# Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour*

Marie-Catherine de Marneffe and Judith Tonhauser
Department of Linguistics, The Ohio State University

March 26, 2016

**Abstract**

Polar questions can be given direct answers (e.g., *Do you want to eat? — No*) and indirect answers (e.g., *Do you want to eat? — I'm not hungry*). Listeners infer positive or negative responses from indirect answers to polar questions with varying degrees of confidence (e.g., Clark 1979, Hirschberg 1985, Green and Carberry 1992, 1994, de Marneffe et al. 2009). For spoken language, the prosodic realization of the indirect answer has been speculated to provide a cue to the intended meaning of the indirect answer (Green and Carberry 1999:fn.34). This paper presents an experiment designed to identify whether and how the prosodic realization of an indirect answer to a polar question influences the response that listeners infer from the indirect answer. The perception experiment explored American English listeners' interpretations of indirect answers with scalar adjectives (e.g., *She's attractive*) realized with a neutral contour (H* L-L%) or the rise-fall-rise contour (L*+H L-H%) in response to polar questions with semantically stronger adjectives (e.g., *Is your sister beautiful?*). Listeners inferred significantly more negative responses to the polar questions when the indirect answer was realized with the rise-fall-rise contour than with the neutral contour. These findings show that the prosodic realization of an indirect answer can provide a cue to the speaker's intended meaning. The paper also discusses implications of our findings for scalar implicature generation and the meaning of the rise-fall-rise contour.

## 1 Introduction

Polar questions like (1Q) can be given direct answers, such as (1A1) or (1A2), or indirect answers, such as (1A3) from Sperber and Wilson 1986:56.

(1) Q: Do you want some coffee?
A1: Yes.
A2: No.
A3: Coffee would keep me awake.

*For helpful feedback on this research, we thank the participants of Speerlab at The Ohio State University and the audiences at CUNY 2013, at the *Questions in Discourse* workshop in Stuttgart in 2013, at the University of Munich in 2014 and at the MXPRAG workshop in Berlin in 2015. We are grateful to Julie McGory and Mike Phelan for recording the stimuli, to Lauren Squires for helpful discussion of the stimuli, and to The Ohio State University for financial support.

A *yes* or *no* answer to a polar question ?$p$ entails $p$ or $\neg p$, respectively (assuming, e.g., Krifka's 2013 anaphoric analysis of response particles). Following Groenendijk and Stokhof 1984, we take a polar question ?$p$ to partition the context set (the set of worlds compatible with the propositions in the common ground of the interlocutors) into two cells: one cell consists of worlds in which $p$ is true, i.e., for (1Q), worlds in which the proposition that the addressee wants some coffee is true, and the other cell consists of worlds in which $\neg p$ is true, i.e., for (1Q), worlds in which the proposition that the addressee does not want some coffee is true. (For alternative analyses of polar questions see AnderBois 2016 and references therein.) *Yes* and *no* are direct answers to polar questions (as are responses that lexicalize the content of $p$ or $\neg p$) because they entail a unique cell of the context set without providing additional information.

We assume that a response to a polar question ?$p$ *addresses* the polar question if and only if the response entails or conversationally implicates either $p$ or $\neg p$. We furthermore assume that a response to a polar question is judged to be acceptable if and only if the response addresses the polar question. The direct answers in (1A1) and (1A2) are thus predicted to be judged to be acceptable in response to the polar question in (1Q) because they entail $p$ and $\neg p$, respectively.

The response in (1A3) does not entail $p$ or $\neg p$ since it is possible to follow (1A3) up with *...and I (therefore) would like some coffee* as well as with *...and I (therefore) wouldn't like some coffee*. If the listener can take the speaker who produced (1A3) in response to the question in (1Q) to be cooperative, then the listener draws the inference that (1A3) must conversationally implicate either $p$ or $\neg p$. As discussed in the literature on indirect answers (e.g., Clark 1979, Hobbs and Robinson 1979, Hirschberg 1985, Green and Carberry 1992, 1994, 1999, de Marneffe et al. 2009), listeners rely on a variety of cues to draw inferences about whether the intended meaning of the indirect answer to the polar question is a positive one (i.e., 'yes, I want some coffee') or a negative one (i.e., 'no, I don't want some coffee'). One cue comes from the discourse goals and plans of the interlocutors, as well as their mutual beliefs and shared knowledge. For instance in (1), if it is mutual knowledge to Q and A that A has an interview early in the morning, (1A3) is likely to be interpreted as a negative answer. However, if Q and A have to finish an important job and it is already late in the day, (1A3) is likely to be interpreted as a positive answer. Another cue comes from utterances subsequent to the indirect answer, as discussed in Green and Carberry 1992, 1994, 1999. For instance, if A elaborates on (1A3) with *I have an important meeting tomorrow*, the listener would conclude that A intended a negative response to the question, but a continuation such as *We need to work a few more hours* would be taken to indicate a positive response.

Green and Carberry (1999:fn.34) also suggest that the prosodic realization of the indirect answer may provide a cue when they state that "it is an interesting question for future research whether it [prosodic information, MCdM & JT] can help in recognizing the speaker's intentions". However, the role of prosody in the interpretation of indirect answers has not yet been systematically explored.[1] In this paper, we start to fill this gap by presenting the results of a perception experiment designed to identify whether and how the prosodic realization of an indirect answer to a polar question influences the inference that listeners draw from the indirect answer. The experiment explored American English listeners' interpretations of indirect answers with scalar adjectives, such as (2B), in response to polar questions with semantically stronger adjectives, such as (2A). The indirect answers were realized either with a neutral contour (H* L-L%)[2] or

[1] See Goodhue and Wagner 2014 on the role of prosody in the interpretation of 'yes' and 'no' answers to polar questions.

[2] In Tone and Break Indices (ToBI) annotation (Beckman and Ayers Elam 1997), a H* is a high tone pitch accent aligned with the stressed syllable and L*+H is a complex pitch accent made up of a low tone aligned with the stressed syllable and an immediately following high tone. A low intermediate phrase accent that is followed by a low intonational phrase boundary tone is indicated by L-L% and one that is followed by a high intonational phrase boundary tone is indicated by L-H%.

with the rise-fall-rise contour (L*+H L-H%), with the pitch accents of the contours realized on the stressed syllable of the adjective. (Both of these contours are explained in section 3.)

(2) A: Is your sister beautiful?
B: She's attractive.

Since *beautiful* is the stronger scalemate of *attractive* in the ⟨*beautiful, attractive*⟩ scale, a listener may infer from B's answer, by standard Gricean reasoning, that B intended to convey that her sister is attractive but not beautiful, i.e., a negative response. To explore whether prosody influences the interpretation of indirect answers, we exploit the fact that a scalar implicature may arise from indirect answers like that in (2): our perception experiment was designed to explore whether and how the prosodic realization of the indirect answer with the weaker scalar adjective influences a listeners' degree of belief in the scalar implicature, i.e., whether the implicature is more or less likely to arise. We chose the rise-fall-rise contour for our investigation since it has been documented extensively with indirect answers and even described to be particularly natural with indirect answers, as opposed to direct answers. Ward and Hirschberg (1985:769), for example, state that the rise-fall-rise contour "occurs precisely when speakers cannot, or do not wish to, commit themselves to direct responses". The contour is also ideally suitable for our investigation since it is generally taken to conventionally contribute meaning to the utterance on which it is realized (e.g., Ward and Hirschberg 1985, Büring 1997, 2003, Wolter 2003, Constant 2012, Wagner 2012, Wagner et al. 2013).

In the next section, we provide relevant empirical and theoretical background on scalar implicatures, including the influence of prosody on scalar implicature generation. Section 3 presents the perception experiment we conducted and section 4 discusses the implications of our findings for analyses of the rise-fall-rise contour and in the context of previous research on prosody and scalar implicatures. Section 5 concludes the paper.

## 2 Empirical and theoretical background on scalar implicatures

Scalar implicatures are inferences that listeners may draw from utterances with expressions that form a scale with stronger expressions (e.g., Gazdar 1979, Horn 1972, 1989). For example, the determiners *some* and *all* form a so-called Horn scale, represented by ⟨*all, some*⟩, with *all* the stronger expression since *all X are Y* entails *some X are Y*, but not vice versa. If a speaker chooses to make a statement that contains the semantically weaker expression rather than a statement that contains the semantically stronger expression, the speaker can be taken to implicate that she does not believe that the statement with the stronger expression is true. Thus, a listener who hears an utterance with *some* may infer that the speaker intended to convey the scalar implicature 'some but not all' under the assumption that the speaker had reasons to not use the stronger expression *all*, for instance because that stronger statement would have been false. For example, a listener who hears an utterance of (3a) may infer that the speaker intended to convey not just the literal meaning of (3a) given in (3b), but rather the meaning in (3c), which is the literal meaning enriched with the scalar implicature.

