

# Dialogue with attitude: The contribution of cue words and prosodic meaning in conversational speech

Catherine Lai

Department of Linguistics  
University of Pennsylvania

April 2011



# Cue words

- ▶ Cue words are short responses which indicate how discourse structures are to be updated with respect to a new utterance.

(1) A: I'm going to California!

B: yeah / right / ok / sure / really

- ▶ They appear frequently in spontaneous dialogue.
- ▶ They help direct the discourse topic and signalling what's in the common ground.
- ▶ That is, they are an important part of how we keep track of what is going on!

# Cue Words and Prosody

- ▶ Prosodic variation affects cue word interpretation.
- ▶ This variation in interpretation is reflected in the different dialogue acts associated with them.
  - ▶ **really, really, really, really.**  
↔ backchannel or question?
  - ▶ **right, right, right, right.**  
↔ backchannel or agreement?
- ▶ To model how these turns are interpreted we need to know not only what dimensions of meaning prosody can work on, but also how this combines with the semantics of the cue word.

# Questions

- ▶ How does prosodic variation relate to the different interpretations associated with cue words.
- ▶ What response variables can we associate with this variation?
  - ▶ Are dialogue acts a good lens through which to study prosodic meaning?
- ▶ How do they affect discourse structures like the common ground and the questions under discussion?
- ▶ (How do we model this in semantic/pragmatic terms?)

# Why do you do what you do?

I will argue that...

- ▶ Cue words express speaker attitude towards (potential) additions to dialogue structures.
- ▶ Effortful prosody intensifies the underlying semantics of these words.
- ▶ Final rises signal that the current question under discussion is unresolved.

To model conversational dialogue we need to model speaker (propositional) attitude, not just their acts.

(But first we need to look at the data)

# Outline

- ▶ Background on cue words and prosodic meaning.
- ▶ Corpus study: *really* as a backchannel and question.
- ▶ Perception experiment: *really*, *right* and surprise.
- ▶ Perception experiment: Rises and uncertainty.
- ▶ Situating cue words and prosody in the dialogue model.
- ▶ Conclusion and outlook.

## Cue words in dialogue

To model dialogue we need to keep track of speaker's public beliefs, the common ground, and the Question Under Discussion (QUD), c.f. (Farkas and Bruce, 2009; Ginzburg, 2009).

- (2)
- a. B: Do you like Lubbock better than Dallas? ( $= ?p_1$ )
  - b. A: Yeah
  - c. B: Why?
  - d. A: Uh, because people are so much nicer ( $= p_2$ )

	Public(A)	QUD	Public(B)
(a)		$p_1?$	
(b)	$p_1$		
(c)		Why $p_1?$	
(d)	$p_2$	$p_2?$	

- e. (Switchboard Corpus: LDC2004T12)
- B: right
- B: yeah
- B: okay
- B: uh-huh
- B: really?
- B: well...
- B: No

(e) depends on the cue word semantics and prosody...

# The many meanings of cue words

- ▶ **Affirmatives:** non-information seeking, certainty,...
  - ▶ e.g. *right, yeah, okay, uh-huh.*
- ▶ **Questions:** information seeking/restructuring, uncertainty,...
  - ▶ e.g. *really?, huh?*
- ▶ **Backchannels:** non-information seeking,...
  - ▶ do not cause the other speaker to cede the floor and that are passive contributions to the discourse.
  - ▶ *uh-huh, okay, yeah, right, really,...*
- ▶ **Spectrum?**
  - ▶ Uncertainty, Questions ← Backchannels → Agreement, Certainty



# Really: Questions, Backchannels, Backchannel Questions

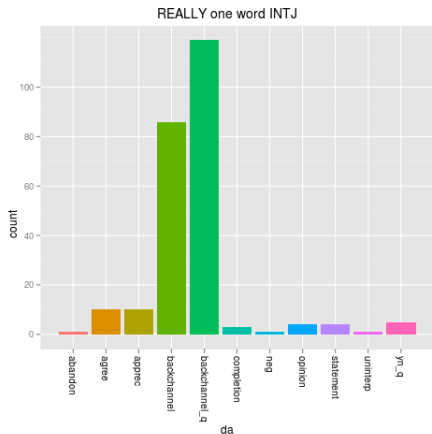
- ▶ Prosody has been found to be helpful for distinguishing different acts associated with affirmatives, e.g. agreement and backchannel (Jurafsky et al., 1998; Gravano, 2009).
- ▶ How about questions and backchannels? e.g. *really*
  - ▶ Is it underlyingly a question? Is it information seeking?
  - ▶ If not, what makes it a question? Prosody?
  - ▶ What does it mean for a question to be used as a backchannel anyway?

# Really: Questions, Backchannels, Backchannel Questions

**Data:** The Switchboard corpus: 2,400 telephone conversations (LDC97S62, Godfrey et al. (1992)), with many annotations including (shallow) discourse structure.

- ▶ SWB-DAMSL ( $\cap$  Treebank): 642 convs. (Jurafsky et al., 1997)
  - ▶ Queried via Switchboard NXT (Calhoun et al., 2010)
- ▶ EARS MDE (RT-03/04): 743 convs. (Strassel, 2003)

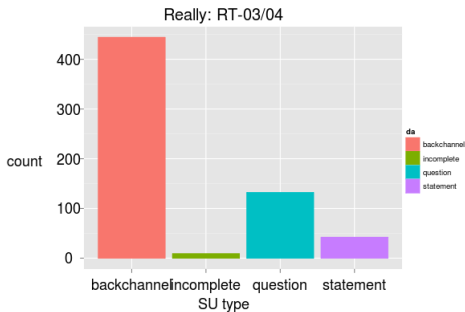
# Really: SWBD-DAMSL



Jurafsky et al. (1997):

- ▶ A backchannel question is ‘a continuer which takes the form of a question’.
- ▶ ‘Unlike rhetorical questions, backchannels lack semantic content’
- ▶ These are separated from the backchannel class ‘because we suspect that they will mess up the prosodic utterance detector’

## Really: RT-03/04 (MDE)



- ▶ RT-03 (Strassel, 2003): smaller set of SU types: statement, question, backchannel, incomplete.
  - ▶ 'Annotators should label only those cases in which these words are functioning in a way that is clearly recognizable as a backchannel.'

