Interpreting Pitch Accents in Online Comprehension: 
H* vs. L+H*

Duane G. Watsona, Michael K. Tanenhausb, Christine A. Gunlogsonc

aBeckman Institute and Department of Psychology, University of Illinois at Urbana-Champaign
bDepartment of Brain and Cognitive Sciences, University of Rochester
cDepartment of Linguistics, University of Rochester

Received 15 December 2006; accepted 6 March 2008; received in revised form 4 December 2007

Abstract
Although the presence or absence of a pitch accent clearly can play an important role in signaling the discourse and information structure of an utterance, whether the form of an accent determines the type of information it conveys is more controversial. We used an eye-tracking paradigm to investigate whether H*, which has been argued to signal new information, evokes different eye fixations than L+H*, which has been argued to signal the presence of contrast. Our results demonstrate that although listeners interpret these accents differently, their interpretive domains overlap. L+H* creates a strong bias toward contrast referents whereas H* is compatible with both new and contrast referents.

Keywords: Prosody; Language; Discourse; Accent; Pragmatics

1. Introduction
It is well known that in English and many other languages, emphasis on select syllables in a sentence can play an important role in communicating how the content expressed fits into the structure of the discourse. Emphasis, or stress, has no single acoustic correlate in English; its presence is signaled by a variety of cues including greater intensity of the sound wave, longer syllable duration, and (often) a pitch movement associated with the stressed syllable (Ladd, 1996). What is less well understood, and difficult to evaluate experimentally, is whether systematic variation in the realization of prominence, particularly in the shapes of pitch accents, can be linked to interpretation. This question is of interest to computer scientists interested in building systems that can process and produce spoken language naturally (Hirschberg, 2002), to psychologists interested in the cognitive mechanisms that underlie prosody and spoken language production and comprehension (Cutler, Dahan, & van Doneselaar, 1997;
Shattuck-Hufnagel & Turk, 1996), and to linguists interested in characterizing the acoustic and semantic properties prosodic information might convey (Ladd, 1996; Pierrehumbert & Hirschberg, 1990). In this article, we examine the real-time comprehension of so-called contrastive accents versus presentational accents, which have been claimed to be characterized by \( L+H^* \) and \( H^* \) pitch accents, respectively. By manipulating the pitch accent using temporarily ambiguous words, and tracking eye movements, we show that it is possible to isolate effects of specific prominence patterns.

The presence or absence of prominence on a particular word in an utterance seems to be linked to the information structure of that utterance (Gussenhoven, 1983; Schwarzschild, 1999; Selkirk, 1995). Consider the sentence “Sarah dissed Bob.” When produced with primary stress on Sarah, as in Example 1a, the utterance is a natural answer to the question, “Who dissed Bob?” When, however, the primary stress is on Bob, as in Example 1b, the utterance more naturally answers the question, “Whom did Sarah diss?” In Example 1a, Sarah provides information that is new to the discourse, whereas in Example 1b, Sarah is part of information that is given and Bob is part of the new information:

1. a. SARAH dissed Bob.
   b. Sarah dissed BOB.

Detailed linguistic analyses, clear intuitions, and a body of psycholinguistic evidence support the claim that the location of prominence in an utterance correlates with the division of the sentence content into categories such as focus/background or given/new, and that speakers and listeners make use of this information as they construct utterance interpretations (see Cutler et al., 1997). In contrast, the claim that the particular form of the pitch accent contributes to the interpretation of an utterance is more controversial and has received far less empirical support. The current report investigates one pair of pitch accents, \( H^* \) and \( L+H^* \), which are argued to make distinctive contributions to interpretation in the theory of Pierrehumbert and Hirschberg (1990) and the ToBI labeling system (Beckman & Ayers, 1994; Beckman & Hirschberg, 1994; Silverman et al. 1992).

In Pierrehumbert and Hirschberg’s treatment, the presence of any pitch accent renders an element both phonologically and informationally salient. In addition, they proposed that the \( H^* \) and \( L+H^* \) accents differ categorically both in their acoustic properties and in how they interact with the discourse.

Phonetically, \( H^* \) is associated with a simple rise to a maximum in \( F_0 \). \( H^* \) is used when a speaker intends for new information to be instantiated into the listener’s discourse model. For example, in response to the question in Example 2a, the accent on Bob in Example 2b signals that this referent is new and that the proposition instantiated with the new referent should be added to the listener’s discourse model:

2. a. Whom did Sarah diss?
   b. Sarah dissed BOB.