(3) a. I like some tango orchestras.
b. I like some (and possibly all) tango orchestras.
c. I like some but not all tango orchestras.

Gradable adjectives also form scales. For instance, the adjectives *beautiful* and *attractive* that were featured in example (2) form the scale ⟨*beautiful, attractive*⟩ since *beautiful* is stronger than *attractive*. (We present two diagnostics for identifying pairs of adjectives that stand in this strength relationship in section 3.2.1.) Given that the two adjectives form a scale, it follows that a listener who hears B's response to A's question in (2), repeated below for convenience, may infer that B intended to convey the scalar implicature 'attractive but not beautiful', for example because a stronger statement (e.g., *Yes* or *She's beautiful*) would have been false.

(2) A: Is your sister beautiful?
B: She's attractive.

Scalar implicatures do not arise obligatorily. In fact, a hallmark property of scalar implicatures is that they are cancelable. Thus, B in (2) could follow up her utterance with *...and, yeah, in fact, she's beautiful*, without contradiction. Recent research shows that several factors are at play in whether an utterance with a weaker scalar expression gives rise to an implicature. One of these factors is the scalar expression itself. Comparisons across different types of scalar expressions reveal that quantificational determiners like *some* are much more likely to give rise to scalar implicatures than scalar adjectives (Doran et al. 2012, van Tiel et al. 2016); scalar implicature rates for adjectives in these studies are as low as 4%. Other factors include what the listener believes the speaker's knowledge state to be (Goodman and Stuhlmüller 2013), the form of the scalar expression (e.g., *some* versus *some of* versus *sm*; Thorward 2009, Degen 2015), the experimental paradigm (Geurts and Pouscoulous 2009, Degen and Goodman 2014), and whether the scalar expression is realized with a pitch accent or as part of the focus of the utterance (e.g., Chevallier et al. 2008, Thorward 2009, Zondervan 2010, Schwarz et al. to appear). Since the prosodic realization of the utterances that give rise to scalar implicatures is of direct concern to this paper, we review these studies in more detail.

Chevallier et al. (2008) and Schwarz et al. (to appear) focused on the disjunction 'or' in French and English, respectively. In Chevallier et al.'s study, French participants read five-letter-strings (a word, e.g., 'TABLE' or a non-word, e.g., 'JAMIS') and would hear a disjunctive or a conjunctive statement about the letters in the string (e.g., 'There is an A or a B' or 'There is an A and a C', translated from French). In some of these utterances, the 'or' or 'and' received "contrastive stress" (p.1747). (Chevallier et al. do not give further information about the prosody.) Participants were asked to judge whether the statement was true or false. In Schwarz et al.'s study, participants heard English sentences such as *Mary will invite Fred or Sam to the barbecue* in which a contrastive pitch accent (L+H*) was realized either on the auxiliary (*will*) or the disjunction *or*. All content words were uttered with a H* accent. Participants were presented in writing with two interpretations ('Mary will invite Fred or Sam but not both' and 'Mary will invite Fred or Sam or possibly both') and were asked to select the right one. Both Chevallier et al. (2008) and Schwarz et al. (to appear) found that the disjunction is more often interpreted as exclusive ('A or B, but not both') rather than inclusive ('A or B, and possibly both') when it is realized with a pitch accent.

Zondervan (2010) explored the hypothesis that scalar implicatures only arise when the scalar expression is part of the information focus of the sentence (van Kuppevelt 1996, van Rooij 2002). He examined the scalar expression 'or' in Dutch. One of his experiments manipulated information focus marked by prosody as well as context. Participants read a context, then heard a target sentence (containing the disjunction), which they had to judge as true or false. Context varied between the focus and non-focus condition. Where the pitch accent was realized differed from Chevallier et al. 2008 and Schwarz et al. to appear: in the focus condition, the two elements of the disjunction received a pitch accent (H*), but not the scalar expression

itself (e.g., 'Paola took an APPLE or a PEAR from the fruit section' translated from Dutch); in the non-focus condition, the subject of the sentence received a pitch accent (H*) (e.g., 'PAOLA took an apple or a pear from the fruit section' translated from Dutch). In line with previous results, Zondervan found a higher implicature rate in the focus condition.

Thorward (2009) investigated how pitch accent interacts with implicature in the case of English *some*. In her experiment 2, participants watched the experimenter move 3 or 4 (out of 4) animals over a fence in a barnyard. A puppet then described the scene, e.g., *Some cats jumped over the fence*, and participants were asked to judge whether the puppet's utterance was an accurate description of the scene. Pitch accent varied in the target sentence: either the determiner *some* received a pitch accent (H* or L+H*) or the noun describing the animal received a pitch accent (L+H*). For the scenarios in which all the animals jumped over the fence, adults rejected the puppet's utterance more often when *some* was realized with a pitch accent than when it was not, confirming that pitch accent plays a role in implicature generation.

These studies show that scalar implicature generation is sensitive to prosody. Specifically, they show that scalar implicature generation is sensitive to whether the weaker scalar expression is realized with a pitch accent or as part of the focus. Building on these results, the perception experiment we describe in the next section explores whether and how the prosodic realization of indirect answers with weaker scalar expressions influences the interpretation of the indirect answer.

# 3 Perception experiment

The perception experiment was designed to explore how the response that listeners infer from an indirect answer to a polar question is influenced by the prosody of the indirect answer.

## 3.1 Participants

63 individuals participated in the experiment. We removed the data of 2 participants who did not consent, of 2 participants who did not self-identify as speakers of American English, of 1 participant who did not give a response to all dialogues, and of 1 participant who failed to answer correctly two out of the three control dialogues. We analyzed the data of the remaining 57 participants (30 female, 26 male, 1 undeclared; ages: 18-68, median age: 31.5).

## 3.2 Stimuli

Stimuli in the experiment consisted of spoken, two-turn dialogues between two individuals, Mike and Julie. In each dialogue, Mike asked Julie a polar question formed from a copula clause with an adjective and Julie responded with a copula clause. The target stimuli were formed from sixteen pairs of adjectives with the following two properties: i) the adjectives were interpreted on the same scale and ii) one adjective was semantically stronger than the other one. In the target stimuli, Mike's question realized the stronger adjective (e.g., *Is your sister beautiful?*) and Julie's indirect answer realized the weaker adjective (e.g., *She's attractive*).

### 3.2.1 Norming studies for the selection of adjectives

We used two diagnostics to identify suitable pairs of adjectives. The first diagnostic is the 'even' diagnostic: two adjectives *W* ('weaker') and *S* ('stronger') are interpreted on the same scale, with *S* stronger than *W*, if

and only if a sentence containing *W* can be followed up with 'even *S*' and a sentence containing *S* cannot be followed up with 'even *W*'. The 'even' diagnostic is applied in (4) to the pair ⟨*beautiful, attractive*⟩:

(4) a. My sister is attractive, even beautiful.
b. My sister is beautiful, #even attractive.

The second diagnostic is the 'but not' diagnostic from Horn 1972: two adjectives *W* and *S* are interpreted on the same scale, with *S* stronger than *W*, if and only if a sentence with *W* can be followed up with a denial of *S* and a sentence with *S* cannot be followed with a denial of *W*. The 'but not' diagnostic is applied in (5) to the pair ⟨*beautiful, attractive*⟩:

(5) a. Samantha is attractive, but not beautiful.
b. Samantha is beautiful, #but not attractive.

We ran two norming studies to identify adjective pairs that pass these two diagnostics with a set of 30 adjective pairs that a semantically trained native speaker of American English had previously judged to pass the two diagnostics.[3] Appendix A gives the 30 adjective pairs, the carrier sentences used in the norming studies and the results of the two norming studies for each pair.

All of the experiments discussed in this paper were run on Amazon's Mechanical Turk platform. To participate, individuals had to have US IP addresses and also had to have completed at least 1,000 Human Intelligence Tasks (HITs) with an overall approval rate of at least 97%. Participants completed a short survey about their age, their native languages and, if English was their native language, whether they spoke American English (as opposed to, e.g., Indian or Australian English). Participants were told that they would be paid regardless of how they responded to the survey questions. We only analyzed responses by participants who self-identified as native speakers of American English.

**Norming study 1** In the first norming study, we used the 'even' diagnostic to identify pairs of adjectives that are interpretable on the same scale, with one adjective stronger than the other.

*Participants*. 41 individuals participated in the experiment. We discarded the data of 1 participant who did not consent, 2 participants who gave incorrect answers to more than 2 of the 5 control stimuli and 8 participants who did not provide an answer to any of the 5 control stimuli. We analyzed the data of the remaining 30 participants (11 female, 19 male; ages: 18-62, median age: 38.5).

