Prosodic cues for backchannels and short questions: Really?

Prosodic Cues for Backchannels and Short Questions: *Really?*

Catherine Lai

Department of Linguistics
University of Pennsylvania, Philadelphia, United States
laic@babel.ling.upenn.edu

Abstract

Short questions can be ambiguous even after considering their preceding contexts. Hence, prosody may be useful for disambiguating different types of questions and their uses. For example, question bias has been linked to the presence of certain pitch accents. This paper presents a corpus study of very short questions and the contribution of prosodic cues to discourse disambiguation. This study focuses on backchannel questions which are by nature highly biased and yet sit between genuine questions and genuine backchannels. The study finds LDA and SVM classifiers do not perform better than random at separating backchannel and question *really* based on these prosodic cues. This means that, while intonation differs between broad categories of questions, theories that try to integrate prosodic cues with semantics and discourse require more than intonation, the final rise and the other usual prosodic suspects like duration and intensity.

1. Introduction

A key factor for question interpretation and detection is the match of form and its intonation. Prototypical question types: wh, yes/no and declarative questions, have been characterized as ending in a final fall, final rise, and a higher final rise respectively [1]. However, linking intonation to question type is much harder if we consider that there is some gradience in what we mean to be a question.

There are clearly more types of questions than the three listed above. Bolinger [2] distinguishes alternative, tag, and reclamatory questions to name a few. Moreover, interrogative forms have many other uses. For example, the broad set of questions can be used to make assertions, clarifications, acknowledge turn control, or to express agreement. We can consider a question to be *genuine* if the speaker does not know the answer. That is, the speaker is seeking information from the hearer. However, the line between information seeking and non-information seeking questions is vague. Even if a question seeks information, there can be ambiguity in what type of information is being sought. In such situations, it has been argued prosodic factors cue the interpretation a question receives and what response is required from the hearer [2, 3, 4].

This paper investigates the connection between prosodic cues such as pitch, intensity and duration, and the interpretation of very short questions (two words or less). Even with previous context, the interpretation and discourse function of these questions may be ambiguous when stripped of sound. However, it is not clear what role prosodic cues play or which prosodic cues are salient. In particular, whether or not so much meaning can be attributed to pitch movement alone. A corpus study of the word *really* was carried out investigate this. *Really* appears in both backchannel and question classes. However, it is not lexically or syntactically marked as either. The result of this study was that something more than standard intensity, duration and pitch movement is required to differentiate these two categories.

The paper is organized as follows. Section 3 describes an investigation of final rises and question types in the corpus data. Section 4 describes attempts to separate backchannel and question *really* with an extended data set and more prosodic features. The next section, however, presents some examples of ambiguous short questions in the light of theories linking intonation and question meaning.

2. Intonation and Short Questions

Interrogatives are often elided in natural speech. This leads to a natural ambiguity in their meaning and use. For example, elided wh-question can be used to seek information (e.g. *What (did you do)?*) or to elicit repetitions (e.g. *What (did you just say)?*). These uses multiply when we consider that questions are not always used to elicit information. Utterances with polar interrogative syntax can be used as backchannels in dialogue [5]. That is, they can be used to express acknowledgement or agreement, as continuers or to mark incipient speakership. However, the short auxiliary-subject questions can also convey surprise, disbelief, or more generally question the truth of the previous utterance. That is, these questions are highly biased. In which case, it is not clear whether they are really questions at all.

In a similar way, *really* is used in dialogue as both a backchannel or as something more like a question. In the following dialogue (1), *really* appears to be a *question* (speaker B justifies his question in the last line).

1. B : You like Lubbock better than Dallas
   A : Yeah
   B : Why?
   A : Uh, because people are so much nicer
   B : Really?
   A : Yes
   B : Well people are nice here in Dallas

This contrasts with the following dialogue where *really* was annotated as a *backchannel* rather than a question.

