Perceiving Surprise on Cue Words: Prosody and Semantics Interact on *Right* and *Really*

Catherine Lai

Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA

laic@ling.upenn.edu

Abstract

Cue words in dialogue have different interpretations depending context and prosody. This paper presents a corpus study and perception experiment investigating when prosody causes *right* and *really* to be perceived as questioning or expressing surprise. Pitch range is found to be the best cue for surprise. This extends to the question rating for *really* but not for *right*. In fact, prosody appears to interact with semantics so ratings differ for these two types of cue word even when prosodic features are similar. So, different semantics appears to result in different surprise/question rating thresholds.

Index Terms: prosody, perception, dialogue, pragmatics, semantics, turn-taking, cue words.

1. Introduction

Dialogue is filled with cue words which are are used to maintain and further the discourse in many different ways. For example, backchannels like *uh-huh* are passive contributions that indicate, amongst other things, that the speaker is still participating in the dialogue, while affirmative words like *yeah* are generally used to signal that the speaker agrees with the information just presented to them by the other interlocutor. Words like *really*, however, seems to have multiple uses uses ranging from a backchannel-like use to something closer to an actual information seeking question or a repair request (see [1] for examples). In general, these words vary in their interpretation depending on the context and their prosodic manifestation [2, 3].

In this paper we investigate the contribution of prosody to the interpretation of *really* and *right*. These two words have very different semantics. On the one hand, short questions like 'Do you really?' suggests that *really*, as a one word turn, should be considered an interrogative which acts as a check on the common ground [4]. On the other hand, *right* expresses agreement with a previous assertion in the discourse. So, if prosody does contribute some compositional meaning we would expect this to interact with the semantics of these particles. Understanding this interaction would help predict the interpretation of other cue words produced in the same manner.

However, before we can understand how prosody interacts with the semantics and pragmatics of these particles we need to find appropriate response variables to associate with this prosodic variation. The prosody of *really* was investigated in [1] with respect to backchannel/question annotations. That study found the prosodic features considered (which included various pitch, intensity and duration features) could not be used to separate the data into these to these two categories. However, re-examination of the data suggests that backchannel/question was just not the right dimension upon which to study cue words removed from context. In fact, it appears that the prosody of isolated *really*s express how surprising the new information is to the speaker. This information would then need to be combined with contextual factors before we can predict the use of these cue words in a proper dialogue.

To this effect, this paper presents an experiment investigating the perception of *really* and *right* in terms of how surprising and how much like a question they sound. Section 2 presents a new corpus study aimed at finding how the acoustic properties of *really* are linked to the notion of surprise. Section 3 presents the perception study informed by that corpus study. From this study we find pitch range to be the feature best correlated with perceived surprise. We also find the question and surprise ratings to be correlated. These results are discussed in Section 4 and Section 5 concludes.

2. Corpus Study

2.1. Data

The data for this study were taken from the MDE 2003 annotations (LDC2004T12) and audio (LDC2004S08) of the Switchboard I corpus. 307 *really*s were analyzed after removing samples that were truncated or unintelligible.

F0 values were extracted using the method described in [5][1]. This method involved manual alignment of glottal pulses, trimming and smoothing of the F0 contour. *Syllable boundaries* were manually annotated using `praat`. F0 *slope* was extracted via linear regression on the F0 data. *Pitch range* in semitones and *duration* were also noted. These measurement were also taken for each syllable. 20 turns from each *really* speaker were taken (from the same conversation) to provide normalization data. Pitch measurements at the 1st and 99th quantiles averaged to approximate pitch range extremes for each speaker (*nmin, nmax*). *Pitch level* was approximated as the *(pmax - nmin)/(nmax - nmin)* where *pmax* was the maximum F0 value of the *really*. Each *really* was also categorized according to whether the speaker sounded like they had just heard something:

(a0) unsurprising.
(a1) new but not particularly unlikely.
(a2) new and undesirable.
(a3) highly improbable but not contradictory.
(r0) contradictory to their beliefs.