*Stimuli*. The written stimuli consisted of two-turn discourses in which the first speaker uttered a sentence with an adjective, and the second speaker uttered a follow-up consisting of *even* and the other adjective, as in Figure 1. For each of the 30 adjective pairs, we created one target stimulus where the first utterance realized the adjective that was hypothesized to be weaker and the second utterance realized *even* and the adjective that was hypothesized to be stronger, and a target stimulus with the reverse order

[3] Some of the adjective pairs that we tested in the norming studies were used in previous studies on scalar items. Not all of these pairs passed both of our diagnostics. Specifically, the pairs ⟨*unforgettable, memorable*⟩, ⟨*horrific, unsettling*⟩, ⟨*ridiculous, silly*⟩, which were used in van Tiel et al. 2016, and the pair ⟨*obnoxious, annoying*⟩, which was used in Doran et al. 2009, did not pass one or both diagnostics. Furthermore, we relied on the adjective pair ⟨*tight, snug*⟩, used in Doran et al. 2009 and van Tiel et al. 2016, in a control stimulus in norming study 1: the sentence *Jeff's pants are snug. They're even tight* was hypothesized to be acceptable, but the corresponding stimulus was judged to be odd by 7 out of the 16 participants that judged that item. While the findings of our norming studies may in part be due to the carrier sentences we used, these observations minimally suggest it is sensible for experimental investigations of scalar implicatures to identify whether native speakers of the relevant language agree that two scalar adjectives stand in the relevant strength relationship.

of the adjectives. In addition to the 60 target stimuli, there were 5 control stimuli formed from clearly acceptable sentences (e.g., *Blair's room is comfortable. It's even cozy.*) and 5 that were formed from clearly unacceptable sentences (e.g., *Elle's exam scores were perfect. They were even below average*).

The 60 target stimuli were distributed across two lists so that each list contained one of the two target stimuli of each adjective pair. Each list had 15 target stimuli that were hypothesized to be acceptable and 15 that were hypothesized to be unacceptable. Both lists were divided into two sub-lists of 15 target stimuli (roughly half of which were hypothesized to be acceptable). Each sub-list also had five control stimuli (roughly half of which were acceptable). The stimuli in each sub-list were presented to the participants in a pseudo-randomized order. Each participant was assigned to a list, and could either complete one or both sub-lists of that list.

*Procedure*. Participants were told that they would read small exchanges between two friends, Mike and Jack, and that Jack is the kind of guy who likes to trump whatever Mike says. They were also told that Jack doesn't master English very well, and so he sometimes says something odd. Participants read two-turn dialogues between Jack and Mike, and were asked to judge whether Jack's response to Mike was odd, by selecting either 'odd' or 'not odd', as illustrated in Figure 1.

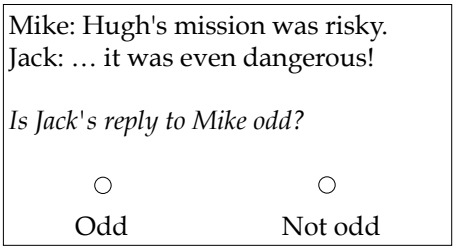



Figure 1: Sample stimulus and response task in norming study 1

At the beginning of the trial, each participant completed two training examples for which the correct answers were given.

*Results*. Each target stimulus received between 6 and 8 judgments (median: 7). For each target stimulus, we determined whether it was judged to be acceptable or unacceptable by identifying whether the majority of judgments were 'not odd' or 'odd', respectively. For an adjective pair to pass the 'even' diagnostic, the target stimulus in which the stronger adjective followed *even* had to be judged to be acceptable, and the target stimulus in which the weaker adjective followed *even* had to be judged unacceptable. Of the 30 adjective pairs, 24 passed the 'even' test.

**Norming study 2** In the second norming study, we used the 'but not' diagnostic to identify pairs of adjectives that are interpretable on the same scale, with one adjective stronger than the other.

*Participants*. 36 individuals participated in the experiment. We discarded the data of 3 participants who did not consent, 1 participant who did not indicate that (s)he was an American English speaker, and 1 participant who did not correctly answer at least 3 of the 5 control stimuli. We analyzed the data of the remaining 31 participants (14 female, 17 male; ages: 25-60, median age: 36).

*Stimuli*. Target stimuli consisted of sentences like those in (5). For each of the 30 adjective pairs, we created one target sentence where the adjective that was hypothesized to be weaker was followed by *but not* and the adjective that was hypothesized to be stronger, as in (5a), and a target sentence with the reverse

order of the adjectives, as in (5b). In addition to the 60 target stimuli, there were 5 control stimuli that were clearly non-contradictory sentences (e.g., *Jeff's mother is happy and creative.*) and 5 that were clearly contradictory sentences (e.g., *Jane's blouse is new and used.*).

The 60 target sentences were distributed across two lists so that each list contained one of the two target sentences of each adjective pair. Each list had 15 target sentences that were hypothesized to be contradictory and 15 that were hypothesized to be non-contradictory. Both lists were divided into two sub-lists of 15 target sentences (roughly half of which were hypothesized to be contradictory). Each sub-list also had five control sentences (roughly half of which were contradictory). The stimuli in each sub-list were presented to the participants in a pseudo-randomized order. Each participant was assigned to a list, and could either complete one or both sub-lists of that list.

*Procedure*. Participants read sentences and were asked to judge whether the sentence were contradictory or not, by selecting either 'contradictory' or 'not contradictory', as illustrated in Figure 2.

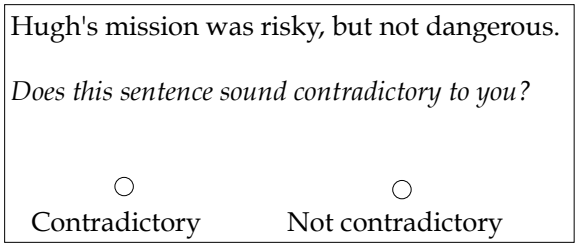



Figure 2: Sample stimulus and response task in norming study 2

At the beginning of the trial, each participant completed two training examples for which the correct answers were given.

*Results*. Each target sentence received between 6 and 9 judgments (median: 7). For each target sentence, we determined whether it was judged to be acceptable or unacceptable by identifying whether the majority of judgments for the target sentence were 'not contradictory' or 'contradictory', respectively. For an adjective pair to pass the 'but not' diagnostic, the target sentence in which the stronger adjective followed *but not* had to be judged to be acceptable, and the target sentence in which the weaker adjective followed *but not* had to be judged unacceptable. Of the 30 adjective pairs, 24 passed the 'but not' test.

**Selection of adjective pairs** Of the 30 adjective pairs, 19 passed both the 'even' and the 'but not' diagnostic. 16 of these adjectives were used in the perception experiment.[4] These adjective pairs are given in Appendix B with the carrier sentences used in the perception experiment.

### 3.2.2 Types of stimuli and preparation of prosodic stimuli

The stimuli in the perception experiment consisted of two-turn, spoken discourses: a polar question formed from a copula sentence with an adjective and a copula sentence answer with an adjective. The stimuli were recorded in a sound-attenuated booth with two speakers of American English trained in American English ToBI. The polar questions across all types of stimuli were produced with a low tone pitch accent on the

[4]The adjective pairs were also used in a separate perception experiment designed to explore whether personal experience influences the listeners' degree of belief in the scalar implicature. Since the relevant manipulation could not be implemented for three adjective pairs, these adjective pairs were excluded from further consideration in the perception experiment discussed in this paper.

stressed syllable of the adjective, a high intermediate phrase accent and a high intonational phrase boundary tone (L* H-H%).

In the target stimuli, the stronger adjective of a pair occurred in the polar question and the weaker adjective of the pair occurred in the indirect answer, as in (2), repeated here for convenience.

(2) A: Is your sister beautiful?

B: She's attractive.

The answers of the target stimuli were recorded with one of two contours, resulting in a total of 32 target stimuli (two for each of the 16 pairs of adjectives). The first, 'neutral' contour consisted of a high tone pitch accent on the stressed syllable of the adjective, a low intermediate phrase accent and a low intonational phrase boundary tone (H* L-L%). The second, 'rise-fall-rise' contour consisted of a complex pitch accent made up of a low tone aligned with the stressed syllable of the adjective and an immediately following high tone, a low intermediate phrase accent and a high intonational phrase boundary tone (L*+H L-H%). The thirty adjective pairs from which the 16 adjective pairs were chosen for the perception experiment were all designed such that the stressed syllable of the weaker adjective was followed by at least one unstressed syllable, and therefore the weaker adjective had sufficient segmental material on which to realize this complex contour. The examples in Figure 3 illustrate the two contours.