## *Really* as a question

An example of a real question? (Switchboard, RT-04, LDC2005T24):

- (3) 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

## *Really* as a backchannel

- (4) 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

Sounds like a difference in prosody...

# Prosody and Questions

Questions are often linked to rising intonation. This has been given various interpretations from a semantic/pragmatic point of view:

- ▶ Forward looking: Pierrehumbert and Hirschberg (1990).
- ▶ Hearer commitment: Steedman (2000), Gunlogson (2002).
- ▶ Contingency: Gunlogson (2008).
- ▶ Uncertainty: Nilsonova (2006), Reese (2007).

**Main theme:** Rising intonation is associated with speaker uncertainty or at least a lowered degree of speaker commitment.

# Prosody and Meaning: Empirically

The theories above are broadly convergent with phonetic studies when it comes to sentential utterances:

- ▶ Rises  $\mapsto$  questioning, uncertainty (Gravano et al., 2008)
- ▶ Higher pitch  $\mapsto$  surprise and questioning (Gussenhoven and Chen, 2000)
- ▶ Greater pitch excursions/delayed peak  $\mapsto$  surprise (Chen et al., 2004)
  - ▶ The effort code (Gussenhoven, 2004).
  - ▶ Flatter contours, i.e. compressed pitch range,  $\mapsto$  backchannels?
- ▶ Rising pitch, greater intensity  $\mapsto$  backchannel interpretations of affirmative cue words in task oriented dialogue. (Benus et al., 2007; Gravano, 2009).



## Prosodically Distinguishing *really<sub>b</sub>* and *really<sub>q</sub>*

Corpus study of *really*:

**Aim:** To determine whether prosodic features distinguish backchannel and question interpretations of *really*.

**Hypothesis:** Final rises signal a question interpretation.

- ▶ If not, can other features make this distinction?

**Data:** SUs = (*oh*) *really* labelled as a backchannel (444) or a question (132).

- ▶ MDE 2003/04 annotations (LDC2004T12, LDC2005T24) of Switchboard-1 Corpus Release 2 audio (LDC2004S08, LDC2005S16).

# Corpus Study: Data

## Data extraction:

- ▶ Timing data, e.g. syllable boundaries, was determined using the Penn Phonetics Lab forced aligner.
- ▶ F0 data were extracted via Praat (autocorrelation).
  - ▶ **Pre-processing:** Input parameter values for Praat were set based on estimated speaker pitch range (Evanini and Lai, 2010).
  - ▶ **Post-processing:** Conversion to semitones (based on speaker F0 median for the conversation) removal of implausible F0 islands, smoothing (Butterworth filter with a normalized cut off frequency of 0.1), interpolation over unvoiced regions (excluding detected pauses).
- ▶ Intensity data was also extracted via Praat and normalized by speaker to z-scores.

# Corpus Study: Data

## Word and syllable level features:

- ▶ Speaker normalized duration (z-score), speaking rate (z-score), raw duration (s).
- ▶ F0, Intensity: mean, standard deviation, slope (linear regression), max, min, range, number of points, 'jitter', relative time of max and min within the time unit.
- ▶ Pitch and intensity curve approximation: orthogonal polynomial curve fitting with order 5 Legendre polynomials giving 5 coefficients (c.f. Kochanski et al. (2005)).  
↪ look at the shape of the contour, i.e. is it peaky or flat?

## How does the data vary?

- ▶ At the word level, most features did show significant differences between *really<sub>b</sub>* and *really<sub>q</sub>*.
    - ▶ Means for duration/rate, F0, F0 jitter, intensity (t-test  $p < 0.01$ ), slope F0 ( $p < 0.05$ ), were higher for the questions.
    - ▶ Pitch range, F0 sd, Legendre coeff. 3 were not significantly different.
  - ▶ At the syllable level,
    - ▶ Both syllables: Mean F0, F0 jitter, mean intensity, intensity range, intensity sd, intensity slope were greater for questions ( $p < 0.01$ )
    - ▶ F0 slope was not significantly different for either syllable ( $p = 0.3667$ ,  $p = 0.0528$ )
    - ▶ Second syllable only: duration/rate was longer for questions ( $p < 0.01$ )
- ↪ *really<sub>q</sub>* can be longer, higher and have greater intensity than *really<sub>b</sub>*.

# Word F0: Mean and Slope

Some differences, but mostly a lot of overlap!