2. B: Oh I've got some Chinese Hollies
      that are just outrageous
   B: They they are very sharp
   A: Oh really
   B: Do you do your own uh lawn maintenance?
   A: Yeah

Speaker A's *really* did not require nor elicit a response from the speaker B. However, although *really* in (3) was marked as a question it is very similar to a backchannel. Even though it apparently required a response (*Yes*), there did not appear to be any need for the speaker to justify their statements any further.

3. B: I kind of enjoyed that boat

```
   I looked at today
B: It's nice and clean
A: Really?
B: It wasn't [interrupted]
B: Yeah
A: Did it have a cabin?
```

Really is used as both a question and backchannel, where questionhood clearly involves some gradience. If *really* is interpreted as a question it must be a highly biased one, since the speaker has the answer from the previous utterance. This sort of bias and the question status of various types of utterances has mainly been investigated in the semantics and pragmatics literature in terms of an intonational lexicon and how this interacts with other linguistic structure. Interpretations of the final rise generally revolve around lack of speaker certainty or commitment to the utterance at hand [6, 7, 8].

By treating rising intonation as carrying its own meaning, Gunlogson [4] is able to analyze rising declaratives as assertions. So, rising and falling declaratives differ in terms of speaker commitment. The latter is a typical assertion that commits the speaker, while the former is like a question in that it commits the addressee to the propositional content of the utterance. In the same manner, rising intonation is treated as an intonational adverb expressing uncertainty by Nilsenova [3]. Both the observed bias and questioning aspects of a rising declarative are then derived as by-products of pragmatic principles.

Reese [9] argues that the outer negation [10] (negatively biased) interpretation of a negative polar interrogative can be triggered by the presence of an L*+H nuclear pitch accent. This allows him to present a unified theory of negative bias in negative polar interrogatives and emphatic focus questions. The contribution of the pitch accent is that of metalinguistic negation: such questions express denial or counterevidence to something in the (immediately) preceding discourse. This contrasts with the inner negation readings which seem to be restricted to checks or confirmations of the common ground.

These proposed links between speaker uncertainty, bias and question intonation predicts final falls/rises should help in determining question status and subtype. In fact, Liscombe *et al*. [11] found that the presence of a final rise in particular to be the most useful cue for a question bearing turn in a student/computer tutor scenario – although additional intensity and timing information also improved performance. Intonation has also been found to differentiate uses of affirmative backchannels like *okay* [12].

One might expect a similar prosodic distinctions to separate backchannel *really* from question *really*. Moreover, we might expect these distinctions to fall into line with the semantic accounts above, particularly Reese's. Intuitively, backchannels act as a confirmation or acknowledgement of the common ground. Question *really?*, on the other hand, seems to indicate that previous utterances were at least very unlikely according to the speakers previous beliefs, as in the denial reading of outer negation polar interrogatives. The differentiability of backchannel and question really is investigated in Section 4. The next section presents a corpus study examining the distribution of final rises/falls in short questions.

3. Final Rises in Short Questions

3.1. Data

This study used the Conversational Telephone Speech component of the MDE RT-04 corpus (LDC2005S16). This comprises of approximately 40 hours of speech from the Switchboard-1 Corpus Release 2. The MDE annotation (LDC2005T24) provides discourse metadata including question and backchannel type turns. 315 questions turns containing two or less words were located.

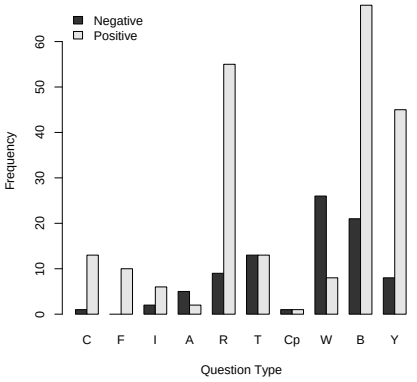

Figure 1: *Proportions different question types with negative and positive slopes in different question types*

Final word F0 contours were extracted using `praat`. F0 values were normalized to a log scale between 0-10 after outlier removal. Outliers were values that fell 1.5 times further away from the mean than the first and third quantiles. The slope for each final word was fitted from the normalized F0 data using the linear regression function `lm` in R. Normalization eliminated five questions, two questions were eliminated due to lack of speaker information, while a further question was removed as its transcription did not match the audio recording, leaving 307 questions. Figure 1 shows proportions of question types.