Annotation was done by the author without reference to the transcript. The first four of these categories can be classed as acknowledgements while *r0* is a possible repair request. Category *a2* reflects desired state bias: situations where information is not necessarily unlikely but simply undesirable to the listener. This category was only selected twice and while this sort of bias

[1] `http://www.phon.ucl.ac.uk/home/yi/tools.html`

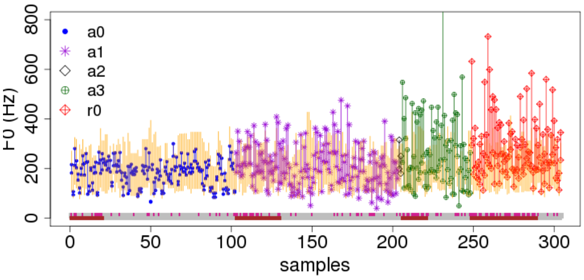

Figure 1: Really *pitch range overlaid on normalization data, grouped by surprise category. The bars at the bottom indicate the presence of an affirmative response in the transcript, and whether it was marked as a question in the MDE annotation.*

is clearly present in dialogue we consider it out of the scope of this paper.

2.2. Data Exploration

The data annotation revealed two basic types of contours: a contour with peak in the first syllable, thereafter a fall or levelling; otherwise, a continuous rise that may plateau at the end. In general, the pitch contours appear to be expanded or compressed in terms pitch range. They also appear to be translatable to different pitch levels and scaled in terms of duration.

The pitch variation in the *really* data is shown in Figure 1. This figure shows pitch range (Hz) of the *really*s overlaid on the normalization data (average minimum to average maximum). The data are sorted grouped according to the surprise categories described above. This shows the range of variation with respect to pitch range and level. This supports the view that prosodic variance in the data is better associated with surprise level rather the the backchannel/question distinction. This also fits the idea that utterances with expanded pitch range should sound more surprised (c.f. the *effort code* [6]) and pitch range is the appropriate response variable for surprise and similar affective states [6, 7, 8]. Similarly, compression of these two contours converges into a flat contour which seems to signal a lack of surprise or a withdrawal of information. Flattening of the final rise may diminish the questioning/uncertainty signal of the *really*. This makes it a better candidate for interpretation as a backchannel. However, this should not be a necessary or sufficient condition because backchannel interpretation still depends on the context and the hearer's model of the dialogue.

Since the categorization was done by only the author, it needs to be as very unrobust. Also, impressionistically it seems that looking at a scale of surprise may be more useful for our end goal. We hypothesize that listeners should be able to rate how surprised and questioning a cue word, like *really*, sounds in isolation while a backchannel interpretation requires the hearer to add these ratings to other contextual factors. The next section looks at the isolated condition.

3. Perception Experiment

A perception experiment was carried out to explore the interpretation of isolated *really*s and *right*s. We wanted to know whether native speakers could reliably rate how surprised these utterance sounded, how much they sounded like a question, and whether these two things are at all orthogonal. We also, we wanted to see if these ratings could be aligned with the MDE question/backchannel annotation.

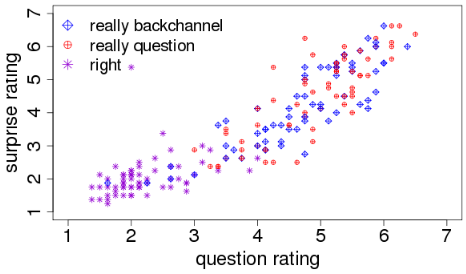

Figure 2: *Average surprise versus question ratings.*

3.1. Experiment Design

The stimuli were selected from the MDE data set analyzed in the corpus study discussed above. The stimuli set consisted of 192 tokens built from three subsets: backchannel *really*s (MDE labelled), question *really*s, and *right*s. Each subset consisted of 64 tokens selected to represent the different features. The data set was split based on quantiles. The data was first split into four groups according to *pitch range*. These four groups were then split similarly further according to *pitch level* (similarly *duration*). One stimuli was then randomly selected from each of the the 64 groupings. 5 female and 3 male University of Pennsylvania affiliates participated in this study. All subjects were native speakers of English and they had an average age of 23 years. The subjects were paid to partipate in this experiment.