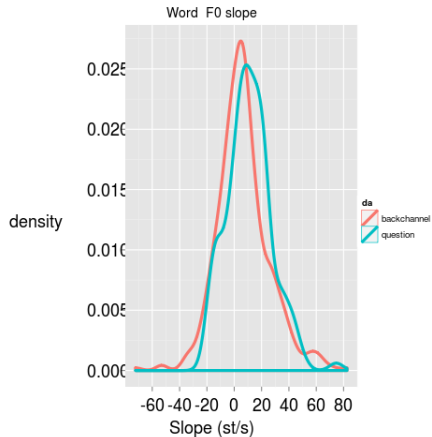


Figure: *F0 Slope*

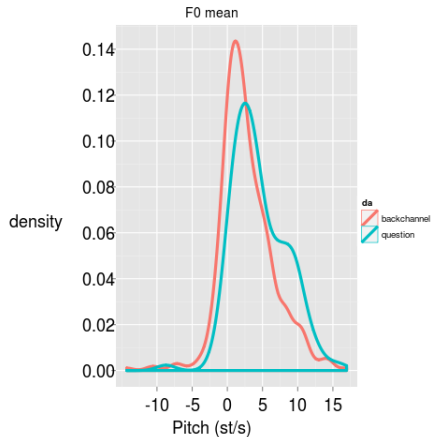


Figure: *F0 Mean*

# Word F0: Mean and Slope

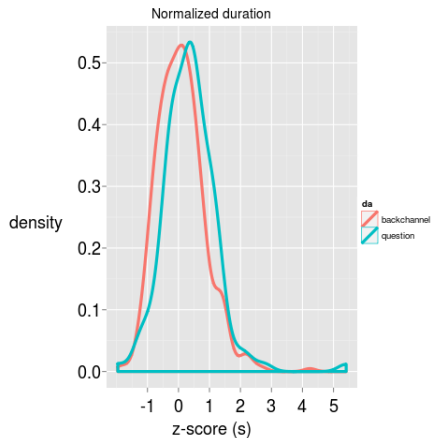


Figure: *Normalized duration*

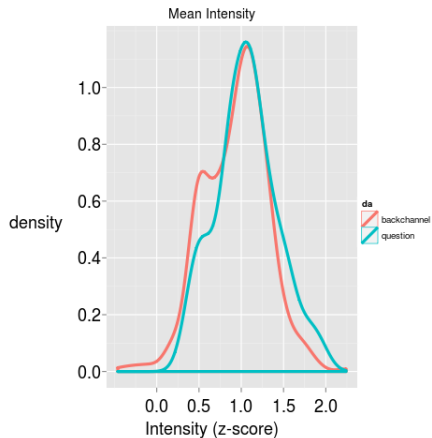


Figure: *Mean Intensity*

# Principal Components Analysis: Word

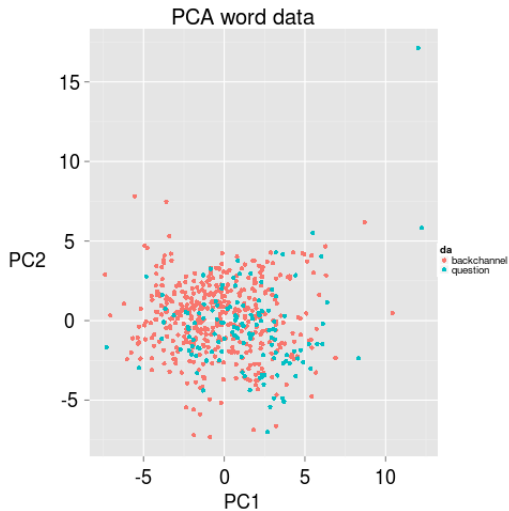


Figure: Projection onto the first two dimensions of the PCA space.

## Separating Prosodic Cues

- ▶ The distributional data above suggests prosodic differences between *really<sub>a</sub>* and *really<sub>b</sub>*.
- ▶ However, the large amount of overlap suggests that it would be difficult to differentiate these two classes based on these features, for any given instance.
- ▶ To further test this hypothesis, two classifiers were built in an attempt to separate the data.
  - ▶ Decision Tree classifier (DTree) (j48 in RWeka).
  - ▶ A Support Vector Machine (SVM) classifier with radial basis function kernel (`libsvm` via R), parameter determined by grid search.



## Classification results: 10 fold cross-validation

- ▶ All the data (444/132): Means over 100 randomizations.

	Error	95 CI	Fmeasure	95 CI
Base	22.92		0.671	
DTree	26.04	(23.96, 28.65)	0.685	(0.666 0.708)
SVM	23.64	(22.92, 24.40)	0.674	(0.671 0.675)

- ▶ The classifiers do worse than the majority class (backchannel) baseline!
- ▶ The highest split in the decision tree was mean intensity.
- ▶ The other split features were also intensity features, plus Legendre coefficient 3.

## Classification results: 10 fold cross-validation

- ▶ **Downsampled (132/132)**: Means of 100 random downsamplings of the backchannel class.

	Error	95 CI	Fmeasure	95 CI
Base	50.00		0.331	
DTree	43.36	(38.63, 48.00)	0.564	(0.516, 0.613)
SVM	36.95	(33.44, 40.75)	0.630	(0.592, 0.665)

- ▶ The classifiers do significantly better than the baseline.
- ▶ The decision trees generated vary greatly for the different downsampled sets in size and the features used.

# Discussion

- ▶ The downsampled classifiers and overall feature distributions suggest that prosodic features can help distinguish *really<sub>b</sub>* and *really<sub>q</sub>*.
- ▶ Question *really<sub>s</sub>* can be longer, have higher pitch and higher intensity (pitch slope seems to be not as important).
- ▶ However, it seems *really<sub>q</sub>* can also be low and short and *really<sub>b</sub>* high and loud. Question status almost certainly depends on other things in the context.
- ▶ It is plausible that these more effortful prosodic features help cue question status as a by-product of a more basic effect they have: underscore/intensify the actual meaning of the cue word.
  - ▶ Cue word *really* is underlying an elided question.

# The Perception of *Really* and *Right*

- ▶ **Hypothesis:** cue words signal speaker attitude towards information entering the dialogue. More effortful prosody intensifies this attitude.
- ▶ Intuitively, bigger *reallys* express something like more surprise.

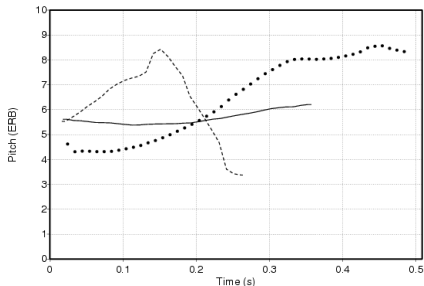


Figure: *Three reallys*.

