. (2004). This result cannot be accounted for in a purely reference based given-new theory of accenting since candle has been referred to in both instances. DTC concluded that accenting can signal that information is moving from a low level of activation to a higher level of activation in the discourse. Thus new information is predicted to be accented, but so is information that has been mentioned but that did not previously occupy a prominent place in the discourse. Whether this proposal can provide a full account
of pitch accent placement remains to be seen. However, it is clear that eye-tracking is a useful tool for investigating these questions.

Watson, Gunlogson and Tanenhaus, and Speer and colleagues, have been using variations of this paradigm to explore processing of different types of pitch accents, specifically the presentational and contrastive pitch accents. Watson, Gunlogson, and Tanenhaus (2005) (WGT) were interested in examining how listeners interpret contrastive accents and presentational accents. As discussed above, it is not clear what contributes to the definition of contrast, despite efforts to formalize it, and it is also not clear what aspect (if any) of the pitch accent influences contrastive and presentational interpretations. The goal of the study was to find out whether any difference in interpretation could be elicited between the two accents in an eye-tracking paradigm. A difference would demonstrate that eye-tracking could be useful for asking deeper questions about the accents’ phonology and interpretation.

In a display similar to that used by DTC, WGT asked listeners to perform three sets of instructions like the one in (6) on the objects in the display in Figure 6.

![Figure 6: The display used in Watson, Gunlogson, & Tanenhaus (2005)](image)

(6) a. Click on the bed and the chair.
   b. Move the chair to the right of the square.
   c. Now, move the bed/bell below the triangle.

Creating a situation that evoked contrast was critical for the study. Example (6a) establishes a set of objects that the listener is working with (the bed and the chair), and in (6b), one of those objects is highlighted. A reference to the bed in
(6c) is potentially contrastive on at least two counts. First, it invokes the remaining member of the set explicitly introduced in the discourse, and so its interpretation is naturally set against a background of mentioned alternatives. Second, we assume that an accent on an element that is already discourse given may function contrastively. A reference to the bell in (6c), on the other hand, explicitly introduces a new entity for the first time and is not contrastive in the ways just mentioned (though there may be also be potential for contrast at a higher level with the content of the preceding instruction).

The critical comparison is the changes in fixations from 0-200ms to 200-400ms after the onset of the cohorts, bed and bell. Because it takes roughly 200ms to program an eye-movement (fixations that occur between 0-200ms after the cohort onset cannot be driven by the cohorts’ acoustic properties, so this time window serves as a useful baseline. The fixations in the baseline region are compared to the 200-400ms time window, a region of time where the cohorts’ acoustic properties should first affect fixations. A comparison of these time regions reveals very different effects for the two accents.
Figure 7: The fixations 0 – 200ms and 200-400ms after word onset for the contrast, new, and unrelated referents for L+H* (a) and H* (b). In the H* condition, fixations to the contrast referent and new referent rise at approximately the same rate in the window 200 to 400ms compared to 0 to 200ms. Looks to the unrelated referent drop off. In the L+H* condition, only looks to the contrast item rise, while looks to the new and unrelated referents decrease.

In the contrastive accent condition, looks to the contrast cohort bed increase in the 200-400ms region compared to the 0-200ms baseline, whereas fixations to the new cohort bell, and the unrelated objects decrease. In contrast, in the presentational accent condition, looks to both cohorts rise during the 200-400ms time window, while looks to the other items decrease. These data suggest that the difference between the two accents can be characterized by differences in their
distribution. The contrastive accent is more specialized, pointing to the presence of a contrast relationship with a contextually salient alternative. The presentational accent functions more generally in non-contrastive and contrastive settings.

As noted above, the initial goal was to simply elicit a difference in interpretation between the two accents, and the initial conclusions of the study are that a major difference between the accents is in their distribution across contexts, with the contrastive accent occurring in a subset of the settings in which presentational accents occur. However, to truly understand how pitch accents are interpreted, a more nuanced approach to the nature of contrastiveness will be needed. The contrastive accent appears deceptively easy to define in informal terms: it signals that the accented word should be instantiated over a highly salient alternative that is comparable along some dimension. However, this relatively straightforward definition seems to make the wrong predictions in a situation like the following. Imagine a display that contains, among other objects, two hearts that are identical except for position. Now consider the instruction consisting of the two sentences in (7) uttered in sequence:

(7) a. Click on the heart on the left.
   b. Now, click on the heart/harp on the right.

Both (7a) and (7b) will naturally have an accent on the final word, left or right respectively. In addition, given the informal definition of contrastiveness above, one might expect a contrastive accent on heart in (7b) to be acceptable since there is a highly salient and comparable alternative that is also distinct in referential terms – namely, the heart referred to in (7a). But intuitively, while a contrastive accent is natural on harp, it is highly unnatural on heart. (A presentational accent seems possible for either.)