More recent work suggests that \( H^* \) can be used to mark given referents if it marks a referent that has changed its discourse status (e.g., Dahan, Tanenhaus, & Chambers, 2002; Terken
& Hirschberg, 1994) or if the information is important or unpredictable in the conversation (Watson, Arnold, & Tanenhaus, 2008).

The “contrastive” accent (L+H*), on the other hand, has an initial drop to a low $F_0$ target followed by a steep rise to a high target. The pitch excursion is higher than that of H*, and systematic differences in peak timing within the syllable have been observed (Bartels & Kingston, 1994). Pierrehumbert and Hirschberg suggested that L+H* conveys that the accented item should be instantiated in the listener’s discourse model rather than a more salient alternative. To illustrate, the accent on Bob in Example 3b indicates that Bob is to be instantiated, and not the other prominent referent in the discourse, Annie:

3. a. Did Sarah diss Annie?
   b. No, Sarah disses BOB.

L+H*

The picture presented by Pierrehumbert and Hirschberg is not universally accepted; researchers disagree about which units constitute intonational morphemes, how they should be described phonologically, and what they mean. With respect to H* and L+H* specifically, there is dispute about whether these accents are two distinct categories or variants of a single accent type. One area of controversy involves the acoustic properties of these accents. For instance, Ladd and Schepman (2003) argued that H* can be preceded by a low pitch target, contrary to claims by Pierrehumbert (1980; see also Beckman & Pierrehumbert, 1986; Pierrehumbert & Hirschberg, 1990) that this is only true for L+H*. Moreover, as noted earlier, differences between realizations of L+H* and H* involve more than one acoustic dimension, and it may be that properties such as peak height are crucial to any interpretative differences. While leaving this possibility open, we continue to frame the discussion in terms of pitch accent shape because that is how claims about meaning have been put forward.

Among researchers in semantics and pragmatics, other accounts of the meaning difference have been given. Steedman (2000) proposed that, although L+H* and H* both mark material that is contrastive in a broad sense, the two differ in marking themes and rhemes, respectively. The line of work pursued by Büiring (1997, 2003) and Roberts (1996) treated contours involving L+H* as marking discourse topics. Although these theories differ substantially from each other and from Pierrehumbert and Hirschberg, they share the assumption that L+H* and H* are distinct categories.

A related issue is whether these two accents are associated with different meanings. Whenever a listener chooses to instantiate some piece of information into the discourse model, it necessarily means some alternatives are left uninstantiated. In fact, within the influential semantic framework of Rooth (1985), the role of prosodically marked “focus” more generally is to evoke alternatives. Thus, L+H* might be a special case of H*, differing only in the size and prominence of the set of alternatives.

Investigating these claims empirically has presented a challenge. On the one hand, investigating the differences between these accents requires not only an appropriate discourse, but also requires that speakers and listeners be engaged in a task where sufficiently conveying pragmatic information is important. On the other hand, understanding the acoustic and pragmatic differences between these accents requires a large degree of experimental control. These
two conflicting constraints have not only created difficulties in understanding these two accent types, but prosodic categories more generally.

To meet this challenge, we use eye tracking in the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm, participants are instructed to manipulate real-world objects or images on a computer while their fixations are monitored. This paradigm has been used successfully to investigate a wide range of psycholinguistic questions including studies of speech perception (McMurray, Tanenhaus, & Aslin, 2002), lexical access (Allopenna, Magnuson, & Tanenhaus, 1998), online pragmatic interpretation (Sedivy, Tanenhaus, Chambers, & Carlson, 1999), and ambiguity resolution (Spivey, Tanenhaus, Eberhard, & Sedivy, 2002). Because fixations are closely time-locked to the acoustic signal and are sensitive to subtle acoustic-phonetic cues, eye tracking is a particularly useful tool for studying the interpretation of pitch accents (Chen, den Os, & de Ruiter, 2007; Ito & Speer, 2008; Watson, Gunlogson, & Tanenhaus, 2006; Weber, Braun, & Crocker, 2006).

In this study, we created conditions where a referring expression (e.g., “the camel”) begins with a stressed syllable whose onset and nucleus are temporarily ambiguous between the name of its intended referent (e.g., a picture of a camel) and a referent whose name is a cohort competitor (e.g., a picture of a candle). We manipulated whether the pitch accent on the stressed vowel is an $H^\ast$ or an $L+H^\ast$ and whether the target and competitor refer to discourse-new (i.e., visible but unmentioned) entities or to entities that have previously been mentioned and paired with a mentioned alternative. Thus, contrast was operationalized in terms of a two-member set of salient entities established by the discourse (i.e., previously mentioned or “given”).