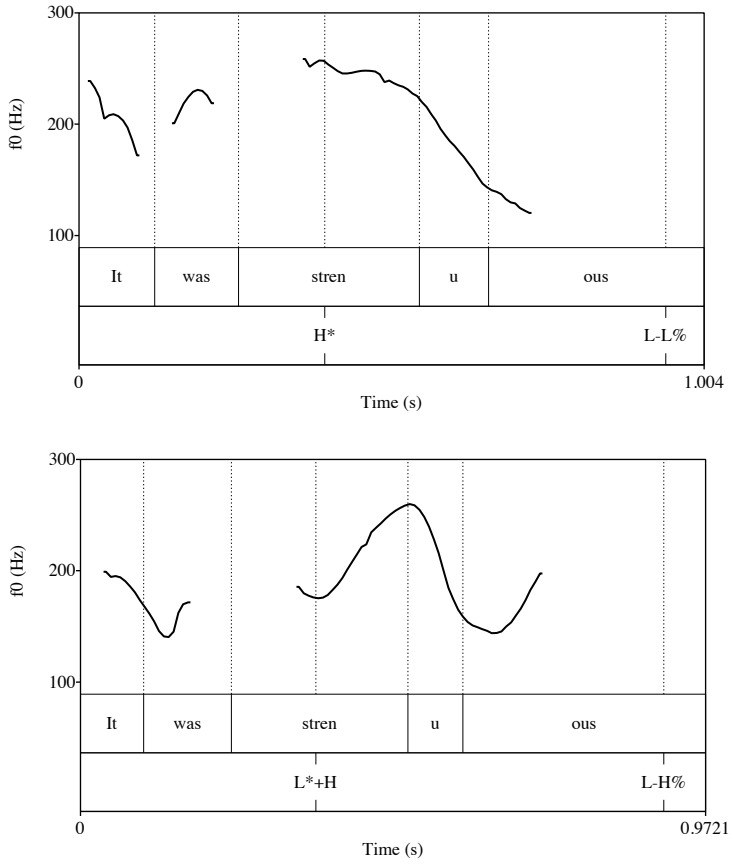



Figure 3: Two utterances of *It was strenuous* with the neutral contour (top panel) and the rise-fall-rise contour (bottom panel)

In the filler stimuli, in contrast to the target stimuli, the weaker adjective of a pair of adjectives occurred in the polar question and the stronger adjective of the pair occurred in the answer, as in (6). Thus each adjective pair gave one filler stimulus, for a total of 16 filler stimuli.

(6) A: Is your sister attractive?
B: She's beautiful.

In the six control stimuli, the adjectives realized in the polar question and the answer were antonyms, as in (7):

(7) A: Was your suitcase light?
B: It was heavy.

The answers of filler and control stimuli were recorded only in the neutral contour. Since in the filler stimuli the meaning of the adjective in the answer was stronger than the meaning of the adjective in the question, we hypothesized that the listeners would be able to infer a positive response solely on the basis of the meanings of the adjectives. Thus, we included the filler stimuli to distract participants from the prosody manipulation present in the target stimuli. The control stimuli, on the other hand, were included to be able to assess whether participants were paying attention to the task. The stimuli are available at `http://www.ling.ohio-state.edu/~mcdm/downloads/2015/IQAPsounds`.

For each adjective pair, there were three stimuli (two target stimuli and one filler stimulus), for a total of 48 target-or-filler stimuli. We distributed these 48 target-or-filler stimuli across three lists so that each adjective pair occurred once in each list (either as a target or as a filler stimulus). Each of the three lists thus had 16 target-or-filler stimuli: 5 or 6 target stimuli with the neutral contour, 5 or 6 target stimuli with the rise-fall-rise contour, and 5 or 6 filler stimuli. Furthermore, each list had 6 control stimuli, for a total of 22 stimuli per list. The order of the stimuli in the three lists was pseudo-randomized. We then created 3 additional lists by reversing the order of the stimuli. Each of these 6 lists were split into two sub-lists (with roughly half of the target, filler and control stimuli of the list). A participant was assigned to a list, and could complete either one or both of the two sub-lists of that list.

### 3.3 Procedure

Participants were told that two friends, Julie and Mike, haven't seen each other for a while, that they are catching up over coffee and that Mike asks Julie questions to which she responds. Participants were instructed to listen to each discourse between Mike and Julie, and then to respond to a question about what Julie meant with her response; specifically, participants were asked whether she meant to convey the meaning of the adjective in Mike's question. Participants could listen to each discourse as often as they wanted.

Participants gave their response on a 7-point Likert scale labeled at four points: 1/Definitely No, 3/Perhaps No, 5/Perhaps Yes, 7/Definitely Yes. We assume that a participant's degree of belief in the scalar implicature is stronger the lower the response is. Figure 4 illustrates a sample stimulus and response task.

Since the adjectives realized in the control stimuli are antonyms, we expected the control stimuli to receive negative responses. (We took a correct answer to a control dialogue to be 1/Definitely No.) For the filler stimuli, on the other hand, we expected positive responses since the answer entails a positive response to the polar question. Given that Doran et al. (2012) and van Tiel et al. (2016) find relatively low implicature rates for scalar adjectives, we expected the target stimuli that were realized with the neutral contour to

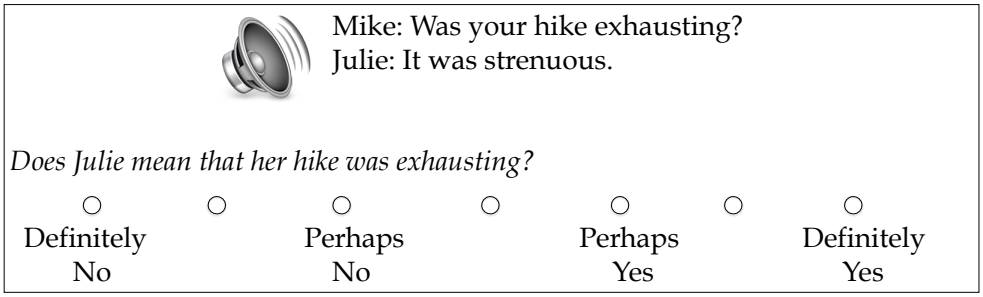



Figure 4: Sample stimulus and response task in the perception experiment

receive less positive responses than the filler stimuli but not overwhelmingly many negative responses. Finally, based on the intuitions reported to us by native speakers of American English, we expected that the target stimuli that were realized with the rise-fall-rise contour would receive more negative responses than the target stimuli with the neutral contour.

### 3.4 Results

Except for one target stimulus for which we only had 9 responses, we obtained 10 responses for each target and filler stimulus. The bar graph in Figure 5 shows how often each response option was chosen for the target stimuli with the neutral and rise-fall-rise (RFR) contours and the filler stimuli.

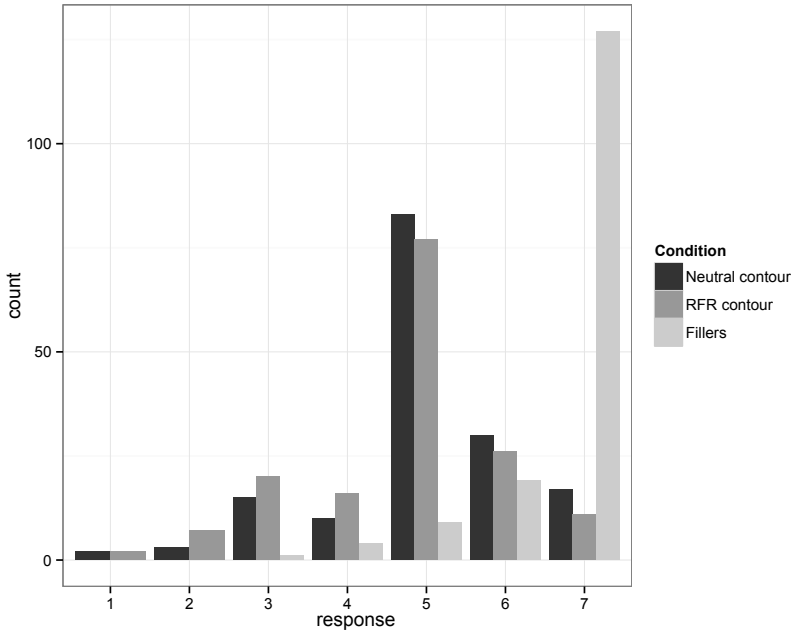



Figure 5: Distribution of responses to filler and target stimuli in the perception experiment

As expected, filler stimuli received highly positive responses (mean response: 6.67) and target stimuli receive less positive responses than the filler stimuli, though still largely positive responses (5/Perhaps Yes, 6 and 7/Definitely Yes). This result is in line with the finding in the previous literature (Doran et al.

2012, van Tiel et al. 2016) that utterances with gradable adjectives have relatively low implicature rates compared to other scalar expressions.[5] Also as expected, the target stimuli with the rise-fall-rise contour received less positive responses than the target stimuli with the neutral contour (mean response 4.77 versus 5.04, respectively). For the target stimuli, we ran an ordinal mixed-effects regression analysis predicting response from contour with by-participant and by-item random intercepts and random slopes for contour, using the `clmm` R `ordinal` package Christensen (2013) (R version 2.15.2). We found a significant effect of contour: responses to target stimuli with the rise-fall-rise contour were significantly lower than responses to target stimuli with the neutral contour ($\beta$ = -0.63, $SE$ = 0.24, $z$ = -2.60, $p < .01$).