The randomized stimuli were presented via a computer interface. The subjects listened to each stimuli through headphones and were allowed to replay the the current stimuli as many times as they liked. Subjects were asked two questions with respect to each stimuli: 'How surprised does the speaker sound?' and 'How much like a real question does this sound like?'. They were then directed to answer these questions on two 7 point sliding scales (1=not at all, 7=extremely). It was decided that a numeric scale would be more reliable than the subjective categorization used in the corpus study. The subjects were given a chance to ask questions and confirmed that they understood the task.

3.2. Results

The average rating for surprise versus question for the stimuli are shown in Figure 2. This shows the correlation between these two ratings (Kendall's $\tau = 0.63, p < 0.001$, distributions are non-normal). The figure also shows a lack of association between either rating and the MDE backchannel/question annotation. The ratings for backchannel/question categories are not significantly different (Mann-Whitney U test: question $p = 0.30$, surprise $p = 0.18$).

Generally, subjects appeared to find the lexical constraint quite strong. None of the *right* stimuli had an average question rating above 4 (the midpoint on the scale). Thus, lexical constraints interacted with how the prosodic cues were interpreted. Along the same lines, subjects did not behave completely uniformly in rating the stimuli. Subject variation and prosodic fea-

Table 1: Correlation coefficient (Kendall's τ) and p-values of the question/surprise ratings and prosodic features for *really* (top) and *right*

Really	τ_q	p-value	τ_s	p-value
pitch range	0.533	0.000	0.581	0.000
pr1	0.339	0.000	0.426	0.000
pr2	0.451	0.000	0.497	0.000
pitch level	0.414	0.000	0.502	0.000
slope	0.172	0.005	0.161	0.008
slope1	0.428	0.000	0.504	0.000
slope2	0.005	0.931	-0.035	0.567
duration	0.285	0.000	0.254	0.000
d1	0.216	0.000	0.230	0.000
d2	0.278	0.000	0.225	0.000
intensity	0.130	0.033	0.272	0.000
Right				
pitch range	0.240	0.007	0.285	0.001
pitch level	0.111	0.210	0.278	0.002
slope	0.234	0.008	0.093	0.299
duration	0.162	0.066	0.154	0.084
intensity	0.198	0.025	0.374	0.000

tures are discussed in the following subsections.

3.3. Prosodic Features

Table 1 show the correlation (Kendall's τ) between the ratings and various prosodic features for the *really*s (overall and by syllable) and the *right*s. We find that the question/surprise ratings to be most highly correlated with pitch range and pitch level. This is, again, inline with the idea that more effortful prosody results in more perception of surprise.

Somewhat unexpectedly, the second syllable slope of *really* was not significantly correlated with either questioning or surprise while the first syllable slope is correlated with the ratings. That is, the final fall/rise does not seem associated with whether the *really* is interpreted as a question. This is not to say that final rise/fall never contributes to question interpretation. However, for *really*, it does not add much to this dimension in the face of expanded pitch range. It is also interesting to note that intensity does not strongly correlate with the ratings. So, the surprise/question value of *really* seems primarily signalled by the size of the pitch excursion in the stressed syllable. Also, interestingly, first syllable pitch range is not highly correlated with that syllable's duration ($\tau = 0.19, p = 0.001$).