We used the action-based version of the visual world paradigm (Tanenhaus et al., 1995), building on previous results by Allopenna et al. (1998) and Dahan et al. (2002). Allopenna et al. monitored participants’ eye movements as they followed spoken instructions, such as:

4. Pick up the candle. Now put it above the diamond.

The displays contained four moveable pictures and four locations. On trials where the display contained a picture of the target (e.g., a picture of a candle) and a picture of a cohort competitor (e.g., a picture of a candy), looks to the target and the cohort competitor began to increase compared to looks to pictures of objects with unrelated names about 200 msec after the onset of the target word, approximately the time it takes to plan and initiate a saccadic eye movement. Looks to the target and cohort began to diverge about 200 msec after the phonemically disambiguating information.

Dahan et al. used the Allopenna et al. (1998) paradigm to investigate whether the presence of a pitch accent biases listeners toward interpreting a temporarily ambiguous noun as referring to a discourse-given or discourse-new entity. In Dahan et al.’s Experiment 1, participants heard instructions such as:

5. a. Put the candy/camel above the diamond.
   b. Now put the camel/CAMEL above the square.

The target in the second instruction was either accented or unaccented, and referred to either the picture referred to in the first sentence or to a previously unmentioned picture. The use of cohorts allowed Dahan et al. to localize effects to the stressed syllable—the primary locus of differences in pitch accent. The pattern of looks to the target and cohort showed that
unaccented nouns were initially biased toward the most salient, given entity, whereas accented nouns were biased toward a new less salient entity.

We investigated three possible hypotheses for how \( L^+H^* \) and \( H^* \) might differ: (a) \( L^+H^* \) signals contrast with a mentioned referent, and \( H^* \) signals a new referent; (b) \( H^* \) and \( L^+H^* \) do not differ in how they are interpreted; and (c) \( L^+H^* \) is a marked accent that is limited to a contrastive interpretation, whereas \( H^* \) is a less marked accent whose interpretation ranges over both the contrastive and new cases.

Sample experimental instructions are given in Example 6:

6. a. Click on the camel and the dog.
   b. Move the dog to the right of the square.
   c. Now, move the camel/candle below the triangle.

\( H^*/L^+H^* \)

The instruction in Example 6a establishes a paired set of salient objects, thus creating a potential contrast set (i.e., the camel and the dog). The instruction in Example 6b makes one of these objects (the dog) highly accessible in the discourse, thus creating the potential for contrast in Example 6c when reference is made to the less accessible mentioned entity. The use of cohorts that are phonemically disambiguated after the vowel allows us to localize effects to the vowel carrying the pitch accent during the time when the input is phonemically ambiguous between the given cohort (camel) and the new cohort (candle).

If \( L^+H^* \) is associated with contrast and \( H^* \) is associated with focusing new information, as the word unfolds, \( L^+H^* \) should create a bias toward the camel in Example 6c, and \( H^* \) should create a bias toward the candle. If \( H^* \) and \( L^+H^* \) do not differ in how they are interpreted, there should be no effect of accent. Fixations to the candle and the camel should rise at the same rate in the \( H^* \) and \( L^+H^* \) conditions as the words unfold. Finally, if the semantic requirements of \( L^+H^* \) are stricter than those of \( H^* \) such that \( L^+H^* \) is associated with contrast and \( H^* \) is associated with both contrastive and new referents, \( L^+H^* \) should create a bias toward fixating on the camel, whereas \( H^* \) should be consistent with either the camel or candle. In the \( H^* \) condition, looks to the camel and candle should rise at roughly the same rate as the word unfolds, whereas in the \( L^+H^* \) condition, looks to the camel should rise more quickly than looks to the candle.

2. Method

2.1. Participants

Sixteen volunteers from the University of Rochester community received $7.50 for participating in this experiment.

2.2. Materials and procedure

Twenty pictures of objects with names that shared initial phonemic segments were selected for this experiment (e.g., camel and candle). In each trial, participants were presented with a
Table 1

$F_0$ and duration means of the stressed word in the H* and L+H* conditions

<table>
<thead>
<tr>
<th>Syllable</th>
<th>H* (Presentational)</th>
<th>L+H* (Contrastive)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Maximum $F_0$</td>
<td>109.0</td>
<td>3.4</td>
</tr>
<tr>
<td>Minimum $F_0$</td>
<td>84.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Duration (msec)</td>
<td>397</td>
<td>15</td>
</tr>
</tbody>
</table>

5 × 5 grid. The grid contained four shapes that did not move on any of the trials, and four objects, two of which were the phonetic cohorts. All pictures came from the colorized version of the Snodgrass and Vanderwart (1980) set (Rossion & Pourtois 2003).

