What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue

Catherine Lai

Department of Linguistics, University of Pennsylvania, USA
laic@ling.upenn.edu

Abstract

This paper investigates how rising intonation affects the interpretation of cue words in dialogue. Both cue words and rising intonation express a range of speaker attitudes like uncertainty and surprise. However, it is unclear how the perception of these attitudes relates to dialogue structure and belief co-ordination. Perception experiment results suggest that rises reflect difficulty integrating new information rather than signaling a lack of credibility. This leads to a general analysis of rising intonation as signaling that the current question under discussion is unresolved. However, the interaction with cue word semantics restricts how much their interpretation can vary with prosody.

Index Terms: Dialogue, cue words, prosody, perception, attitude.

1. Introduction

Dialogue involves co-ordination of the beliefs of its participants. In particular, participants need be able to gauge levels of certainty with respect to what is being said and plan further discourse accordingly. Detection of attitudes like surprise and uncertainty is clearly important for developing dialogue models from both a theoretical and practical standpoint [1, 2].

Speakers may use overt lexical/semantic markers to indicate levels of certainty, e.g. 'I know' versus 'I doubt'. However, this sort of attitude is also expressed via prosody [3]. Prosodic features have been shown to help in the automatic detection of uncertainty [4, 5]. However, it is not clear from such classification based tasks what and how the prosodic features map to uncertainty, or what this uncertainty is directed at. That is, given that we can detect uncertainty, we would like to know how this affects structures that direct how the dialogue proceeds, such as participants' public beliefs and the Questions Under Discussion (QUD), a stack tracking what is being talked aboutl. This crucially determines relevance and hence whether or not a dialogue can move forward [6].

One prosodic feature that is closely bound to these dialogue structures is rising intonation. Rises have also been widely associated with uncertainty [7]. However, final rises do not always seem to signal uncertainty. For example, [8] find that backchannels interpretations of affirmative cue words, e.g. *okay*, are distinguished by rising pitch. Similarly, [9] find that pitch upturn is employed to encourage the interlocutor to continue speaking. So, it is not clear how these two uses of rising intonation can be reconciled. One possibility is that the underlying semantics of the utterance constrains how a rise is interpreted.

As such, this paper investigates the how rising intonation interacts with cue words. That is, one word responses like *yeah*, *right* and *really*. Like rising intonation, these discourse markers also seem to express a range of speaker attitudes. At one end of the spectrum, affirmatives like *right* and *yeah* primarily express agreement with the utterances they are produced in response to. Other affirmatives like *okay* and *sure* express acceptance of a request, which may simply be to accept last utterance in the dialogue. As such, they do not seem to express as strong agreement as *right*. At other end, responses like *really* and *well* generally express an inability or unwillingness to admit the utterance at issue into the common ground [10]. However, the interpretation of these cue words does also seem to vary significantly with prosody [11, 12]. So, one goal is to see how rising intonation varies the expression of these attitudes in order to determine how rising intonation relates to dialogue structures. Looking at this variation also sheds light on the semantics of cue words themselves, and how dialogue updates proceed in general.

The paper is structured as follows. The data and method employed in the perception experiment are described in Sections 2 and 3. The results are presented in Section 4. These results suggest, as discussed in Section 5, that rising intonation signals that the current question under discussion is unresolved and that this accounts for the affect of uncertainty and the rises associated with backchannels. Section 6 concludes.

2. Data

The stimuli for for this experiment were drawn from the Switchboard I Release 2 corpus of telephone conversational speech (LDC97S62). Each stimulus consisted of a (textual) context and (resynthesized audio) cue word response pair. The goal was to see if different types of rises and falls would affect cue word interpretation. So, the rises and falls were varied in pitch range and the presence of a peak/valley. It was expected that stimuli with larger pitch range would signal greater surprise (as in [11]), while higher peaks/lower valley would produce more emphatic interpretations. Context types were chosen to represent different levels of certainty. The goal here was to test whether lexical markers of uncertainty in the context would be mirrored in interpretation of the response.