This finding suggests that the rise-fall-rise contour strengthens the degree of belief in the scalar implicature over the neutral contour: listeners are more likely to infer that the speaker, Julie, intended a negative answer to Mike's question when Julie realizes her indirect answer with the rise-fall-rise contour than with the neutral one. Thus, in our dialogues, the prosodic realization of an indirect answer to a polar question affected its interpretation, with the rise-fall-rise contour leading listeners more toward negative answers than the neutral contour.

### 3.5 Discussion

Our perception experiment investigated the interpretation of indirect answers with scalar adjectives in response to polar questions with semantically stronger adjectives. The result of the perception experiment shows that when the indirect answer is realized with the rise-fall-rise contour, a listener is more likely to interpret the answer as conveying a negative response to the polar question than when the answer is realized with the neutral contour.

More generally, what these results show is that the prosodic realization of an indirect answer can indeed be a cue to the intended interpretation of the indirect answer with respect to the polar question. Our result provides evidence that a polar question $?p$ that is responded to with an indirect answer with meaning $q$ that neither entails $p$ nor $\neg p$ may be more likely to be taken to mean $\neg p$ when the indirect answer is realized with the rise-fall-rise contour than with the neutral contour. We have thus provided confirmation for the speculation in Green and Carberry 1999:fn34 that the prosody of an indirect answer "can help in recognizing the speaker's intentions".

One avenue for future investigation is whether our result generalizes to pairs of questions and indirect answers that do not involve scalar adjectives. Do the two contours we have investigated lead to similar effects whenever the two contours are realized on indirect answers to polar questions, regardless of whether the indirect answer realizes a gradable adjective, or are these effects particular to the kinds of question-answer pairs we have investigated? For instance, is the indirect answer in (1A3), repeated below for convenience, more likely to be interpreted as a negative response when realized with the rise-fall-rise contour than with the neutral contour?

(1) Q: Do you want some coffee?
A3: Coffee would keep me awake.

[5] To compare the rate with which implicatures arose in this experiment with the implicature rates for scalar adjectives reported in the previous literature, we recoded the responses as a '1' (implicature) if participants responded 1/Definitely No, 2 or 3/Possibly No, and as a '0' (no implicature) if they responded 5/Possibly Yes, 6 or 7/Definitely Yes (responses of '4' were discarded). With this recoding, we found an implicature rate of 20% for target stimuli with the rise-fall-rise contour and of 13% for target stimuli with the neutral contour. This implicature rate is somewhat higher than that reported for gradable adjectives in van Tiel et al. 2016 but comparable to that reported in Doran et al. 2012.

An answer to this question depends to a large extent on the answer to a question we have largely side-stepped in this paper until this point, namely the question of which meaning the rise-fall-rise contour contributes to an utterance, and how that meaning interacts with the interpretation of indirect answers. In the next section, we discuss the meaning of the rise-fall-rise contour in view of our experimental finding, and also address the contribution of our work to scalar implicature generation more generally.

# 4 Implications of the experimental finding

## 4.1 The meaning of the rise-fall-rise contour

The key finding of our perception experiment was that indirect answers are interpreted more negatively in response to polar questions when the answer is realized with the rise-fall-rise contour than when it is realized with the neutral contour. In this section, we review three analyses of the meaning of the rise-fall-rise contour with respect to whether they predict this experimental finding.[6]

### 4.1.1 Ward and Hirschberg 1985

In their seminal work on the meaning of the rise-fall-rise contour, Ward and Hirschberg (1985) propose that the contour conventionally implicates speaker uncertainty with respect to a scale evoked in the discourse (p.765): uncertainty about whether it is appropriate to evoke a scale at all (type I uncertainty), uncertainty about which scale to choose, given that some scale is appropriate (type II), or, given some scale *S*, uncertainty about the choice of some value on *S* (type III). Ward and Hirschberg (1985) also propose two felicity conditions for the rise-fall-rise contour (p.747): some item in the utterance with the contour must be perceived as linked by a scalar relationship with its context, and it must be contextually plausible that the speaker is uncertain with respect to the scale or scalar value evoked.

The first felicity condition is satisfied by our target stimuli since the scalar adjectives in the polar question and the indirect answers were chosen to be interpreted on the same scale, with the adjective in the question stronger than the adjective in the answer; cf., example (2).

(2) A: Is your sister beautiful?
B: She's attractive.

From the fact that the gradable adjectives in our target stimuli evoke a conspicuous scale, we conclude that it is not type I or II uncertainty that is conveyed by the rise-fall-rise contour in our target stimuli. While none of the examples discussed in Ward and Hirschberg 1985 mirror our target stimuli exactly, there are some examples that also feature a polar question with a semantically stronger expression than that realized in the indirect answer. In (8), their (25a), the expression *the first chapter* in the polar question is stronger than the expression *the first half of the first chapter* in the indirect answer. (We follow Ward and Hirschberg's convention of annotating examples uttered with the rise-fall-rise contour using \ to indicate the L*+H pitch accent and / to indicate the final rise, i.e., L-H%.) And in (9), their (61a), the expression *a hundred* in the polar question is stronger than the expression *a ninety-eight* in the indirect answer.

[6]For reasons of space, we discuss papers whose primary focus is the rise-fall-rise contour, to the exclusion of Büring 2003, Oshima 2008 and Wagner 2012. See Wolter 2003 for a proposal that the rise-fall-rise contour is licensed only on declarative utterances that fail to resolve an issue.

(8) A: Did you read the first chapter?
B: I read the first \half/ of it. (Ward and Hirschberg 1985:758)

(9) A: Did she get a hundred on the midterm?
B: She got a ninety-\eight/. (Ward and Hirschberg 1985:767)

Example (9) is discussed in Ward and Hirschberg 1985 as illustrating type III uncertainty: "B might be uncertain about whether *ninety-eight* and *a hundred* are close enough (on a cardinal scale) for A's purposes" (p.767). Although Ward and Hirschberg do not discuss the type of uncertainty involved in example (8), it is plausible that here, too, B conveys uncertainty about whether reading the first half of the first chapter is enough for A's purposes (see Figure 4 in Ward and Hirschberg 1985 for the scale involved in this example).

Applied to our target stimuli, Ward and Hirschberg's proposal predicts that, when B realizes their utterance with the rise-fall-rise contour, B is uncertain about the choice of the weaker adjective on the scale evoked by the two adjectives. Specifically, given the examples in (8) and (9), one possibility is that B conveys that she is uncertain about whether the weaker adjective is close enough, for A's purposes, on the scale evoked by the adjectives to the stronger adjective. Thus, under Ward and Hirschberg's account, B's utterance with the rise-fall-rise contour in (2) might convey that B is uncertain whether *attractive* is close enough to *beautiful* for A's purposes. This meaning (a conventional implicature according to Ward and Hirschberg) would be absent when B realizes their utterance with the neutral contour.

Does this account predict our experimental finding? That is, when B conveys this kind of uncertainty, does that strengthen a listener's degree of belief that B meant to deny the stronger adjective over and above what the utterance with the neutral contour conveys? It could, possibly, if we make the additional assumption that it only makes sense for B to convey this kind of uncertainty when the weaker adjective is the strongest value on the relevant scale that she is committed to. This assumption is plausible since it is implausible for a speaker to convey a lower value on a scale and uncertainty about whether that value is close enough to the higher value if the speaker was committed to the higher value and could have just uttered the stronger adjective to convey the higher value. So, in example (2), this would mean that using the rise-fall-rise contour not only conveys that B is uncertain whether *attractive* is close enough to *beautiful* for A's purposes, but also that *attractive* is the strongest value on the relevant scale that B is committed to. Thus, under this additional assumption, B's use of the rise-fall-rise contour conveys that B is not committed to the stronger adjective.

There are two ways to take the implication that B conveys that she is not committed to the stronger adjective. A first possibility is to strengthen this implication to the implication that B conveys that she is committed to the stronger adjective being false (cf., Sauerland's 2005 epistemic step). But, if this was the case, we would expect our experiment participants to have given many negative responses to the target stimuli with the rise-fall-rise contour in our perception experiment (cf., Figure 5). Since that is not the case (recall that the response mean for these stimuli is 4.77), it does not seem plausible that listeners take B to convey that the stronger adjective is false. A second possibility is that the implication that B conveys that they are not committed to the stronger adjective is taken as just that: B may or may not be committed to the stronger adjective. But in this case we do not expect listeners to interpret B's utterances with the rise-fall-rise contour as more negative than B's utterances with the neutral contour, contrary to our experimental finding. We therefore conclude that Ward and Hirschberg's (1985) account does not predict our experimental finding.