Genuine yes/no questions (Y) and wh-questions (W) were manually identified along with a number of other question types. Reclamatory (R) questions elicited repetitions of the previous utterance. Confirmation (C) questions clarified the current topic of discourse. Incomplete (I) questions attempted to elicit non-specific speech from the hearer (e.g. *Hello?*). In tag questions (T), speakers questioned their own prior statement. Backchannel questions were utterances were the speaker questioned an immediately prior statement of the hearer. Speakers offered possible but indefinite options to the hearer in suggestion (S) questions. Complementary (Cp) questions elicit the same responses as wh-questions but without the wh-word. Alternative (A) questions present a list of possible alternatives.

3.2. Observations

Although the amount of data is relatively small, we can make some general observations. Yes/no questions do appear to have final rises while wh-questions have final falls in the majority of cases. Confirmation, incomplete, and suggestion questions tend to end in a rise. These questions are really declarative questions so this result is inline with Haan's functional hypothesis. The ambivalence associated with these types of questions also agrees with the association of rising and speaker uncertainty. However, it appears that wh-reclamatories are distinguished from genuine wh-questions by their final rise.

This categorization shows that broad categories of ques-

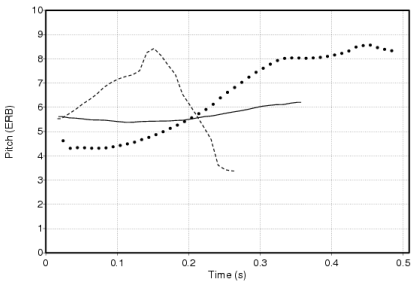

Figure 2: *Two types of* really*: dashed, speckled lines are surprised, the solid line is a backchannel.*

tions do have different final rise characteristics. However, it is not at all clear that pitch movement can make finer grained distinctions. In particular, it is not clear what the presence of a final rise means for disambiguating the different uses of backchannel like questions discussed above. Backchannel questions were used to convey a range of signals from acknowledgement to denial and surprise and with varying pitch curves. Figure 2 shows pitch tracks for examples (1) and (2) and one additional 'surprised' *really*. It seems plausible that pitch movement could be an indicator of deviation backchannel status alongside a number of other prosodic factors. However, the following study suggest this is not the case.

4. Prosodically Distinguishing Backchannel and Question *Really*

This section argues that the usual suspects of prosody – intonation, duration and intensity – do not provide useful cues for distinguishing backchannel questions from pure backchannels. The second subsection explores the prosodic differences, and the final subsection tests whether these differences provide useful cues.

4.1. Data

These experiments expanded the previous data set to include MDE 2003 annotations (LDC2004T12) and audio (LDC2004S08) from the Switchboard I. Instances of *really* labelled as a backchannel (450) or a question (130) ($really_q$, $really_b$ in the following) were extracted using timing information from the MDE annotations.

Pitch and intensity measurements were made using `Praat` with samples at 1ms intervals (to provide enough points for curve fitting). The pitch data was normalized to a log scale from 0-10 as previously. The mean intensity of the speaker for their entire conversation was subtracted from the intensity measurements. Pitch and intensity curves were approximated using orthogonal polynomial curve fitting with order 5 Legendre polynomials (c.f. [13]). Six coefficients were recorded for each pitch and intensity curve (`p0-p5`, `i0-i5` resp.). Legendre polynomial fitting has the nice property that coefficients derived from this process are not fraught with the correlation problems of those from standard polynomial fitting. They also have intuitive interpretations: the first coefficient indicates general bias, the second indicates overall slope, the third indicates convexity and so on. Information about the utterance may be signalled by from non-linear characteristics of pitch and intensity the curve, such as convexity, so this is a desirable property.