Cue words. The resynthesized responses were derived from 6 base cue words: *really*, *well*, *okay*, *sure*, *yeah*, and *right*. Two tokens of each base word were randomly selected for resynthesis. Tokens were drawn from occurrences of the cue words in one word turns according to the transcripts. Base tokens were checked for modal voice quality. Resynthesized contours were set with respect to the start, end, and the midpoint of the stressed vowel (nucleus for diphthongs). F_0 values for the stylized contours were based on quantiles derived from F_0 values from other turns of that speaker in the same conversation. Each base token was resynthesized in 8 ways as shown in Figure 1, so that the start point was always the median value and the gradient between the mid- and endpoints remained the same. So, the stimuli in each group varied in pitch range but maintained the same

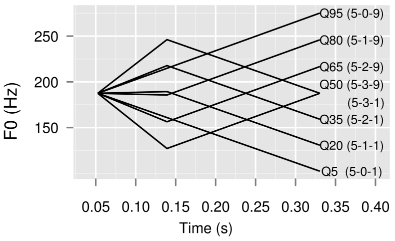

Figure 1: Stylized pitch contours for *right*, with quantiles and contour mnemonics.

slope at the end of the word. The stimuli were resynthesized using PSOLA via praat. The resynthesized versions were checked for naturalness and that each contour was audibly different.

Contexts. The contexts were drawn from turns that occurred immediately prior to one of the cue words. Four different types of context were selected for each cue word. As mentioned previously, these types were chosen to represent different levels of certainty, although they clearly do not exhaust the possible categories. Four types of context were used: (i) factual, e.g *X is Y*, (ii) evaluative, e.g. *X is good*, (iii) attributed, e.g. *I heard that X*, (iv) inferred e.g. *probably X*. Four turns were selected for each context type. So, the stimuli consisted of 6x2x8 = 96 cue words and 6x4x4 = 96 contexts in total.

3. Method

14 native speakers of American English, undergraduate students, participated in this experiment. Subjects were paid for their participation. The experiment was presented via a web interface formulated using WebExp.[1] Subjects were told that they were going to be presented with snippets from real telephone conversations. They were presented with a written context and audio (with text) response which they could listen to as many times as they chose by pressing a button. Contexts and responses were randomly paired. Subjects were asked to provide ratings on a 1-7 scale as answers to the following questions:

1. How expected does what A said seem to B? (1=completely unexpected, 7=completely expected)
2. How credible does what A said seem to B? (1=not at all credible, 7=completely credible)
3. Given B's reaction, how much would you expect A to explain or provide more evidence for what they say/why they said it? (1=wouldn't expect a follow up, 7=definitely expect a follow up).

Rather than ask directly about uncertainty, the idea was to relate uncertainty to different aspects of dialogue structure. Question 1 (EXPECTEDNESS) reflects certainty with respect to B's prior beliefs. Question 2 (CREDIBILITY) reflects how willing B is to believe A, i.e. add the content of A's utterance to their public beliefs. Question 3 (EVIDENCE) reflects the status of the QUD, i.e. whether A's utterance has been resolved/accepted or whether it is still contentious.

Subjects first completed 4 practice slides to familiarize themselves with the task. All participants reported that they understood the task before moving on to the main experiment, which consisted of 64 more slides in the same format. Note: due to a calculation error not all contexts and cue words were presented to each subject. However, the unbalanced nature of the data set is not a problem for the multilevel model used to analyze the data in the following section.

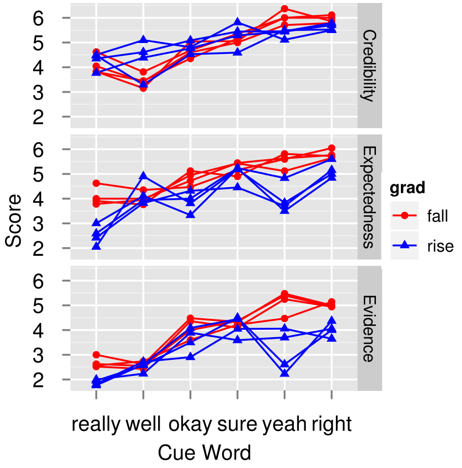

Figure 2: Mean scores for each cue word by question (question 3 reversed).

4. Results

4.1. General Trends

The mean scores for each cue word, grouped by question, are shown in Figure 2. The EVIDENCE scale has been reversed so that low scores indicate a lack of resolution of the question under discussion, hence uncertainty. Scores are generally higher for affirmative cue words with falling intonation over all of the questions. It also appears that scores increase with the affirmative strength of the cue word. With falling intonation, agreement markers *yeah* and *right* seem to convey more certainty than action accepters, *okay* and *sure*. As expected, *really* and *well*, which mark discord in the dialogue, have lower scores.

