Thinking & Reasoning, 2014
http://dx.doi.org/10.1080/13546783.2014.965746

Some but not all dispreferred turn markers help to interpret scalar terms in polite contexts

Jean-François Bonnefon[1], Ethan Dahl[2], and Thomas M. Holtgraves[2]

[1]CLLE, University of Toulouse and Centre National de la Recherche Scientifique, Toulouse, France
[2]Department of Psychology, Ball State University, Muncie, IN, USA

In polite contexts, people find it difficult to perceive whether they can derive scalar inferences from what others say (e.g., does "some people hated your idea" mean that not everyone hated it?). Because this uncertainty can lead to costly misunderstandings, it is important to identify the cues people can rely on to solve their interpretative problem. In this article, we consider two such cues: Making a long Pause before the statement, and prefacing the statement with Well. Data from eight experiments show that Pauses are more effective than Wells as cues to scalar inferences in polite contexts—because they appear to give a specific signal to switch expectations in the direction of bad news, whereas Well appears to give a generic signal to make extra processing effort. We consider the applied value of these findings for human–human and human–machine interaction, as well as their implications for the study of reasoning and discourse.

Keywords: Scalar inference; Politeness; Discourse marker; Everyday reasoning.

Imagine that a company spokesperson announced that "the company will have to let go of some of the employees at the Paris branch". Later, it is revealed that the plan had always been to let go of all the employees. Was the spokesperson honest about the intentions of the company?

Whether in ordinary conversation or in official discourse, what is left unsaid is often as important as what is actually said. To consider what the speaker could have said but chose not to, allows one to derive inferences about what is really meant. From the perspective of linguistic pragmatics,

Correspondence should be addressed to Jean-François Bonnefon, CLLE, Maison de la recherche, 5 allées A. Machado, 31058, Toulouse Cedex 9, France. E-mail: bonnefon@univ-tlse2.fr

The third author acknowledges support for this research from the National Science Foundation [grant number BCS-1224553].

these inferences typically follow the Standard Recipe of quantity implicatures (Geurts, 2010): When a speaker makes a claim x rather than a stronger claim y, then the speaker can be assumed to believe that y is false.

Scalar inferences (Horn, 1984) are a special and important instance of this standard recipe. Let us consider a claim x such that there exists one claim y which is broadly the same length as x and more informative than x, in the sense that y entails x but x does not entail y. When a speaker makes the claim x, she is assumed not to believe y as per the Standard Recipe. For example, the claims (1-a-c) respectively invite the scalar inferences that the speaker believes (2-a-c) to be false:

(1)
a. Some guests brought wine to the party;
b. I will visit you on Monday or Tuesday;
c. I will probably be in Paris next week.

(2)
a. All guests brought wine to the party;
b. I will visit you on Monday and Tuesday;
c. I will most definitely be in Paris next week.

Scalar inferences are extremely robust and compelling—so much so that it could be tempting to consider them as default inferences, automatically retrieved with the lexical content of words such as some, or, and possibly (Levinson, 2000). Experimental data, though, convincingly demonstrated that scalar inferences were not retrieved by default (although see Grodner, Klein, Carbary, & Tanenhaus, 2010; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). Research generally showed that scalar inferences required time and processing effort (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; De Neys & Schaeken, 2007; Huang & Snedeker, 2009), and that different contexts could increase or decrease their likelihood (Breheny et al., 2006; Chevallier et al., 2008).

One important class of such contexts has to do with politeness. It has been repeatedly demonstrated that scalar terms are less likely to generate a scalar inference when they contain a face-threatening content such as a criticism, an imposition, or a piece of bad news (Bonnefon, Feeney, & Villejoubert, 2009; Bonnefon & Villejoubert, 2006; Feeney & Bonnefon, 2013; Juanchich & Sirota, 2013; Juanchich, Sirota, & Butler, 2012; Pighin & Bonnefon, 2011; Sirota & Juanchich, 2012). Consider the following examples:

(3)
a. Some of the guests hated your recipe;
b. We will take away your annual bonus or your company car;
c. Your bad breath is possibly an issue in social settings.

In all these examples, it is not quite clear whether the speaker intends the listener to derive a scalar inference, or whether the speaker is tactfully communicating the stronger claim (all the guests hated the recipe; both the bonus and the company car are going away; the listener's bad breath is most definitely an issue).

The confusion arises because of the challenge of communicating threatening information. Consider the situation where the listener cooked dinner, and wants feedback from the speaker. As it happened, all the guests hated the listener's recipe. A perfectly sincere speaker would go for a bold, on-record statement such as "everyone hated your recipe". A perfectly polite speaker would rather go for an evasive statement such as "I don't really know what people thought" or even a white lie such as "you did great". Using a scalar statement such as "some of the guests hated your recipe" is not as clear-cut as a sincere or polite statement: It could either be meant sincerely or politely, and people are typically unsure about how to interpret these statements (see Bonnefon, 2014, for a review).