## Questions:

- ▶ Does surprise differentiate *really's* interpretation?
- ▶ Are surprise and questioning meanings orthogonal?
- ▶ Do ratings match MDE annotations?

# Perception Experiment

- ▶ **Stimuli:** 64 backchannel *reallys*, question *reallys*, and *rights* (192 tokens total), each representing different quantiles: pitch range  $\times$  level  $\times$  duration (MDE 2003)
  - ▶ F0 values: manual alignment of glottal pulses, trimmed and smoothed, normalized to semitones. (Xu, 1999).
  - ▶ Manual labelling of boundaries.
- ▶ **Subjects:** 8 Penn students, native English speakers, paid.
- ▶ **Method:** The randomized stimuli were rated on 1-7 scales (1=not at all, 7=extremely) via a computer interface:
  - 'How surprised does the speaker sound?'
  - 'How much like a real question does this sound?'

# Results

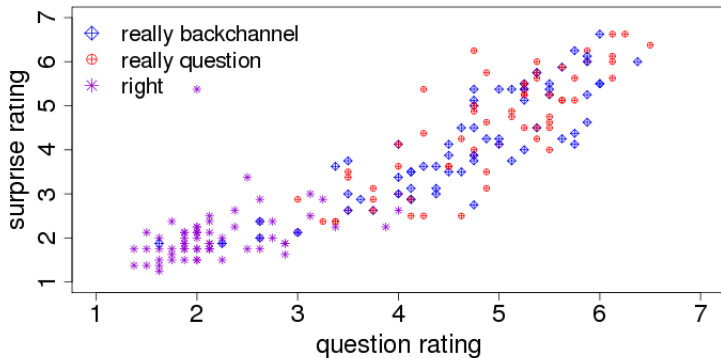


Figure: Average surprise v. question ratings.

# Results

- ▶ Backchannel/question MDE categories are not significantly different with respect to ratings.
  - ↳ Mann-Whitney U test:
    - ▶ question  $p = 0.30$ ,
    - ▶ surprise  $p = 0.18$ .
- ▶ Surprise/questioning are correlated.
  - ↳ Kendall's  $\tau = 0.63$ ,  $p < 0.001$  (non-normal dists).

# Prosodic Features

<i>Really</i>	$\tau_q$	<i>p</i> -value	$\tau_s$	<i>p</i> -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
<i>Right</i>				
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

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



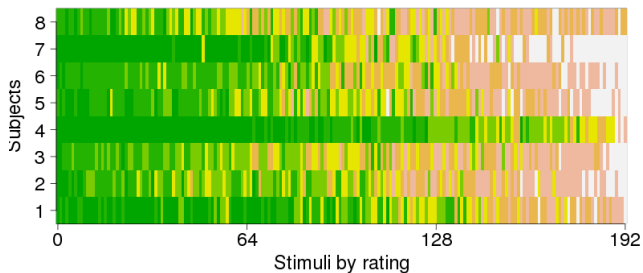
# Prosodic Features

- ▶ Question/surprise ratings are most highly correlated with pitch range and pitch level (*really*), pitch range and slope (*right*).  
↳ *more effortful prosody?*
- ▶ First syllable slope of *really* was significantly correlated with the ratings, but not the second syllable.  
↳ *final fall/rise does not seem associated with question interpretation.*  
↳ Perhaps 'questioning' is confounding...
- ▶ Pitch range 5-10 st:  $\text{mean}_q \text{ right} = 2.41$ ,  $\text{really} = 4.93$ .
- ▶ *really!*, *really?*, *really*, *really*,
- ▶ *right!*, *right?*, *right*, *right*,

## Subject Variation

Subjects could perform the task but had different rating biases.

→ Krippendorff's agreement  $\alpha$  for ordinal data (Artstein and Poesio, 2008): above chance  $\alpha_s = 0.58$ ,  $\alpha_q = 0.50$ , but still not great.



**Figure:** Stimuli ordered by mean average rating (increasing rightwards) by subject. Subject 4 was significantly different from the rest (Pairwise U tests:  $p < 0.001$ ).

# Experiment Summary

- ▶ Perception of surprise is a good way to look at how *really* varies.
- ▶ Effortful features intensify the underlying meanings:
- ▶ *Right* ( $p$ )  $\approx$   $p$  is (now) in the speaker's public beliefs
  - effort  $\propto$  agreement.
    - ▶ Affirmatives are not underlyingly response seeking.
- ▶ *Really* ( $p$ )  $\approx$   $p$  is new information,
  - effort  $\propto$  surprise
    - ▶ *Really* is underlying response seeking.

# What do rises do then?

As mentioned previously,

- ▶ Rises have been linked to the perception of uncertainty both formally and empirically. (Pon-Barry, 2008; Litman et al., 2009; Gravano et al., 2008; Nilsenova, 2006)
- ▶ However, rises have been found to be characteristic of affirmative backchannels in task oriented speech (Benus et al., 2007)
  - ↪ Not really cases of speaker uncertainty
- ▶ In all these cases, the rise-speaker seems to want the hearer to talk more.
- ↪ Rises tell us more about the state of the dialogue rather than the state of the speaker.

# What do rises do then?

How does cue word semantics interact with rising intonation?

▶ **Hypotheses:**

- ▶ Rises signal that the current question under discussion is unresolved.
- ▶ The underlying semantics of the utterance constrains how a rise is interpreted.
- ▶ Rather than ask directly about the QUD, we consider:
  - ▶ **EXPECTEDNESS** reflects certainty with respect to B's prior beliefs.
  - ▶ **CREDIBILITY** reflects how willing B is to believe A, i.e. add the content of A's utterance to their public beliefs.
  - ▶ **EVIDENCE** reflects the status of the QUD, i.e. whether A's utterance has been resolved/accepted or whether it is still contentious.