On inspection, rising intonation appears to have less of an affect on CREDIBILITY than the EXPECTEDNESS or EVIDENCE scales. In the later two cases, rising intonation pushes scores towards the uncertain end, most strikingly for *yeah*, but also for and *really*, *okay* and *right*. However, this does not seem to be the case for *well*, which seems to have the opposite trend.

4.2. Multilevel Model

A multilevel model was developed to help sort out the effects of cue word and contour, as well as context and subject variability. Following the approach outlined in [13], observed scores, y, were modelled as follows.

$$y_i \sim \mu + \alpha^{cw}_{j[i]} + \alpha^{ct}_{k[i]} + \alpha^{cx}_{l[i]} + \alpha^{s}_{m[i]} + \alpha^{cw.ct}_{j[i],k[i]} \quad (1)$$

$$\alpha^{cw}_{j} \sim N(0, \sigma^2_{cw}) \text{ for } j = 1, \dots, 6 \quad (2)$$

$$\alpha^{ct}_{k} \sim N(0, \sigma^2_{ct}) \text{ for } k = 1, \dots, 8 \quad (3)$$

$$\alpha^{cx}_{l} \sim N(0, \sigma^2_{cx}) \text{ for } l = 1, \dots, 4 \quad (4)$$

$$\alpha^{s}_{m} \sim N(0, \sigma^2_{s}) \text{ for } m = 1, \dots, 14 \quad (5)$$

$$\alpha^{cw.ct}_{j,k} \sim N(0, \sigma^2_{cw.ct}) \text{ for } j = 1, \dots, 6, k = 1, \dots, 8 \quad (6)$$

One of the benefits of this approach is that we do not have

[1] http://www.hcrc.ed.ac.uk/web_exp/

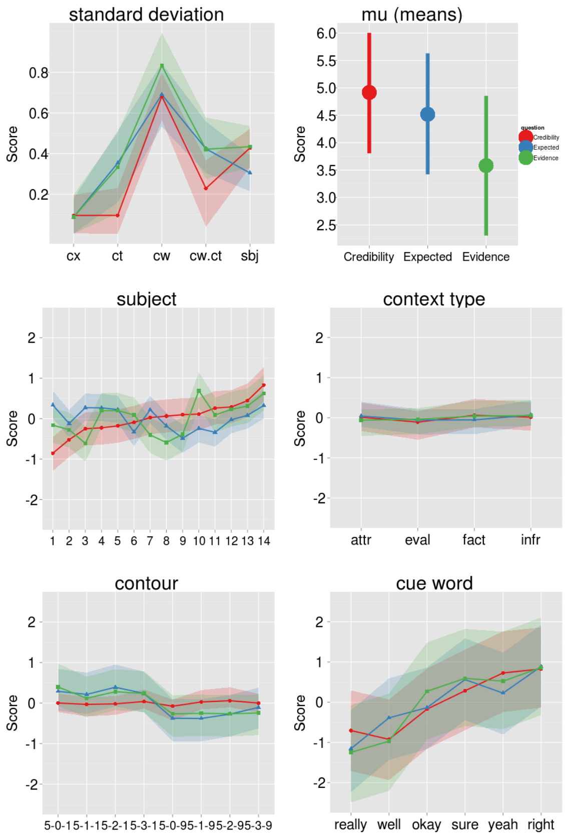

Figure 3: Parameter estimation medians. The shaded range represents 2.5th-97.5th quantiles. Red: CREDIBILITY, blue: EXPECTEDNESS, green: EVIDENCE.

to treat any of elements of each group as a baseline. So, α_k^{cw} is a parameter representing the effect of cue word k holding the other variables constant. Contour (ct), context (cx) and subject (s) and the interaction between cue word and contour (cw.ct) were similarly modelled as separate groups. The coefficients within each group were modelled as arising from different normal distributions, however their means are pulled out into a grand mean μ. The model parameters, along with finite population standard deviations for each group, were estimated using the Markov Chain Monte Carlo technique as implemented in JAGS[2] via the R package `rjags`. The model estimation passed Gelman-Rubin and Geweke convergence diagnostics.