### 4.1.2 Constant 2012

Constant (2012) proposes that the rise-fall-rise contour is a focus quantifier over alternatives of the uttered sentence that remain assertable after the utterance has been made; the relevant alternatives that are quantified over are called 'post-assertion alternatives'. Specifically, Constant proposes that the contour presupposes that there are post-assertion alternatives and conventionally implicates that the speaker cannot safely claim any of these alternatives. In (10), for instance, *John* is the focus of B's utterance and alternatives to the uttered sentence are of the form *X liked the movie*, with *X* an individual from a contextually restricted set (e.g., the set consisting of B and his friends). The asserted content of (10B) is that John liked the movie and this proposition is removed from the set of alternatives to obtain the post-assertion alternatives, which is a set of propositions of the form *X liked the movie*, with *X* a friend of B other than John. By using the rise-fall-rise contour, the speaker conventionally implicates that none of the post-assertion alternatives can be safely claimed, for example because the speaker knows that the other friends didn't like the movie, or because the speaker doesn't know whether the other friends liked the movie. (We continue to annotate the rise-fall-rise contour using the \ / notation.)

(10) A: Did your friends like the movie?
B: \John/ liked it. (adapted from Constant 2012:414)

In Constant's example (26a), given in (11), the rise-fall-rise contour is realized on the adjective *good*, which is a weaker adjective than, for instance, *perfect* (no context is given for this example):

(11) The food was \good/. (adapted from Constant 2012:414)

According to Constant 2012:415, the relevant alternative propositions for this example are of the form *The food was X*, including *The food was* {*perfect* | *good* | *mediocre* | *bad*}, i.e., alternatives on a scale of quality. The rise-fall-rise contour is felicitous in (11) because "the fact that the food was good doesn't resolve a relevant alternative, like *The food was perfect*" (p.415).[7] In other words, the rise-fall-rise contour is felicitous in (11) because the set of post-assertion alternatives is not empty: the literal meaning of (11) does not resolve all of the relevant alternatives; only the alternatives *The food was* {*good* | *mediocre* | *bad*} are resolved. By using the rise-fall-rise contour in (11), the speaker conventionally implicates that the post-assertion alternative, here the proposition that the food was perfect, cannot be safely claimed, for example, because the speaker takes it to be false.

Applied to our target stimuli, Constant's account predicts that, when B realizes their utterance with the rise-fall-rise contour, B conventionally implicates that post-assertion alternatives cannot be safely claimed. In example (2), for instance, the relevant alternatives could be propositions of the form *She is* {*beautiful* | *attractive* | *plain* | *ugly*}. The literal meaning of B's utterance of *She's attractive* resolves all of the alternatives except for {*She's beautiful*}, which therefore is the post-assertion alternative that B conventionally implicates cannot be safely claimed, e.g., because the speaker takes it to be false. When B realizes their utterance with the neutral contour, this conventional implicature is absent.

Does this account predict our experimental finding? If the post-assertion alternative with the stronger adjective (e.g., *She's beautiful*) cannot be safely claimed because it is false, we expect B's indirect answer utterances with the rise-fall-rise contour to be interpreted by our experiment participants as denying the

[7] According to Constant 2012:415, a proposition $p$ resolves a proposition $q$ if $p \rightarrow q$ or $p \rightarrow \neg q$.

stronger adjective. Instead, what we found is that B's indirect answer utterances with the rise-fall-rise contour receive largely positive interpretations, as already mentioned in connection with our discussion of Ward and Hirschberg's analysis in the previous section. This observation suggests that Constant's analysis may make too strong predictions for our target stimuli. What if, instead, the post-assertion alternative cannot be safely claimed by B because B does not know whether it is true? In that case, Constant's analysis does not predict that B's utterances with the rise-fall-rise contour are interpreted as more negative than B's utterances with the neutral contour.[8] We therefore conclude that Constant's 2012 account does not predict our experimental finding.

#### 4.1.3 Wagner et al. 2013

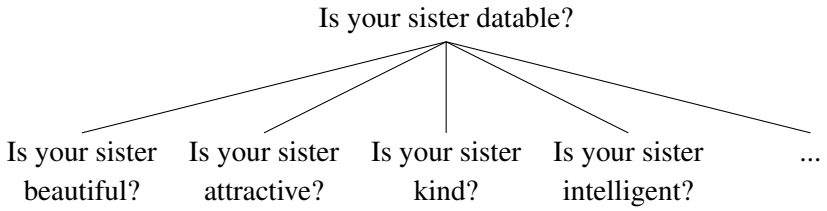
According to Wagner et al. (2013:142), when a speaker utters a sentence that denotes proposition *p* and realizes that utterance with the rise-fall-rise contour, the speaker asserts *p* and considers *p* to be only an incomplete answer to the question under discussion. In (12), for instance, the question under discussion is the content question uttered by Q. Since A realizes the answer with the rise-fall-rise contour, A conveys that *p*, the proposition that John solved the problem, is a partial answer to the question under discussion, i.e., other people may also have solved the problem. Wagner and his colleagues assume that if A had realized the utterance *John did* with the neutral contour (with a pitch accent on *John*), this would have implied that the answer was exhaustive, i.e., that "no one else but John solved the problem" (p.142).

(12) Q: Who solved the problem?
A: \John/ did. (adapted from Wagner et al. 2013:142)

To apply Wagner et al.'s analysis to our target stimuli, we first have to determine what the question under discussion might be for which the utterance with the rise-fall-rise contour is a partial answer. (On questions under discussion and their relation to prosody see, e.g., Roberts 2012, Büring 2003, 2016, Simons et al. to appear.) In contrast to (12), where the overt interrogative is the relevant question under discussion, it is not plausible to assume that the polar question in our target stimuli (e.g., *Is your sister beautiful?* in (2)) is the relevant question under discussion since the indirect answer (e.g., *She's attractive* in (2)) is not a partial answer to that question. We assume, instead, that the question under discussion in our target stimuli is implicit and, in particular, is a question to which the polar question is a sub-question (i.e., the implicit question under discussion entails the polar question; cf. Roberts 2012). What this means is that a complete answer to the implicit question under discussion entails an answer to the polar question, and an answer to the polar question is a partial answer to the implicit question under discussion. For instance, we assume that the implicit question under discussion for the dialogue in (2) is *Is your sister datable?*, with possible sub-questions as in (13). A combination of an answer to each of the sub-questions constitutes an answer to the question under discussion. We assume that both the question 'Is your sister beautiful?' and the question 'Is your sister attractive?' are part of this strategy of questions: the first is part of the strategy because it is the overt interrogative uttered by A; the second is part of the strategy because it is the question to which B's answer is a direct answer.

[8] The same reasoning applies when the set of post-assertion alternatives contains more than one alternative, e.g., *She is* {*drop dead gorgeous* | *beautiful*}, where we take *drop dead gorgeous* to be stronger than *beautiful*.

(13)




According to Wagner et al.'s analysis, when B responds to the polar question (e.g., *Is your sister beautiful?*) with the indirect answer (e.g., *She's attractive*) realized with the rise-fall-rise contour, B conveys that their answer to the question under discussion (e.g., *Is your sister datable?*) is partially answered. For instance, in our running example (2), the question under discussion is partially answered because the answer *She's attractive* is an answer to a sub-question of the question under discussion (namely to the implicit question 'Is your sister attractive?'). At the same time, other sub-questions are not yet answered, thereby rendering B's answer a partial answer to the question under discussion. When B realizes their indirect answer with the neutral contour, the implication that the answer is a partial answer to some question under discussion does not arise.

Does this account predict our experimental finding? That is, does the fact that B's use of the rise-fall-rise contour implies that the question under discussion is only partially answered strengthen a listener's degree of belief that B meant to deny the stronger adjective over and above what the utterance with the neutral contour conveys? That does not seem to be the case: B being committed to this implication is compatible both with B being committed to the stronger adjective and with B denying the stronger adjective. Consider example (2): if B conveys, by using the rise-fall-rise contour, that the proposition that the sister is attractive is only a partial answer to the question under discussion of whether the sister is datable, then B may still be committed to the sister being beautiful or to the sister not being beautiful. Since the meaning of the rise-fall-rise contour, under Wagner et al.'s analysis does not influence B's commitment state to the meaning of the stronger adjective, we conclude that Wagner et al.'s analysis does not account for our experimental finding.

#### 4.1.4 Summary

Given our experimental finding in section 3, a fruitful avenue for future research on indirect answers is to look for cues to the interpretation of an indirect answer in the prosodic realization of the indirect answer. However, in order to study how the prosodic realization of an indirect answer may influence the interpretation of the indirect answer, it is vital to understand the contribution to meaning that a particular prosody makes. In the previous sections, we have reviewed three accounts of the rise-fall-rise contour with respect to the question of whether these accounts predict our experimental finding. We found that Ward and Hirschberg's (1985) account and Constant's (2012) account either make too strong predictions for our data or no predictions, and that Wagner et al.'s (2013) account seems to make no predictions. A full assessment of accounts of the meaning of the rise-fall-rise contour must, of course, go beyond an assessment of how these analyses account for our experimental finding. We cannot do so here for reasons of space, but conclude that, in the context of research on the meaning of the rise-fall-rise contour, our experimental finding is a novel data point on the basis of which accounts of the meaning of the contour can be assessed.