Each test trial contained three instructions similar to those in Example 6. The first utterance instructed the participant to click one of the cohorts and one of the other objects, thus establishing a potential contrast set. The second instruction was to move the previously mentioned non-cohort object to a location near one of the shapes, thus establishing it as the more salient entity. The final instruction was to move either the previously mentioned cohort or the unmentioned cohort to a location near one of the shapes.

Twenty test trials were randomly intermixed with 24 filler trials. Eleven filler trials contained phonetic cohorts. Whether a cohort was referred to or not was counterbalanced to discourage participants from focusing on cohorts as expected targets.

Instructions were produced by a male speaker with the desired intonation. The appropriateness of each accent was checked by a ToBI trained coder. Each recording was sampled at 44.1 kHz. Acoustic measures of $F_0$ and duration were analyzed to confirm the annotator’s coding. The means are presented in Table 1. Within the stressed word, the L+H* items had reliably greater pitch excursions than the H* accents, $F(1, 39) = 27.64, p < .001$; and started from a lower $F_0$ point, $F(1, 39) = 30.26, p < .001$. The target word in L+H* items was also longer than in the H* condition, $F(1, 39) = 95.24, p < .001$. Fig. 1 provides a stylized view of the H* and L+H* contours across items. The stressed syllable of each item was divided into four parts, and the average $F_0$ for each quartile is presented in the figure. Note that there is a steeper rise to the high target in the L+H* condition than in the H* condition.1

A version of each critical sentence was used as a baseline sentence. Each condition was edited so that the target word and the remainder of the sentence were spliced out of their original context and into the corresponding point in the baseline. The baseline condition was counterbalanced across conditions to control the prosody of the pre-target material. The pre-target material in all conditions was “Now move the.” The word now was produced with either an H* or L+H* accent (5 H* and 15 L+H*). Critically, the same preamble was used across conditions for each item. The word now was always followed by an intonational phrase boundary (i.e., a brief pause).

Four lists were created by rotating each set of test instructions through the four conditions created by crossing pitch accent and target type. Two orders were then created for each list, a forward and backward order, resulting in eight lists. Participants were assigned to one of the lists.
3. Results and discussion

A research assistant naïve to the purpose of the study coded the fixations using the crosshairs generated by the ISCAN software to record what items were being fixated in each frame of video. Fig. 2 presents the proportion of fixations for the L+H* and the H* accents for given and new targets. Factoring in the 200 msec it takes to program and launch a saccadic eye-movement in displays such as the ones we used, the region from 0 to 200 msec should reflect only baseline biases due to the discourse. Fixations in the region from 200 msec to 400 msec should largely reflect input from the onset and the vowel (the first 200 msec of the target word), as the disambiguating information is downstream. Thus, effects of pitch accent should be most marked in this region.

Figure 3 presents the combined looks to potential referents that are contrast members (given targets and cohorts), potential referents that are discourse new (new targets and cohorts), and the unrelated distracters for the 0 to 200 msec and 200 to 400 msec regions for the L+H* and
Fig. 2. The graphs of the fixation proportions for the contrast condition and the new condition. Note: The onset of candle or camel is 0 msec. Panel A presents the data for L+H*, and Panel B presents the data for H*. For L+H*, input-driven fixations begin to increase to the contrast member (given targets and given cohorts), with delayed looks to the target when it is new. For H*, there is an overall bias toward the given referent; however, input-driven fixations rise to all potential referents, regardless of whether they are given or new.
Fig. 3. The fixations 0 to 200 msec and 200 to 400 msec after word onset for the contrast, new, and unrelated referents for L+H* and H*. Note: In the H* condition, fixations to the contrast referent and new referent rise at approximately the same rate in the window of 200 to 400 msec compared to 0 to 200 msec. Looks to the unrelated referent drop off. In the L+H* condition only looks to the contrast item rise, whereas in the H* condition, looks to the new and unrelated referents decrease.
the H* pitch accents. For the L+H* accent, fixations clearly increase to contrast members that are consistent with the unfolding target word, whereas they decrease somewhat to the new referents. This data pattern is exactly what is expected on the hypothesis that L+H* prefers a contrastive interpretation.

The pattern for H* is clearly different. In the 200 to 400 msec region, fixations increase to all potential referents with names consistent with the input, regardless of whether they are contrast members or discourse new, although later on there is a given bias. These results are clearly inconsistent with the hypothesis that H* prefers a discourse new referent, but consistent with the hypothesis that H* is an unmarked pitch accent.