We can then also relate uncertainty to different aspects of dialogue structure.

# Stimuli

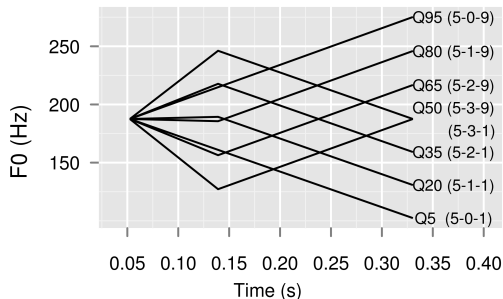
In this experiment, subjects evaluated context + resynthesized cue word pairs with respect to EXPECTEDNESS, CREDIBILITY, EVIDENCE.

- ▶ Cue words from Switchboard:
  - ▶  $2 \times \{really, well, okay, sure, yeah, \text{ and } right\}$
  - ▶ one word turns according to the transcripts.
  - ▶ checked for voice quality
- ▶ Contexts were drawn from turns immediately preceding one of the cue words, representing different levels of certainty (not exhaustive!)
  - ▶ factual, e.g. *X is Y*,
  - ▶ evaluative, e.g. *X is good*,
  - ▶ attributed, e.g. *I heard that X*,
  - ▶ inferred e.g. *probably X*.

## Resynthesis 8 ways

For each base token:

- ▶  $F_0$  values were based on quantiles of  $F_0$  values of the speaker for that conversation.
- ▶ The start point was the median value and the gradient between the mid- and endpoints remained the same.
- ▶ Timing was set with respect to the start, end, and the midpoint of the stressed vowel (manually identified).



- ▶ Varies overall pitch range and peak height but not slope.
- ▶ Test whether pitch range  $\propto$  unexpectedness.

# The Task

14 native speakers of American English, undergraduate students, paid, were asked to:

- ▶ Read the context: e.g. *the book was just ever so much better*
- ▶ Listen to the response: e.g. *really* (right)
- ▶ Answer the following questions (1-7 scale):
  - ▶ How expected does what A said seem to B?  
(1=completely unexpected, 7=completely expected)
  - ▶ How credible does what A said seem to B?  
(1=not at all credible, 7=completely credible)
  - ▶ 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).

⇒  $6 \times 2 \times 8 = 96$  cue words and  $6 \times 4 \times 4 = 96$  contexts



# Experiment Design

- ▶ Written context and audio (with text) response with replay enabled. (WebExp)
- ▶ Contexts and responses were randomly paired.
- ▶ 4 practice slides, 64 main experiment slides (human error!)

# Multilevel Model

- ▶ Model the effects of cue words, contours, contexts, subjects and the cue word/contour interaction as arising from different normal distributions (groups).
- ▶ The model parameters, along with finite population standard deviations for each group, were estimated using the Markov Chain Monte Carlo technique (JAGS)
- ▶ This gives us distribution rather than a point estimate!

## Multilevel Model

- ▶ Following Gelman and Hill (2007), the observed scores,  $y$ , for each question were modelled as follows.

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

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

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

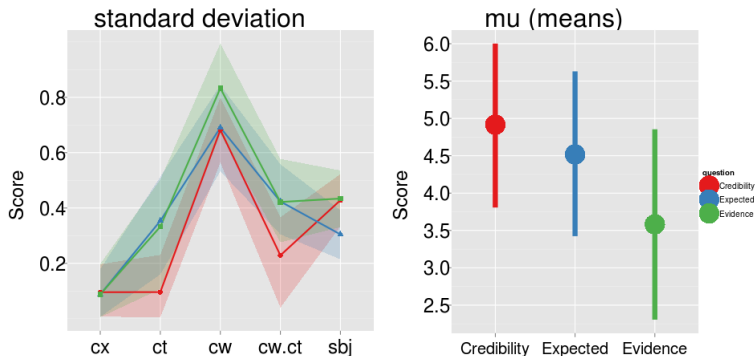
$$\alpha_l^{cx} \sim N(0, \sigma_{cx}^2) \text{ for } l = 1, \dots, 4 \quad (4)$$

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

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

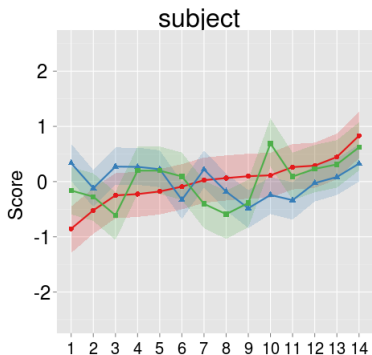
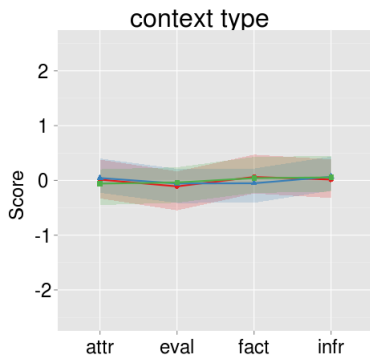
- ▶ e.g.  $\alpha_k^{cw}$  is a parameter representing the effect of cue word  $k$  holding the other variables constant.
- ▶ Let's look at estimated medians and 95% intervals for the different parameters for each of the scales.

# Parameter estimates



- ▶ Dot = median, shaded region = 2.5th-97.5th quantiles.
- ▶ Biggest standard deviation estimate comes from the cue word itself.
- ▶ Contour has more of an effect on EXPECTEDNESS and EVIDENCE.