## 4.2 The influence of prosody on scalar implicature generation

Our perception experiment built on prior research on the contributions of prosody to scalar implicature generation. As discussed in section 2, prior literature has already established that whether a listener takes an utterance with a weaker scalar expression to give rise to a scalar implicature depends on the prosodic realization of the scalar expression as well as on whether the scalar expression is part of the focus of the uttered sentence. In our perception experiment, in contrast, the target stimuli did not vary with respect to whether the scalar expression was realized with a pitch accent or with respect to whether the scalar expression was part of the focus of the indirect answer utterance: the weaker scalar adjectives in our target stimuli were always realized with a pitch accent and were always part of the focus. Rather, what varied in our target stimuli was the prosodic contour of the indirect answer. Our experiment thus established that it is not just the presence of a pitch accent or the realization as part of the focus that matters for whether a listener takes an utterance to give rise to a scalar implicature, but the prosodic contour of the utterances that contains the weaker scalar item.

However, our perception experiment differed in one crucial aspect from the experiments reported on by Chevallier et al. (2008), Thorward (2009), Zondervan (2010) and Schwarz et al. (to appear): whereas their experiments involved truth value judgment paradigms, our perception experiment relied on an implication judgment paradigm (i.e., experiment participants had to judge whether a particular implication arose; for discussion of implication judgment paradigms see Tonhauser et al. 2013 and Tonhauser and Matthewson 2015). To confirm that the finding that scalar implicature generation is sensitive to the prosodic contour of the utterance with the scalar expression, we conducted a follow-up experiment to the perception experiment.

### 4.2.1 Methods

**Participants** 35 individuals participated in the experiment. We discarded the data from one participant who did not consent and one who did not indicate that (s)he was a native American English speaker. This left us with data from 33 participants (12 female, 21 male; ages: 20-59, median age: 31).

**Stimuli** The target stimuli were Julie's utterances from the 32 target stimuli in the perception experiment described in section 3. We also used six of the control stimuli from the perception experiment to assess whether participants were paying attention to the task.

**Procedure** Experiment participants were told that Julie thinks that the stronger adjective is true (e.g., that her sister is beautiful). They then listened to Julie uttering the sentence with the weaker scalar adjective to Mike (e.g., Mike was told by Julie about her sister: *She's attractive*). Participants were asked to judge whether what Julie said is true. A 'no' response meant that the participant took Julie to convey that the stronger adjective is false (e.g., 'attractive but not beautiful'), i.e., the participant generated the scalar implicature. A 'yes' response, on the other hand, meant that the participant did not generate the implicature.

### 4.2.2 Results

11 of the 132 responses to target stimuli with the neutral contour were 'no', in contrast to 22 responses of the 132 responses to target stimuli with the rise-fall-rise contour.[9] Thus, there was a numeric increase

[9]The implicature rate in this experiment is 16% for utterances with the rise-fall-rise contour and 8% for utterances with the neutral contour. These implicature rates are lower than those observed in the perception experiment described in section 3. This

in implicature rate with the rise-fall-rise contour over the neutral contour. A logistic regression model predicting response from contour with by-participant and by-item random intercepts with random slope for contour for the by-item intercept revealed a significant difference between the contours: Julie's answers realized with the rise-fall-rise contour receive significantly more 'no' responses than answers realized with the neutral contour ($\beta = 1.79$, $SE = 0.62$, $z = 2.87$, $p < .01$). However, it is important to note that this effect is not consistently found for all experiment participants: 16 of the 33 participants replied 'yes' to all stimuli, thus not generating any implicatures, and two participants categorically generated an implicature when the indirect answer was realized with the rise-fall-rise contour, but not when the indirect answer was realized with the neutral contour. In the regression model, we can take into consideration that participants differ in their sensitivity to the two contours by including random slope for contour in the by-participant random intercept. And, indeed, a logistic regression model predicting response from contour with by-participant and by-item random intercepts and random slopes for contour did not show a significant effect of contour.

#### 4.2.3 Discussion

These results show that whether an utterance with a weaker scalar adjective is realized with the neutral contour or with the rise-fall-rise contour may influence whether a listener generates a scalar implicature. Thus, scalar implicature generation may not only be sensitive to whether the scalar item is realized with a pitch accent or in focus, but may also be sensitive to the contour of the utterance that realizes the weaker scalar expression.

## 5 Conclusions

Polar questions frequently receive indirect answers: Stenström (1984) and de Marneffe et al. (2009) find that 13% versus 14% of the polar questions in their data receive indirect answers, and Hockey et al. (1997) report an even higher percentage, namely 38%. The interpretation of an indirect answer to a polar question is influenced by a number of factors, including the literal meaning of the answer, what the listener knows about the speaker and utterances following the indirect answer. In this paper, we have provided empirical evidence that the prosodic realization of the indirect answer may provide a cue to the intended meaning of the indirect answer. Theoretical models of the interpretation of indirect answers thus need to incorporate information about the prosodic realization of the indirect answer. We have also discussed the implications of our experimental finding for the meaning of the rise-fall-rise contour and for scalar implicature generation. Our finding also has implications for natural language processing: dialogue generation systems, for instance, may benefit from taking into consideration the prosodic realization of the indirect answer to a polar question.

finding suggests that scalar implicature rate is influenced by the experimental paradigm, as also noted in Geurts and Pouscoulous 2009 and Degen and Goodman 2014.

# A Materials of the norming studies

The 30 pairs of adjectives tested in the norming studies are given in Table 1 with sample carrier sentences. An asterisk (∗) identifies those pairs of adjectives that did not pass a diagnostic.

| Adjective pairs | even | but not | Sample carrier sentences |
|---|---|---|---|
| <obsessive, dedicated> | | * | Ann's sister is dedicated, even/but not obsessive |
| <hostile, contentious> | * | * | Bart's debate was contentious, even/but not hostile |
| <faithful, devout> | * | * | Carl's wife is devout, even/but not faithful |
| <terrifying, intimidating> | | | Chris' opponent was intimidating, even/but not terrifying |
| <beautiful, attractive> | | | Don's sister is attractive, even/but not beautiful |
| <prominent, famous> | | | Finn's boss is famous, even/but not prominent |
| <fatal, malignant> | | * | Gwen's cancer is malignant, even/but not fatal |
| <dangerous, risky> | * | | Hugh's mission was risky, even/but not dangerous |
| <obnoxious, annoying> | * | | Jack's uncle is annoying, even/but not obnoxious |
| <life-threatening, serious> | | | Kaye's illness was serious, even/but not life-threatening |
| <hysterical, anxious> | | | Kim's patient was anxious, even/but not hysterical |
| <decrepit, shabby> | | | Lee's house is shabby, even/but not decrepit |
| <exhausting, strenuous> | | | Lynn's hike was strenuous, even/but not exhausting |
| <starving, hungry> | | | Mark's dog was hungry, even/but not starving |
| <brilliant, intelligent> | | | Nick's professor is intelligent, even/but not brilliant |
| <unforgettable, memorable> | * | * | Pam's honeymoon was memorable, even/but not unforgettable |
| <ridiculous, silly> | * | | Pat's costume was silly, even/but not ridiculous |
| <horrific, unsettling> | * | | Rick's nightmare was unsettling, even/but not horrific |
| <hot, good-looking> | | | Sam's date was good-looking, even/but not hot |
| <petrified, scared> | | | Stu's daughter was scared, even/but not petrified |
| <hideous, ugly> | | | Sue's bridesmaid dress was ugly, even/but not hideous |
| <toxic, inedible> | | | This mushroom is inedible, even/but not toxic |
| <corrupt, shady> | * | | This politician is shady, even/but not corrupt |
| <deadly, harmful> | | | This toxin is harmful, even/but not deadly |
| <disgusting, unpleasant> | | | Tim's bathroom was unpleasant, even/but not disgusting |
| <articulate, understandable> | | | Tom's interview was understandable, even/but not articulate |
| <impeccable, polished> | | | Tom's speech was polished, even/but not impeccable |
| <delicious, tasty> | | * | Val's cake was tasty, even/but not delicious |
| <hilarious, funny> | | | Will's blog post was funny, even/but not hilarious |
| <filthy, dirty> | | | Zack's carpet was dirty, even/but not filthy |

Table 1: Pairs of adjectives tested in the norming studies

The controls used in the two norming studies are given in Table 2.

| 'even' study | Sentences used to create "not odd" control stimuli |
|---|---|
| | Blair's room is comfortable. It's even cozy. |
| | Troy's mother is pleasant. She's even charming. |
| | Jeff's pants are snug. They're even tight. |
| | Eve's relationship is special. It's even unique. |
| | Paige's soup is warm. It's even hot. |
| | **Sentences used to create "odd" control stimuli** |
| | Elle's exam scores were perfect. They were even below average. |
| | Jane's dog is well-behaved. He is even aggressive. |
| | Grant's proposal was well-received. It was even dismissed. |
| | Pierce's car was cheap. It was even pricy. |
| | John's uncle is rich. He's even poor. |

| 'but not' study | "non-contradictory" control stimuli |
|---|---|
| | Jeff's mother is happy and creative. |
| | Eve's husband is shy and handsome. |
| | Blair's sofa is comfortable and old. |
| | Paige's cat is long-haired and lazy. |
| | Troy's beer is warm and flat. |
| | **"contradictory" control stimuli** |
| | Jane's blouse is new and used. |
| | Elle's door is open and closed. |
| | Grant's towel is wet and dry. |
| | Pierce's car was expensive and cheap. |
| | This nail is bent and straight. |

Table 2: Control stimuli used in the two norming studies

# B Materials of the perception experiment

The 16 question-answer pairs that were recorded as target stimuli are given in Table 3, with the stressed syllable of the adjective in the answer capitalized. The filler stimuli differ from the target stimuli in that the weaker adjective of the pair is realized in the question and the stronger adjective is realized in the answer.