We evaluated the data patterns statistically by conducting an analysis of variance with three factors: region (0–200 and 200–400), accent type (L+H* and H*), and discourse status (contrast member and discourse new). The three-way interaction of region, accent, and contrastive status was reliable by participants and by items, $F_1(1, 15) = 5.313, p < .05$ and $F_2(1, 19) = 7.21, p < .05$. A sub-analysis on the L + H* conditions showed an interaction of region and contrast, $F_1(1, 15) = 16.427, p < .01$ and $F_2(1, 19) = 8.18, p < .05$; with a higher proportion of fixations to the contrast members in the 200 to 400 msec region compared to the 0 to 200 msec region, but no difference between looks to discourse-new referents across region. For the H* condition, there was a main effect of region, with greater looks to the objects in the 200 to 400 msec region than in the 0 to 200 msec region, $F_1(1, 15) = 10.29, p < .01$ and $F_2(1, 19) = 13.56, p < .01$. Crucially, there was no hint of an interaction by region and discourse status, $F_s < 1$.

The pattern of results during the vowel is clearly most consistent with the hypothesis that L+H* signals a contrastive interpretation, whereas H* is an unmarked pitch accent. However, the fixation pattern over a longer duration suggests a bias for the given referent even for the H* accent. This bias is reflected in a higher proportion of looks to given targets compared to new targets and given cohorts compared to new cohorts (for similar results, see Experiment 2 of Dahan et al., 2002).

To evaluate the pattern of results over a larger region, we conducted an analysis that included the proportion of looks from 200 to 800 msec after the onset of the target word, conducting separate analyses for the targets and the competitors. There was an overall preference for looking at the previously mentioned contrast object when it was the target, $F_1(1, 15) = 26.074, p < .001$ and $F_2(1, 17) = 10.14, p < .01$; and when it was the competitor, $F_1(1, 15) = 14.10, p < .01$ and $F_2(1, 19) = 8.21, p < .05$. However, the preference for looking at the contrast object was greater in the L+H* than in the H* condition. This was reflected in a marginally reliable interaction with pitch accent and discourse status for the participant analyses and fully reliable interactions for the items analyses, both for fixations to the target, $F_1(1, 15) = 3.69, p = .07$ and $F_2(1, 13) = 4.462, p < .05$; and for fixations to the competitor $F_1(1, 15) = 3.986, p = .064$ and $F_2(1, 19) = 4.45, p < .05$.

4. Conclusion

The results strongly support two conclusions. First, our results add to the growing body of evidence indicating that H* is not preferably interpreted as marking a discourse new entity.
Rather H∗ was consistent with both contrast and discourse new referents, suggesting that the accent may simply signal reference to less salient information (see Dahan et al., 2002), independent of that information’s “givenness.” Second, L+H∗ is more narrowly interpreted as signaling contrast, differing from H∗.

Although these results alone cannot tell us whether L+H∗ and H∗ accents are categorically different or two extremes of a single accent type, they provide a clearer picture of how patterns involving these pitch accents are interpreted. The results clearly rule out a simple form of the hypothesis that the domains of interpretation of H∗ and L+H∗ are mutually exclusive, with the former signaling only new referents. We can also reject the hypothesis that there is no difference between these accents because they clearly elicited different patterns of fixation. We are left to conclude that the domain of interpretation of H∗ and L+H∗ overlap. More specifically, L+H∗ is associated with contrastive interpretation whereas H∗ can mark either contrastive or discourse-new referents.

The discourse status of referents is only part of the complex story of information structure. A more comprehensive investigation can be expected to shed light on the role of accent in processing of constituents larger than NP and the issue of “focus projection” (Chomsky, 1971) with respect to different accent types.

The present study represents an important step in this research program, presenting clear evidence that listeners can decode the constellation of properties associated with prominence as a vowel unfolds, and immediately use that information to guide real-time interpretation. We hope that in future research this kind of framework can be employed to sift out contributions of specific acoustic elements such as pitch range and peak timing. Moreover, our approach to studying the online processing of pitch accents can be naturally extended to investigate extant questions about the relationship between information structure, discourse structure, and prosody.

Note

1. All sound files are available for download at http://www.psych.uiuc.edu/~dgwatson/

Acknowledgments

This project was supported by National Institutes of Health Grant HD27206. Duane G. Watson was supported by National Science Foundation Grant SES-0208484. We thank Priscilla Leung for help with data collection and analysis.

References