# Parameter Estimates

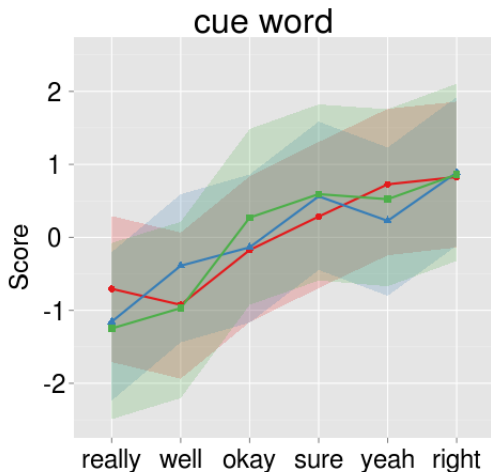


- ▶ Contexts don't have much of an effect: estimates are small and fall well inside the 95% intervals of the other type.
- ▶ Subjects have different strategies/biases.

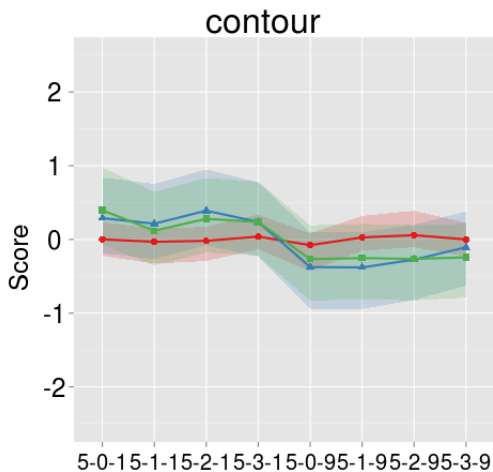
Now abstracting away from this...

# Parameter Estimates

- ▶ We get a credibility ordering over cue words.
- ▶ e.g. *right* is a strong agreement word.



# Parameter Estimates

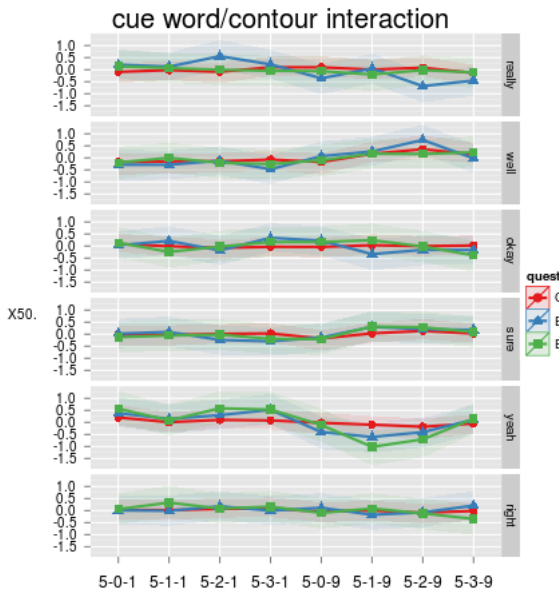


- ▶ Rising intonation lowers EXPECTEDNESS and EVIDENCE scores, but not CREDIBILITY.
- ▶ Posteriors associated with falls and rises appear quite distinct, medians for rises generally lying below the 2.5th quantile of the falls.

# Parameter Estimates

Variation across cue words:

- ▶ *yeah* can express more unexpectedness than *right*.
- ▶ *yeah*'s semantics is not as strong/specific  $\rightsquigarrow$  prosody more influential.
- ▶ *really*: variation appears to be mostly on the EXPECTEDNESS scale (c.f. previous experiment).





# The Interpretation of Rises

- ▶ Intonation did not have much of an effect on the credibility scale.
  - ▶ Rises on these cue words reflect difficulty integrating the new information rather than expressing disbelief.
  - ▶ Credibility is clearly reflected in the choice of cue word
- ▶ Rises signal that question under discussion is unresolved, implicitly signalling that resolution depends on the hearer:
  - ▶ This is congruent with the rising intonation of affirmative backchannels (turn passing),
  - ▶ It signals the expectation that more evidence will be presented
  - ▶ Though rises point to the QUD, they do not necessarily make an utterance an interrogative!

# Unexpectedness

- ▶ For cue words, inability to resolve the QUD may arise due to an addition being
    - ▶ epistemically unexpected (i.e. it doesn't fit their world view)
    - ▶ unexpected from the point of view of relevance.  
e.g. *right*: the respondent may agree with the content, while still feeling that it does not resolve the current QUD (different from surprise).
  - ▶ Greater overall pitch ranges were not really associated with the perception of more unexpectedness/surprise.
- ↪ the connection between pitch range and surprise may be more to do with slope or peak position rather than a max-min measure.
- But** resynthesis was based on quantiles, so no strong conclusions about individual contours across cue words.

## Summary and Implications

How can we analyze prosody? We need to look at what part of the discourse structure it acts on.

- ▶ Rising intonation works at the discourse/dialogue management level: it signals that the current QUD is unresolved.
  - ↪ Co-operative interlocutors should try to resolve it!
  - ↪ Conversational dialogue systems should evaluate utterances with rising intonation with respect to the QUD
- ▶ Cue words form a scale of CREDIBILITY, reflecting speaker attitude.
  - ↪ Track other conversational participants public beliefs.
  - ↪ Determine which type of cue word to use and when.

# Situating cue words in a discourse model

Roughly following (Portner, 1975).