The question-answer pairs used for the control stimuli are given in Table 4, with the stressed syllable of the adjective in the answer capitalized.

| | Question (Mike) | Answer (Julie) |
|---|---|---|
| 1 | Was your date hot? | He was good-LOOKing. |
| 2 | Was your hike exhausting? | It was STREnuous. |
| 3 | Was your patient hysterical? | She was ANxious. |
| 4 | Was your opponent terrifying? | He was inTImidating. |
| 5 | Was your interview articulate? | It was underSTANdable. |
| 6 | Is your sister beautiful? | She's atTRACtive. |
| 7 | Was your speech impeccable? | It was POlished. |
| 8 | Was your carpet filthy? | It was DIRty. |
| 9 | Is your house decrepit? | It's SHABby. |
| 10 | Was your bathroom disgusting? | It was unPLEAsant. |
| 11 | Was your blog post hilarious? | It was FUNny. |
| 12 | Was your dog starving? | It was HUNgry. |
| 13 | Is your professor brilliant? | She's inTELligent |
| 14 | Was your bridesmaid dress hideous? | It was UGly. |
| 15 | Is your boss famous? | She's PROminent. |
| 16 | Was your sister's illness life-threatening? | It was SErious. |

Table 3: Question-answer pairs for target stimuli in perception experiment

| Question (Mike) | Answer (Julie) |
|---|---|
| Is your cell phone plan cheap? | It's exPENsive. |
| Was your suitcase light? | It was HEAvy. |
| Is your boss energetic? | He's leTHARgic. |
| Was your brother rude? | He was reSPECTful. |
| Was your sister innocent? | She was GUILty. |
| Was your umbrella big? | It was TIny. |

Table 4: Question-answer pairs for control stimuli in perception experiment

# References


AnderBois, Scott. 2016. Negation, alternatives, and negative polar questions in American English. In E. Onea, M. Zimmermann, and K. von Heusinger, eds., *Questions in Discourse*. Leiden: Brill.

Beckman, Mary E. and Gayle Ayers Elam. 1997. *Guidelines for ToBI labelling, version 3.0*. The Ohio State University. http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf.

Büring, Daniel. 1997. *The Meaning of Topic and Focus — The 59th Street Bridge Accent*. London: Routledge.

Büring, Daniel. 2003. On D-trees, beans and B-accents. *Linguistics & Philosophy* 26:511–545.

Büring, Daniel. 2016. Focus, questions and givenness. In E. Onea, M. Zimmermann, and K. von Heusinger, eds., *Questions in Discourse*. Leiden: Brill.

Chevallier, Coralie, Ira A. Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti, and Dan Sperber. 2008. Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology* 61(11):1741–1760.

Christensen, Rune Haubo Bojesen. 2013. "ordinal": Regression models for ordinal data. R package. Version 2013.9-30.

Clark, Herbert H. 1979. Responding to indirect speech acts. *Cognitive Psychology* 11:430–477.

Constant, Noah. 2012. English rise-fall-rise: A study in the semantics and pragmatics of intonation. *Linguistics & Philosophy* 35:407–442.

de Marneffe, Marie-Catherine, Scott Grimm, and Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 136–143. Association for Computational Linguistics.

Degen, Judith. 2015. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics & Pragmatics* 8(11):1–55.

Degen, Judith and Noah D. Goodman. 2014. Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Doran, Ryan, Rachel E. Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1:211–248.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel paradigm for distinguishing between what is said and what is implicated. *Language* 88:124–154.

Gazdar, Gerald. 1979. *Pragmatics: Implicature, Presuppositions, and Logical Form*. New York: Academic Press.

Geurts, Bart and Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics & Pragmatics* 2(4):1–34.

Goodhue, Daniel and Michael Wagner. 2014. It's not just what you say, it's how you say it: Intonation, *yes* and *no*. Talk presented at 45th Annual Meeting of the *North East Linguistics Society*.

Goodman, Noah D. and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5:173–184.

Green, Nancy and Sandra Carberry. 1992. Conversational implicatures in indirect replies. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 64–71. Newark, Delaware, USA: Association for Computational Linguistics.

Green, Nancy and Sandra Carberry. 1994. A hybrid reasoning model for indirect answers. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 58–65.

Green, Nancy and Sandra Carberry. 1999. Interpreting and generating indirect answers. *Computational Linguistics* 25:389–435.

Groenendijk, Jeroen and Martin Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Hirschberg, Julia B. 1985. *A Theory of Scalar Implicature*. Ph.D. thesis, University of Pennsylvania.

Hobbs, Jerry R. and Jane J. Robinson. 1979. Why ask? *Discourse Processes* 2:311–318.

Hockey, Beth Ann, Deborah Rossen-Knill, Beverly Spejewski, Matthew Stone, and Stephen Isard. 1997. Can you predict responses to yes/no questions? Yes, no, and stuff. In *Proceedings of Eurospeech 1997*.

Horn, Lawrence. 1972. *On the Semantic Properties of Logical Operators in English*. Bloomington, Indiana: Indiana University Linguistics Club.

Horn, Laurence. 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.

Krifka, Manfred. 2013. Response particles as propositional anaphors. In *Proceedings of Semantics and Linguistic Theory (SALT) XXIII*, pages 1–18.

Oshima, David Y. 2008. Morphological vs. phonological contrastive topic marking. In *Proceedings of the Chicago Linguistic Society*, vol. 41-1, pages 371–384. Chicago: Chicago Linguistic Society.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5:1–69. Reprint of 1996 publication.

Sauerland, Uli. 2005. The epistemic step. Talk presented at *Experimental Pragmatics*, Cambridge University, April 2005.

Schwarz, Florian, Chuck Clifton, and Lyn Frazier. to appear. Strengthening 'or': Effects of focus and downward entailing contexts on scalar implicatures. In F. S. J. Anderssen, K. Moulton and C. Ussery, eds., *Semantics and Processing (UMOP 37)*. Amherst, MA: GLSA Publications.

Simons, Mandy, David Beaver, Craige Roberts, and Judith Tonhauser. to appear. The Best Question: Explaining the projection behavior of factive verbs. *Discourse Processes* .

Sperber, Dan and Deirdre Wilson. 1986. *Relevance. Communication and Cognition*. Cambridge, MA: Harvard University Press.

Stenström, Anna-Brita. 1984. *Questions and Responses in English Conversation*. Lund: Lund University Press.

Thorward, Jennifer. 2009. *The interaction of contrastive stress and grammatical context in child English speakers' interpretations of existential quantifiers*. B.A. thesis, The Ohio State University.

Tonhauser, Judith, David Beaver, Craige Roberts, and Mandy Simons. 2013. Toward a taxonomy of projective content. *Language* 89:66–109.

Tonhauser, Judith and Lisa Matthewson. 2015. Empirical evidence in research on meaning. Unpublished ms., The Ohio State University, University of British Columbia.

van Kuppevelt, Jan. 1996. Inferring from topics. Scalar implicatures as topic-dependent inferences. *Linguistics & Philosophy* 19:393–443.

van Rooij, Robert. 2002. Relevance implicatures. Ms., ILLC, Amsterdam.

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33:137–175.

Wagner, Michael. 2012. Contrastive topics decomposed. *Semantics & Pragmatics* 5(8):1–54.

Wagner, Michael, Elise McClay, and Lauren Mak. 2013. Incomplete answers and the rise-fall-rise contour. In *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*.

Ward, Gregory and Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language* 61(4):747–776.

Wolter, Lynsey. 2003. Fall-rise, topic, and speaker noncommitment. In *Proceedings of Western Conference on Linguistics (WECOL)*, vol. 14, pages 322–333.

Zondervan, Albert Jan. 2010. *Scalar implicatures or focus: An experimental approach*. Ph.D. thesis, University of Utrecht.