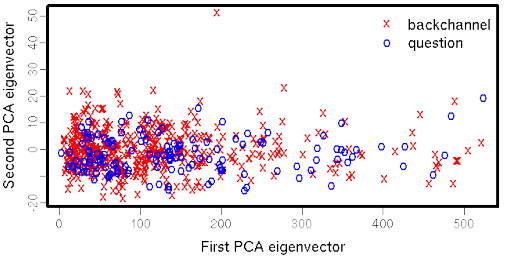

Figure 3: *Projection on to the first two dimension of the PCA space.*

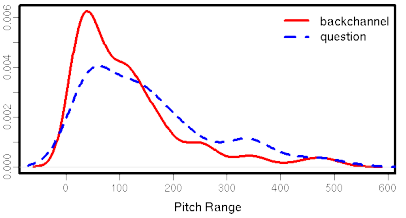

Figure 4: *Probability densities for the* `prange` *feature.*

Beyond this, the correlation between raw intensity and pitch at 10ms intervals was also derived (`Corr`) for each utterance, as well overall pitch range (`prange`). Duration (`Dur`) and relative time position of pitch minimum (`p.min.d`) and maximum (`p.max.d`) were also recorded.

4.2. Exploring the Differences

Principal components analysis was carried out on all the numeric features using the R function `prcomp`. The principal component with the largest standard deviation (109.4) was dominated by `prange` The second component (standard deviation, 7.7) points predominantly in the direction of `i0` (intensity bias) and `i1` (linear coefficient of intensity). However, it does not appear that these components differentiate backchannel and question *really*. This can be seen from Figure 3 which shows the overlapping distribution of the data transformed to the space spanned by the principal components and then projected onto the first two components. Analysis of this data feature by feature suggests that there are differences in how $really_q$ and $really_b$ are produced. However, the amount of distributional overlap leaves the hypothesis that listener can actually differ-

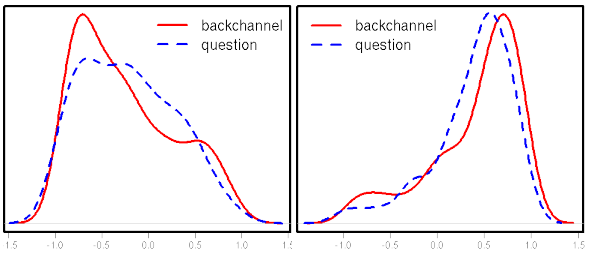

Figure 5: *Empirical probability densities of pitch/intensity correlation for rising (left) and falling (right) pitch.*

	Error (Std. Error)	95% CI
Baseline	0.245 (0.018)	(0.21,0.28)
LDA	0.244 (0.019)	(0.21,0.28)
SVM	0.267 (0.019)	(0.23,0.30)

Table 1: *Estimates for classification errors and 95% confidence intervals from bootstrapped bias-corrected cross-validation.*

entiate the data based on these cues, somewhat weak. For example, density plots for the pitch range data are shown in Figure 4. The non-parametric bootstrap was employed to see if the means of these two distributions differed as the data is clearly not normally distributed. Both sample means (115Hz, 154Hz) fall outside the 99% bootstrap confidence intervals for the other class ($(102.7, 128.3)$ and $(124.3, 184.4)$ resp.). However, it is clear from the density plot that the overlap in distributions is great.

In a similar vein, Figure 5 shows empirical density plots for the pitch/intensity correlation data. These are separated according to whether pitch rising or falling (`p1`) to account for the fact that intensity generally falls at the end of an utterance. This plot indicates that the distributions for rising pitch backchannels and questions are actually quite different. That is, a speaker is more likely to maintain intensity when producing *really*$_q$ than *really*$_b$. However, once again there is almost total overlap in the distributions, so it is unlikely that listeners use this cue to determine that a given *really* is a backchannel or question.