In general, it seems the mapping from prosodic feature values to surprise/question levels do not match between *really* and *right*. For example, the mean question rating for *really*s with pitch range between 5 and 10 semitones is 4.93, while the corresponding value for *right*s is only 2.41. However, the data shows that these 'barriers' can still in fact be passed for *right*. One *right* stimulus received a relatively high average surprise rating of 5.375. In fact, this stimulus had the highest ranking pitch range, pitch level and duration of the *right* set (5.68, 3.08 and 2.49 standard deviations from respective means). Note, however, that the same stimulus still received a low question rating (2.0). So, for *right*, expanded pitch range, level and duration appears to contribute to the perception of surprise independent of questionhood.

Subject	mean$_q$	sd$_q$	mean$_s$	sd$_s$
1	3.49	1.93	2.97	1.83
2	4.39	1.85	3.98	1.73
3	4.14	1.11	4.06	1.56
4	2.39	1.60	2.11	1.32
5	4.64	1.86	3.88	1.85
6	4.31	1.84	3.91	1.76
7	4.11	2.54	3.52	2.40
8	3.94	1.64	4.19	1.57

Table 2: *Means and standard deviations for question and surprise ratings, by subject*

	1	2	3	4	5	6	7
2	$***$						
3	0.18	0.16					
4	$***$	$***$	$***$				
5	$***$	1.00	$**$	$***$			
6	$***$	1.00	0.81	$***$	1.00		
7	0.12	1.00	1.00	$***$	1.00	1.00	
8	0.74	0.09	1.00	$***$	$**$	0.35	1.00

Table 3: *Pairwise Mann-Whitney U tests for question rating by subject with Bonferroni correction (** = $p < 0.01$, *** = $p < 0.001$).*

This leads to the question of whether rising intonation can be interpreted as questioning with an agreement particle like *right*. Of the 64 *right* stimuli, 33 had F0 contours with overall rising slope. Athough, the stimuli with average question rating greater than 3 did have positive slope (6/33 items), most rising *right*s did not sound questioning to the subjects. Inspection of one pair of stimuli with similar features, shows that closer fitting of F0 than simple linear regression may be necessary to sort this out. For example, one the lower rated item had a perceptible final fall even though the general trend was positive. We will return to question of rises and *right* in the Section 4.

3.4. Subject Variation

We also need to consider how well these results apply to the range of subjects. Using Krippendorff's α for ordinal data, we see that agreement between raters was above chance but still not extremely high ($\alpha_s = 0.58$, $\alpha_q = 0.50$). Closer examination of the data suggests that subjects had different rating biases. Table 2 shows means and standard deviations of the ratings by subject. The responses of subjects 1 and 4, in particular, appear significantly lower than the other participants. This is confirmed via pairwise Mann-Whitney U tests (Table 3).

However, these subjects did not simply avoid the higher end of the scale. Figure 3 shows the distribution of pitch range versus first syllable slope for the *really* data used in the experiment. It also highlights the stimuli which were high (> 5) and low (< 3) rated as questions by subject 4. In this case, the ratings seem based on the same prosodic inputs as for the other subjects but the rating cutoff points appear to be higher.

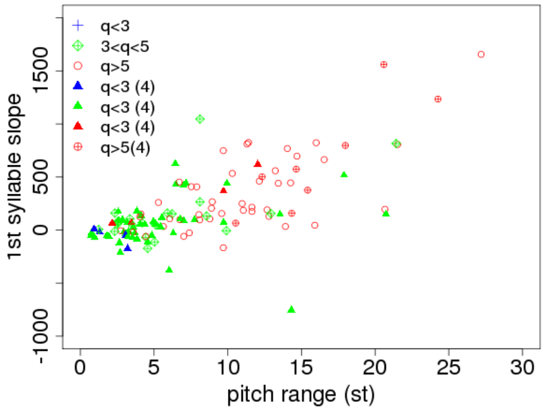

Figure 3: *Pitch range versus first syllable slope. The colors indicate different average ratings. Points highlight the ratings of our most conservative subject (4).*

4. Discussion

Really and *right* appear to induce different mappings from prosodic features to surprise/question ratings. That is, prosodic features do interact with semantic content. It is plausible that surprise features have the effect of intensifying what is already there. So, the addition of surprise features, like expanded pitch range, appears to intensify the existing interrogative semantics of *really*.