In an experiment currently underway, WGT are using the eye-tracking paradigm to test these intuitions. The critical question is how the contrastive accent is interpreted. The pairs to be tested, like heart/harp, are cohort competitors, identical in initial segments and thus not disambiguated phonemically until relatively late in the word. This delay (potentially) allows the effect of pitch accent to emerge, during the time that both heart and harp are consistent with the unfolding phonetic input. If, as suggested, the contrastive accent is natural with harp but not with heart, then a contrastive accent on the italicized element in (7b) should result in more early looks to the harp, whether the actual word turns out to be heart or harp. Presentational accents, by hypothesis possible with either choice in (7b), should not skew the early looks toward the harp in the same way.
Pending specific results of the experiment just described, several more general points can be made about investigating pitch accents. First, assuming that the intuitive judgments about (7) are borne out, it is clear that any formulation of contrastiveness that calculates the availability of alternatives based solely on reference will go astray in contexts like (7), where the presence of a referentially distinct alternative does not suffice to license a contrastive accent. Second, the intuitive unavailability of the contrastive accent on heart in (7b) suggests there is indeed content to the distinction between presentational and contrastive marking, even if (as remains possible) they are two ends of a spectrum rather than distinct categories. Finally, we note the usefulness of the eye-tracking paradigm both for testing and refining intuitive judgments in general, and for providing a way in particular to test a hypothesis about an “unacceptable” linguistic choice without recourse to meta-linguistic judgments.

A recent study by Ito and Speer (2005) adopts a “targeted language game” in which spontaneous utterances are first collected in a well-defined task. Then carefully controlled utterances are used to evaluate hypotheses that emerge from that task, using eye movements as the dependent measure. Ito and Speer’s work is based on the Christmas tree decorating task developed by Speer and her colleagues.

In this task, the director, a naïve participant, instructs a confederate about how to decorate a Christmas tree using ornaments that need to be placed on the tree in a specified sequence. Ornaments differ in type, e.g., bells, hats, balls, houses, etc. and in color, e.g., orange, silver, gold, blue, etc. Recordings demonstrated that participants typically used a presentational accent (H*) when a color was new to the local discourse. For example, “orange” typically received a presentational accent in the instruction, “First, hang an orange ball on the left” when an orange ornament was being mentioned for the first time) for a particular row. However, if the instruction to place the orange ball followed placement of a ball of a different color, e.g., a silver ball, then “orange” was more likely to be produced with a contrastive accent (L+H*).

Ito and Speer recorded utterances that were modeled on those produced by the naïve speakers, and played them to naïve participants. Participant’s eye movements were monitored as they followed the instruction. The dependent measure was the time to find the appropriate ornament. An L+H* accent on the color facilitated target recognition as compared to an H* accent, when an instruction followed placement of the same type of object of a different color. Thus, the preferred pitch accent pattern used by naïve participants facilitated performance, as measured by eye movements.
4. Eye-Tracking and timing

Given the temporal precision of eye-tracking, perhaps one of the areas where use of the paradigm is most exciting is in questions related to the prosodic timing of a sentence. By timing, we mean to refer to the sentence prosodic phrase structure, segmental duration, and rhythmic structure. Although researchers have pointed out that timing information may be of use to a listener in processing a sentence (for review, see Cutler, Dahan & van Donselaar, 1997), there have been few ways of investigating the degree to which this information is utilized by a listener as a sentence is heard on-line. Recently, researchers have begun to use eye-tracking to investigate these questions.

For example, Salverda, Dahan, & McQueen (2003) investigated the effects of vowel duration and prosodic boundaries on the disambiguation of word pairs in which one of the words (e.g. ham) is embedded in a longer carrier word (e.g. hamster). Previous work had suggested that fine-grained sub-phonological properties of the words, specifically vowel length, could be used in disambiguation (Davis et al., 1997). The vowel in monosyllabic embedded word tends to be longer in duration than the vowel in the first syllable of a carrier word, and listeners used this information to resolve the ambiguity. Salverda et al. proposed that prosodic structure drives these vowel differences. Because the offset of the word “ham” corresponds with a larger prosodic boundary then the offset of the syllable “ham-” in “hamster”, lengthening of the vowel, which often occurs immediately before a prosodic boundary, is greater for the embedded word than the carrier word.

To test this hypothesis, Salverda et al. presented native speakers of Dutch with displays like Figure 8, which contained four shapes and four picturable nouns, two of which were ambiguous pairs (e.g. ham and hamster).

![Figure 8. The display from Salverda et al. (2003).]
The participants’ task was to identify the target word that was mentioned in the test sentences. The target was always the carrier word (e.g. hamster). The critical manipulation was the manner in which the first syllable in “hamster” was acoustically realized. It was either spliced from a production of the monosyllabic word “ham” or from another version of the word “hamster”.