4.3. Parameter Estimates

Figure 3 shows estimated medians and 95% intervals for the different parameters for each of the scales. The finite population standard deviations give us a measure of how variation in the actual data is associated with each factor. We can immediately note that a source of variation was subjects themselves. Subjects appeared to have different strategies for the different scales. Abstracting away from this, we can consider estimated effects of cue word, contour and context.

The actual parameter estimates for context type are very small. While there are small differences, estimates fall well inside the 95% intervals of the other types and thus do not appear to be significant. So, the interpretation of these responses did not seem to depend on the semantic context types provided in this experiment. This is also reflected in the small standard deviation estimates. The greatest standard deviation is associated with the cue words themselves. That is, cue word semantics appear to have a large effect on the perception of response credibility, unexpectedness and the need for more evidence. Again, the effect goes in the same direction as the strength of affirmation for all scales. This is not the case for the contour results, where we see a clear distinction between the CREDIBILITY rating and the other scales. For the EXPECTEDNESS and EVIDENCE scales, rising intonation pushes scores towards the low end of the scale. The posteriors associated with falls and rises appear quite distinct with with medians for rises generally lying outside the 2.5th quantile of the falling contours. This effect is not present for CREDIBILITY.

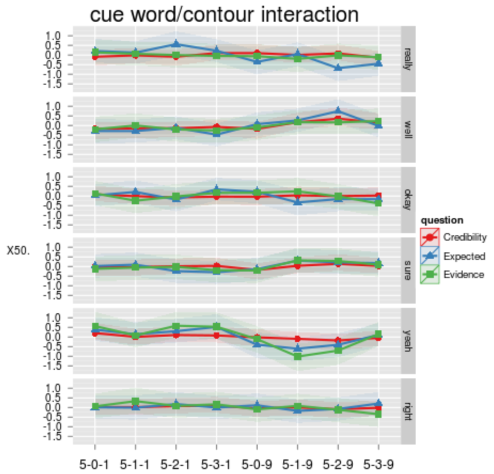

Figure 4: Cue word/contour interaction

Figure 4 shows the results for the cue word/contour interaction term. We can see that the effect of rising intonation varies across cue word. As in Figure 2, the greatest effect appears to be with respect to *yeah*, with rising contours pushing scores downwards for EVIDENCE and EXPECTEDNESS, while falling contours pull the scores up. A similar trend is observed with *really*, although interestingly variation appears to be mostly on the EXPECTEDNESS scale. On the other hand, rising contours appear to raise *well* scores.

Although the general trends for rises and falls seem fairly robust with respect to unexpectedness, we do not see much of a distinction between the different types of falling and rising contours (c.f. Figure 3). Greater pitch ranges were not really associated with the perception of more unexpectedness. This is somewhat unexpected given previous results linking pitch range to the perception of surprise [14, 11]. So, the connection between pitch range and surprise may be more to do with slope rather than peak height or overall pitch range. Note, since the resynthesis was based on quantiles, we cannot really draw strong conclusions based on the individual contours across cue words. However, given that the stimuli were generated over 90% of the speaker's pitch range, the general within-word trends seem clear. Exploration of this is left for future work.

[2] `http://www-fis.iarc.fr/~martyn/software/jags/`

5. Discussion

5.1. The Interpretation of Rises

These results give us a window into how the uncertainty associated with rising intonation is interpreted in a dialogue. Credibility is clearly reflected in the choice of cue word. The fact that intonation did not have much of an effect on the credibility scale suggests that rises reflect difficulty integrating the new information proffered by the other speaker, rather than expressing some sort of disbelief about the content in general.

In terms of dialogue structures, a natural way to frame this integration difficulty is to say that rises signal that question under discussion is unresolved. That is, the speaker is unable to confirm or deny the at-issue content. By signaling that the QUD is unresolved, the speaker implicitly signals that resolution depends on the hearer. This turn passing behavior is congruent with the rising intonation of affirmative backchannels noted previously. This explains the association between rises and the expectation that more evidence will be presented. More generally, while rises are response seeking, they do not necessarily make an utterance an interrogative.

For cue words, the inability to resolve the QUD appear may happen when the utterance under scrutiny does not fit with the respondent beliefs. The content may be epistemically unexpected (i.e. it doesn't fit their world view). However, another possibility is that the content is unexpected from the point of view of relevance. This experiment did not differentiate these two cases. However, the latter case seems to apply to strong agreement words like *right* pronounced with a rise are interpreted. That is, the respondent may agree with the content, while still feeling that it does not resolve the current QUD.