4.3. Testing Prosodic Cues

The overlapping distributional data above suggests that it unlikely that the prosodic features described above can differentiate *really*$_q$ and *really*$_b$. To further test this hypothesis, two classifiers were built in an attempt to separate the data. The first was a classifier based on Linear Discriminant Analysis (LDA) as implemented in R (`lda`). The second was a Support Vector Machine (SVM) classifier with radial basis function kernel (`libsvm` via R) . The 10-fold cross-validation error rates are shown in Table 1 alongside bootstrap estimates, standard error, and 95% confidence intervals (1000 bootstrap samples).

The classifiers clearly do not fare much better than the baseline that simply categorizes everything as a backchannel. Indeed, the SVM classifier appears to do worse! They certainly do not reach error rates outside the 95% confidence interval for the cross-validation error of the baseline. This supports the hypothesis that these two categories are not separable on the basis of these features.

5. Conclusion

Short questions can be ambiguous in way that is not always resolvable from the previous context. This suggests that prosody plays a part in disambiguating uses of different question types. As we saw, backchannel questions can be used both as an acknowledgement and as real questions. They also project many shades of meaning in between. However, they all seem to be geared towards expressing speaker uncertainty of previous utterances in the discourse. Hence, backchannel questions form an interesting testing ground for theories positing an intonational lexicon and its ramifications for discourse.

The second part of this paper was an attempt to find out if some element of this intonational lexicon could systematically differentiate *really*$_q$ and *really*$_b$ similar to the way final rises change the interpretation of declarative sentences. It seems clear at this point that the prosodic features considered (including intonation) are not enough to make this distinction. This does not mean that there is no role for prosody in distinguishing the different uses of *really*. Of course, there may be latent cues that were not covered by this analysis. The results of this *really* study strongly suggest, however, that theories that try to integrate prosodic cues with semantics and discourse should go beyond intonation, the final rise and the other usual prosodic suspects: plain duration and intensity.

However, the fact that listeners can differentiate between *really*$_q$ and *really*$_b$ suggest many avenues for further study. In particular, further perception studies on *really* and other short questions with multiple uses will help tease out prosodic cues.

6. References

[1] J. Haan, "Speaking of Questions: An Exploration of Dutch Question Intonation," Ph.D. dissertation, Utrecht: LOT Graduate School of Linguistics, 2002.

[2] D. Bolinger, *Intonation and its uses : melody in grammar and discourse*. Stanford, Calif.: Stanford University Press, 1989.

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

[4] C. Gunlogson, "Declarative questions," in *Proceedings of Semantics and Linguistic Theory XII*, B. Jackson, Ed. CLC Publications, 2002.

[5] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, Prosodic, and Syntactic Cues for Dialog Acts," in *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, 1998, pp. 114–120.

[6] R. A. Hudson, "The meaning of questions," *Language*, vol. 51, no. 1, pp. 1–31, Mar 1975.

[7] M. Steedman, "Information Structure and the Syntax-Phonology Interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, September 2000.

[8] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgen, and M. Pollack, Eds. Cambridge: MIT Press, 1990.

[9] B. Reese, "Bias in questions," Ph.D. dissertation, University of Texas at Austin, 2007.

[10] D. Ladd, "A First Look at the Semantics and Pragmatics of Negative Questions and Tag Questions," in *CLS 17*, 1981, pp. 164–171.

[11] J. Liscombe, J. Venditti, and J. Hirschberg, "Detecting Question-Bearing Turns in Spoken Tutorial Dialogues," in *Interspeech'06*. ISCA, 2006.

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

[13] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *The Journal of the Acoustical Society of America*, vol. 118, p. 1038, 2005.