In fact, the interrogative semantics of *really* appears to trump its final fall/rise characteristics in terms of how questioning the utterance is perceived. Hence, second syllable slope was not correlated with the question rating. However, impressionistically, a high second syllable plateau can make the *really* sound more skeptical along the lines of the 'repair' category mentioned in Section 2. This sort of skepticism is not necessary in order to get a questioning interpretation and so cannot be addressed by the results of this particular experiment. However, this is certainly dimension that needs to be considered in the interpretation of this cue word.

Unlike *really*, *right* does not have anything interrogative in its semantics. So expanded pitch features push the interpretation in a more exclamative direction, although we have seen that the thresholds for high surprise ratings seem higher for *right* than for *really*. However, *right* can take on an interrogative meaning, as, for example, a tag question with rising intonation. We also saw that the rising *right*s in our data set contained the more questioning stimuli. In a sense *right* is more sensitive to a final rise than *really*. However, a final rise has been argued to be a marker of speaker uncertainty [9]. Since *right* is generally used to agree with a previous utterance, adding uncertainty about that agreement is difficult without the right contextual conditions. A case where this may be possible is when the speaker agrees with the information presented but is uncertain about how it relates to the question under discussion. Similarly to *really*, this may signal that a repair is necessary in the dialogue. A detailed study of the prosody of repairs and cue words is left for future work.

5. Conclusions and Future Work

This paper investigated how prosody affects the perception of two cue words with very distinct meanings. The perception experiment show that effortful prosodic features, like expanded pitch range, cue an interpretation of surprise. Moreover, perceived level of surprise was correlated to whether the stimulus was perceived as questioning. However, the ratings achieved by *really* did not translate to similar ratings on the agreement marker *right* for similar prosodic values. That is, the semantics/pragmatics of the cue word appeared to change the thresholds with which a cue word would be considered surprising. In fact, without an underlyingly questioning semantics, it seems that surprise prosody on its own will not lead to the perception of questioning. Instead this seems to be something closer to hearing an exclamative.

Clearly, the speaker affect that is perceived on the isolated cue words needs to be integrated with contextual cues before we can make a properly predictive model of the interpretation of these cue words. In particular we need to further investigate the prosody of repairs and how expressions of uncertainty are interpreted on agreement cue words. It seems that the fine detail of final rise/fall and voice quality [10] have an important part to play here.

6. Acknowledgements

Thanks to the members of the Penn Phonetics lab/splunch, especially Jiahong Yuan, for many helpful comments.

7. References

[1] C. Lai, "Prosodic Cues for Backchannels and Short Questions: *Really?*" in *Proceedings of Speech Prosody 2008, Campinas, Brazil, May 2008*, 2008.

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

[3] 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.

[4] M. Romero, "Biased Yes/No Questions: The Role of VERUM." *SPRACHE UND DATENVERARBEITUNG*, vol. 30, no. 1, p. 9, 2006.

[5] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, pp. 55–105, 1999.

[6] C. Gussenhoven and T. Rietveld, "The Behavior of H and L Under Variations in Pitch Range in Dutch Rising Contours," *Language and Speech*, vol. 43, no. 2, pp. 183–203, 2000.

[7] A. Shimojima, Y. Katagiri, H. Koiso, and M. Swerts, "Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses," *Speech Communication*, vol. 36, no. 1-2, pp. 113–132, 2002.

[8] L. Yang and N. Campbell, "Linking Form to Meaning: The Expression and Recognition of Emotions Through Prosody," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. ISCA, 2001.

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

[10] C. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality," *Speech Communication*, vol. 50, no. 6, pp. 531–543, 2008.