5.2. Rises and cue word semantics

Although *right* has higher scores than *really* on all the scales, we still see similar distinctions between rising and falling contours in terms of EVIDENCE and EXPECTEDNESS scales. However, the interaction with rises sheds light on how cue words with more similar semantics vary in meaning. With respect to the affirmatives, we see that *yeah* appears to be able to express more unexpectedness then *right*. This seems to be attributable to the fact that *right* conveys that the respondent already believed the content at-issue (hence the high credibility scores associated with it). However, while *yeah* conventionally expresses agreement, it does not reveal so much about the respondents previous beliefs. Thus, *yeah* with rising intonation can be interpreted as conditional acceptance while simultaneously asking for more evidence. This sort request for more evidence would be pragmatically odd when the speaker is already known to believe the content, as would be the case for *right*. So, in a sense, prosody is able to influence the interpretation of *yeah* more than *right* because its semantics is not as specific.

A similar contrast is evident between *really* and *well*. The latter signals that downdate is not possible for that speaker given the current state of the dialogue. As part of this, *well* marks the QUD as unresolved and so the addition of the rise is redundant in that respect. On the other hand, while *really* does act as a check on the dialogue, its underlying question status passes responsibility for downdate back to the *really*-hearer. That is, *well* appears to be a stronger disaffirmative. However, like *yeah*, *really*'s semantics appear to allow for more shades of meaning. In summary, the interpretation of rising intonation with cue words depends on the (discourse) semantics of the cue word. In particular, how much this reveals about the speaker's beliefs.

6. Conclusion and Further Work

Speakers use various means to signal certainty and uncertainty in dialogue. This paper presented a study of two such means, intonation and cue words, in the hope that understanding their interaction would shed light on how both of these affect dialogue structure and maintenance. In general, understanding how these parts of dialogue compose has implications for both formal theories of dialogue and for determining how a dialogue system should respond to such cues of speaker attitude.

The results of the perception experiment suggests that rising intonation does not cast uncertainty on the credibility of the content the cue word is directed at. Rather, such rises signal the QUD is unresolved, hence the other participants should provide further evidence/explanation of the current claim. Cue words themselves vary on the credibility scale. That is, the meaning of rises/falls reveals the status of the QUD, while cue words signal the level of belief of respondent. An inability to downdate the QUD could be due to epistemic clashes or misunderstandings about an utterance's relevance. Further work will focus on teasing apart these two sources of uncertainty, the effect of varying pitch slopes, as well as the effect of prosodic cues in contexts.

7. References

[1] M. Nilsenova, "Rises and falls. studies in the semantics and pragmatics of intonation," Ph.D. dissertation, University of Amsterdam, 2006.

[2] J. Liscombe, J. Hirschberg, and J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[3] C. Gussenhoven, "Intonation and interpretation: phonetics and phonology," in *Proceedings of the Speech Prosody'02*, 2002, pp. 47–57.

[4] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech'08*, 2008.

[5] D. Litman, M. Rotaru, and G. Nicholas, "Classifying Turn-Level Uncertainty Using Word-Level Prosody," in *Proceedings of Interspeech'09*, 2009.

[6] D. Farkas and K. Bruce, "On Reacting to Assertions and Polar Questions," *Journal of Semantics*, 2009.

[7] A. Gravano, S. Benus, J. Hirschberg, E. S. German, and G. Ward, "The effect of prosody and semantic modality on the assessment of speaker certainty," in *Proceedings of 4th Speech Prosody Conference, Campinas, Brazil*, 2008.

[8] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of ICPhS 2007*, 2007, pp. 1065–1068.

[9] N. G. Ward and R. Escalante-Ruiz, "Using Subtle Prosodic Variation to Acknowledge the User's Current State," in *Proceedings of Interspeech'09*, 2009.

[10] D. Schiffrin, *Discourse markers*. Cambridge Univ Pr, 1988.

[11] C. Lai, "Perceiving Surprise on Cue Words: Prosody and Semantics Interact on *Right* and *Really*," in *Proceedings of Interspeech'09*, 2009.

[12] A. Gravano, S. Benus, H. Chavez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of okay," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 2007, pp. 800–807.

[13] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge, 2007.

[14] C. Gussenhoven, "Intonation and Interpretation: Phonetics and Phonology," in *Speech Prosody 2002, International Conference*. ISCA, 2002.