- ▶ Conversational participants need to keep track of (at least):
  - ▶ Shared/accepted knowledge  $\rightsquigarrow$  the common ground (CG),
  - ▶ The tasks and goals  $\rightsquigarrow$  participants' to-do list (TD),
  - ▶ The discourse topic  $\rightsquigarrow$  the question under discussion stack (QUD).
- ▶ For a proposition/instruction to be added to the CG/TD it needs to pass some sort of credibility or *quality* threshold.
  - ▶ e.g. subjective probability given what's in the current conversational background (Davis et al., 2007), utility for the todo list...
- ▶ Cue words comment on the relationship between an utterance, the standards associated with discourse structures, and the component of the conversational background that is relevant for evaluation.

## Discourse structures

- ▶ Different affirmatives underlyingly associate with different structures...

	Common ground	QUD	To-do list
Sentence Type	Declarative	Interrogative	Imperative
Accept	<i>yeah, right, sure</i>	<i>(yes, no)</i>	<i>okay</i>
Reject	<i>no</i>		<i>no</i>
Check/Modify	<i>really, well</i>		<i>really, well</i>

- ▶ Task oriented dialogue: focus is on the to-do list.
- ▶ Conversational dialogue: focus is on the common ground.
- ▶ QUD tracks the discourse topic in both cases.
- ▶ Final rises signal that the QUD/Task is unresolved.

## Task oriented vs conversational speech

Cue word distributions in different corpora reflect the different conversational expectations.

Corpus	Yeah	Right	Sure	Okay	Really	Well
Columbia Games Corpus	903	189	-	2247	-	-
HCRC Maptask	1642	≈1500	3	2360	0	960
Let's Go	113	54	0	93	1	30
ICSI Meeting	11482	4420	286	4766	218	2499
Switchboard (NXT)	11922	2797	308	1540	535	5364

- ▶ *Really* is indicative of conversational speech, where a bit of surprise is a good thing (Gricean Information/Relevance!).
- ▶ *okay* is indicative of task oriented speech where completing the task is the priority.

## Work for semantics

- ▶ This approach helps shed light on the adjectival core of cue words as well as the gradable nature of truth values.
  - (5)
    - a. John's an okay dancer
    - b. Jane is right on the center point.
    - c. Mary really is an alien!
- ▶ In fact, we can bring them into the fold with other formal treatments of gradability (Lai, 2010).
- ▶ i.e. We can analyze cue words as fundamentally doing the same thing when they modify adjectives/nouns/verbs as when they are used as a response to a whole proposition.

# Conclusion

- ▶ We investigated how prosody interacts with cue words.
- ▶ The corpus study of *really*: DA backchannel/question categorization was too indirect a level upon which to try to tease out the contribution of prosody to meaning.
- ▶ The perception experiment of *really* and *right*: 'Effortful' prosody intensifies the underlying meaning rather than necessarily overlaying affect like surprise.
- ▶ The perception experiment of cue words and rises: Final rises signal that the question under discussion is unresolved.



## Conclusion and Futher Work

- ▶ Methodologically, this line of inquiry brings a more instrumental and empirical approach approach to the traditional semantic/pragmatic analysis.
- ▶ This approach helps shed light on how we can model gradability at the propositional level.
- ▶ Other parts of this project are on similarly discourse regulating sentential constructions, such as verum focus and negative polar questions and how they interact with prosody.





# Acknowledgments

# Thanks!

Jiahong Yuan, Mark Liberman, Florian Schwarz, Ani Nenkova, Aviad Eilam, C.E.A. Diertani, Yanyan Sui, Brittany McLaughlin, Yi Xu, Nigel Ward, Agustín Gravano, Chris Kennedy, Maribel Romero, Andrew Clausen, the Penn Splunchers, and many more...

## References

- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Benus, S., Gravano, A., and Hirschberg, J. (2007). The prosody of backchannels in American English. In *Proceedings of ICPHS 2007*, pages 1065–1068.
- Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, pages 1–33.
- Chen, A., Gussenhoven, C., and Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning. *Language and Speech*, 47(4):311–349.
- Davis, C., Potts, C., and Speas, M. (2007). The pragmatic values of evidential sentences. In Gibson, M. and Friedman, T., editors, *Proceedings of the 17th Conference on Semantics and Linguistic Theory*, pages 71–88. CLC Publications, Ithaca, NY.
- Evanini, K. and Lai, C. (2010). The importance of optimal parameter setting for pitch extraction. In *Presented at the 2nd PanAmerican/Iberian Meeting on Acoustics, Cancun, Mexico, 15-19 November 2010*.
- Farkas, D. and Bruce, K. (2009). On Reacting to Assertions and Polar Questions. *Journal of Semantics*.
- Fernandez, R. (2006). *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. PhD thesis, Department of Computer Science, Kings College London, University of London.
- Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge.
- Ginzburg, J. (2009). *The Interactive Stance: Meaning for Conversation (forthcoming in 2009)*. Studies in Computational Linguistics. CSLI Publications.
- Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP-92*.
- Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. PhD thesis, Columbia University.
- Gravano, A., Benus, S., Hirschberg, J., German, E. S., and Ward, G. (2008). The effect of prosody and semantic modality on the assessment of speaker certainty. In *Proceedings of 4th Speech Prosody Conference, Campinas, Brazil*.
- Gunlogson, C. (2002). Declarative questions. In Jackson, B., editor, *Proceedings of Semantics and Linguistic Theory XII*. CLC Publications.
- Gunlogson, C. (2008). A question of commitment. *Belgian Journal of Linguistics*, 22(1):101–136.

- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.
- Gussenhoven, C. and Chen, A. (2000). Universal and Language-Specific Effects in the Perception of Question Intonation. In *Sixth International Conference on Spoken Language Processing*. ISCA.
- Ishi, C., Ishiguro, H., and Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech Communication*, 50(6):531–543.
- Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard-DAMS Labeling Project Coders Manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science.
- Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120.
- Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118:1038.
- Lai, C. (2010). What does really really mean?: Evidence, standards and probability in dialogue. In *Presented at NELS 41, Philadelphia, October 2010*.
- Litman, D., Rotaru, M., and Nicholas, G. (2009). Classifying Turn-Level Uncertainty Using Word-Level Prosody. In *Proceedings of Interspeech'09*.
- Nilsenova, M. (2006). *Rises and Falls. Studies in the semantics and pragmatics of intonation*. PhD thesis, University of Amsterdam.
- Ohala, J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41(1):1–16.
- Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*. MIT Press, Cambridge.
- Pon-Barry, H. (2008). Prosodic manifestations of confidence and uncertainty in spoken language. In *Proceedings of Interspeech'08*.
- Portner, P. (1975). Instructions for Interpretation as Separate Performatives. In Schwabe, K. and Winkler, S., editors, *On Information Structure, Meaning and Form*, pages 407–426. John Benjamins.
- Reese, B. (2007). *Bias in Questions*. PhD thesis, University of Texas at Austin.
- Schlagen, D. (2004). *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. PhD thesis, School of Informatics, University of Edinburgh.
- Steedman, M. (2000). Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry*, 31(4):649–689.
- Strassel, S. (2003). Simple Metadata Annotation Specification V5.0. *Linguistic Data Consortium, Philadelphia*    

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f<sub>0</sub> contours. *Journal of Phonetics*, 27:55–105.

## Affect, Attitude and Biological Codes

Gussenhoven (2004) interprets pitch in terms of biological codes:

- ▶ **Frequency Code:** Higher voices are more submissive, friendly, polite (Ohala, 1984).
  - ▶ Rises in questions  $\mapsto$  uncertainty.
  - ▶ Gravano et al. (2008): downstepped contours sound more certain, rising contours sound uncertain for declaratives.
- ▶ **Effort Code:** important information is produced with more articulatory effort.
  - ▶ Chen et al. (2004): pitch peak height, peak delay and register correlate positively with *surprise*.
  - ▶ Ishi et al. (2008): surprise is associated with short rises and non-modal voice for Japanese particles.
- ▶ **Production Code:** energy diminishes during exhalation
  - ▶ high beginnings signal new topics while high endings signal continuations

Teasing apart the relative contribution of each code is not straightforward.

## Cue words meanings

- ▶ *yeah*: that is acceptable (CG).  
↳ add proposition p to the common ground,
- ▶ *right*: that is an accurate characterization, according to my beliefs.  
↳ add p to the common ground; add p to public beliefs.
- ▶ *okay*: that is acceptable (TD).  
↳ add instruction i to your to-do list
- ▶ *sure*: It is for certain that p should accepted.  
↳ add p to CG (TD) with certainty.
- ▶ *really*: Is that acceptable if you raise the quality standard?  
↳ Request a stricted quality assessment.
- ▶ *well*: We need more information to make the assessment  
↳ p's evaluation is contingent on other information.

## With a rise

- ▶ *yeah, right, okay*: Speaker is willing to accept, but the QUD/TD remains unresolved.
- ▶ *sure*: Check addressee's certainty, before accepting p.
- ▶ *really*: Check the quality of p's evaluation before accepting p.
- ▶ *well*: Signal that p's evaluation is contingent on information to be added.

*really, yeah*  $\rightsquigarrow$  conventionalized meaning: that's new.



# Cue word frequencies

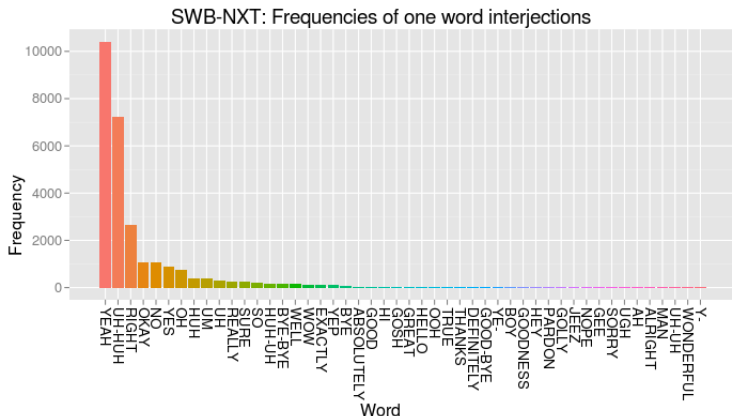


Figure: Most frequent one word interjections (INTJ) in Switchboard NXT (Penn Treebank)

# Dialogue acts

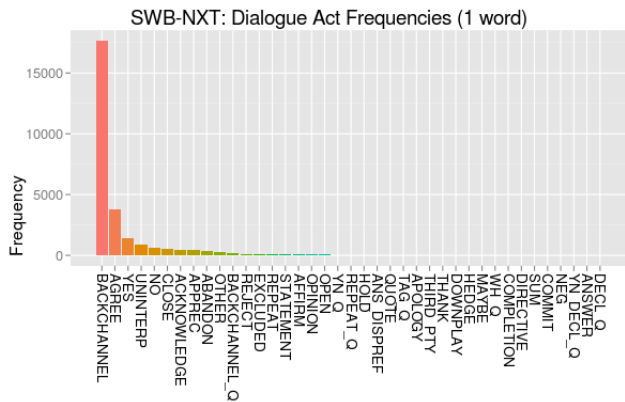


Figure: Dialogue acts for one word turns in Switchboard

# NSUs, Dialogue Frameworks

- ▶ Work on formal dialogue models usually focuses on situating these sorts of utterances within similar sorts of taxonomies.
  - ▶ Fernandez (2006): e.g. *really* ↪ Clarification request or backchannel.
  - ▶ Schlangen (2004): e.g. *really* ↪ Comment Question.
  - ▶ Previous empirical work on cue words has focused sense/dialogue act classification.
- ▶ What makes something a question? Its form? Its function?