![Proportion of fixations from Salverda et al. (2003)](image)

**Figure 9.** Proportion of fixations from Salverda et al. (2003)

Figure 9 shows the proportion of fixation from the onset of the target, where “hamster” is the target and “ham” is the competitor. There are clear effects of the origin of the syllable. Conditions in which the syllable originated from the embedded word created increased looks to the competitor and conditions in which the syllable originated from the carrier word increased looks to the target. Critically, Salverda et al. found no effect when they re-ran the study with syllable tokens extracted from the embedded and carrier words that were matched for length, suggesting that duration is critical for listener disambiguation, and it is likely that these durational differences are driven by prosodic position.

Evidence consistent with this hypothesis comes from a similar eye-tracking study by Crosswhite et al. (in revision), who found that embedded and carrier words were less confusable in intonational phrase final positions, further
suggesting that the listeners’ use of vowel duration is influenced by overall prosodic structure. Consistent with this claim, Salverda et al (2005) showed that position in a prosodic domain affects the degree to which different lexical competitors become active in memory. Comparing instructions such as *Put the cap below the square* and *Click the on the cap*, in four-picture displays that included a picture of a cat, a cap, and a captain, Salverda et al. found that in medial position, participants were more likely to look at the captain than the cat, as they heard the word *cap*. This pattern reversed in utterance final position.

Other experiments investigating prosodic timing have explored the degree to which listeners use timing to predict upcoming linguistic structure. A study by Arnold, Tanenhaus, Altmann & Fagnano (2004), while not a direct investigation of prosody, used eye-tracking to understand the processing of disfluencies. Using displays similar to those in Figure 6, they found that the presence of a disfluency such as “…the uh…” influenced the initial interpretation of the following noun. In the fluent utterance, listeners were biased towards fixating a given referent, replicating the pattern found in DTC. After a disfluency, however, listeners were biased towards fixating a referent that had not been previously mentioned. The authors concluded that this result might arise because listeners are sensitive to the statistical relationship between disfluencies and new information (Arnold & Tanenhaus, in press) or, alternatively, because listeners might have inferred that speaker production difficulty was likely to co-occur with information that was new to the discourse. Research that attempts to discriminate among these alternatives is in progress.

One could easily use this paradigm for investigating similar claims about aspects of sentence timing such as intonational boundaries. Researchers have argued that the location of intonational boundaries correlate in part with speaker production difficulty (Watson & Gibson, 2004). Specifically, it has been claimed that intonational boundaries may be a point at which utterance planning may occur. Listeners may interpret intonational boundaries in certain contexts in the same way in which they interpret disfluencies: as a possible marker of upcoming complex or inaccessible material that may require extra planning time.

The use of eye-tracking in these sets of experiments demonstrate the paradigm’s effectiveness for investigating how listeners use information that arises as a by-product of prosodic timing. First, it allows for experimental manipulation of relatively natural speech. If the listener is using disfluencies or intonational phrase boundaries to make inferences about processes in production, test utterances need to realistically reflect those processes. Secondly, the close time-locking of the response measure to the signal allows one to test strong hypotheses about how listeners combine prosodic information with other information in the utterance; the experimenter has a measure of interpretation from the point of processing a
specific event such as the duration of an ambiguous word, a disfluency, or an intonational boundary.

5. Final Thoughts

Although we argue that monitoring eye movements can provide a useful measure of prosodic interpretation, we do not mean to imply that eye-tracking provides a direct window into language comprehension. A number of different factors contribute to what a participant ultimately fixates on, including the goals of the task, the visual display, and the language that is used. Of course, problems of one type or another arise with all experimental measures. In visual world studies, it is important for the experimenter to design the task such that the visual display is controlled for visual interest and the task is closely matched to the communicative signal.

Nonetheless, the sample eye-tracking studies that we have reviewed would seem to avoid some of the problems we noted in Section 2. The eye-tracking paradigm provides a measure of interpretation without resorting to difficult metalinguistic judgments. It also allows the experimenter to control and operationalize the context of an utterance and its potential communicative goals. At the same time, the response measure is sensitive to even small systematic acoustic variations in the signal. Finally, the paradigm can be used in complex interactive discourse environments. Thus the visual world paradigm may prove to be a powerful tool in the study of prosody.

6. References


Crosswhite, K., Masharov, M., McDonough, J.M., & Tanenhaus, M.K. (in revision) Phonetic cues to word length in the on-line processing of onset-embedded word pairs.


Yee, E., Blumstein, S., & Sedivy, J. C. (2000). The time course of lexical activation in broca's aphasia: Evidence from eye movements. *13th annual CUNY conference on human sentence processing, La Jolla, CA,*