As argued in Bonnefon, Feeney, and De Neys (2011), this uncertainty increases the risk of costly misunderstandings in high-stake situations, which are the very situations in which people tend to express themselves tactfully. As a consequence, it is important to identify the cues that people use to detect whether a scalar term truly invites a scalar inference, or whether it is only used politely (Demeure, Bonnefon, & Raufaste, 2009). This article focuses on two such potential cues, specifically two discourse markers that can preface scalar statements: The word "Well", and a long Pause.

We focus on these two markers because they can signal dispreferred conversation turns. Dispreferred conversation turns consist of speech acts that take the opposite direction to what would be ideally expected from a cooperative interaction (Davidson, 1984; Fox Tree, 2010; Pomerantz, 1984). Typical examples are disagreements, rejections of offers, and expressions of blame. Dispreferred turns can be marked by delays and hesitations ("hehh", silent pauses), space takers ("well", "you know"), or more specific forms, like token agreements before contradiction ("yes, but … "). Dispreferred turn markers can facilitate the detection of polite criticism (Holtgraves, 2000), and we reasoned that they might likewise facilitate the resolution of scalar inferences in face-threatening contexts. Consider for example the following exchange:

(4)
a. Student: What did you think of my term paper?
b. Teacher: Well, that was a very difficult assignment.

The presence of Well signals a dispreferred conversation turn, and makes it easier to interpret (4-b) as "I did not like your paper" (see Holtgraves,

2000, for experimental evidence of the impact of "Well" in such statements). We expect a similar effect of dispreferred markers on exchanges featuring the scalar term "some". For example, we expect that the markers will help people differentiate between responses (5-b) and (5-c), or between responses (6-b) and (6-c):

(5)
a. What did people think of my idea?
b. Well, some people loved your idea.
c. Well, some people hated your idea.

(6)
a. What did people think of my idea?
b. [Long Pause] Some people loved your idea.
c. [Long Pause] Some people hated your idea.

We expect that the markers will emphasise the difference between the two responses, helping people to interpret "some people hated your idea" as "some and possibly everyone hated your idea", while helping them to interpret "some people loved your idea" as "not all people loved your idea". That is, we expect the markers to amplify the valence effect already obtained in previous research (Bonnefon et al., 2009). We know that negative valence pulls interpretation away from the scalar inference, compared to positive valence, but this pull effect is not strong enough and results in ambiguity. We predict that dispreferred turn markers should increase the pull and disambiguate the meaning of the utterance. In the rest of this article, this prediction is tested in a first series of experiments. Experiments 1 and 2 investigate the effect of Wells and Pauses, respectively, and Experiment 3 directly compares the effects of the two markers. This first series of experiments concludes that only Pauses appear to have the expected effect, and an interim discussion considers a potential explanation for this asymmetry. This explanation is then tested in a second series of five experiments.

EXPERIMENT 1

Method

Participants (93 men and 73 women, age range 18−75, $M = 32.8$) were recruited in the USA through the Mechanical Turk platform, and compensated \$0.10 for their participation. In all experiments, participants indicated whether or not they were native speakers of English and all analyses were restricted to those reporting English as their first language. All experiments were run exclusively on US samples. Participation in more than one of the

experiments was technically possible but rare—when it occurred, only the data from a participant's first experiment were used, and his or her data from subsequent experiments were excluded. In all experiments, participants saw only one vignette.

In all experiments, the attrition rate was 0% in all conditions, that is, no participant ever stopped before completing a study. This is quite probably due to the brevity of the task, since only one critical item was presented to participants. In addition to the critical item, participants in all experiments also responded to a single item designed to serve as an attention check. Participants were instructed to leave this item blank and those who failed to do so were excluded from the analysis. The percentage of participants failing the attention check ranged between 4% and 15% ($M = 8\%$) over the experiments. Analyses of the distribution of attention check failures over conditions were conducted for each experiment. There were no significant differences in the distribution of attention check failures for any of the experiments (all $p > .25$).

Participants were randomly assigned to one group of a 2 × 2 between-participant design manipulating the valence of the target statement (Love vs. Hate) and the presence or absence of "Well" before the target statement. They read a vignette featuring the target statement and were asked for their agreement with the scalar inference based on this statement. For example, participants in the Hate−Well group read the following vignette and question (emphasis added):

> Yesterday, you pitched an idea to a group of five persons. Today, you ask Bob (who was in the group) what people thought of your idea. Bob replies: "Well, some people hated your idea." How likely is it that everyone in the room hated your idea?

In the Love condition, both instances of "hated" were replaced with "loved". In the No Marker condition, the word "well" was omitted. Participants responded on a scale from −5 totally unlikely, to +5 totally likely. Higher ratings thus denote the rejection of the standard scalar inference, and lower ratings denote acceptance of the scalar inference.

Results

Figure 1 (left panel) displays the mean perceived likelihood of "all loved/hated" in each experimental group. We conducted an analysis of variance in which the predictors were the valence of the target statement and the presence or absence of Well. This analysis detected a main effect of valence (consistent with Bonnefon et al., 2009), $F(1, 162) = 19.1$, $p < .001$, $\eta_p^2 = .11$; and an interaction between valence and the presence of Well, $F(1, 162) =$

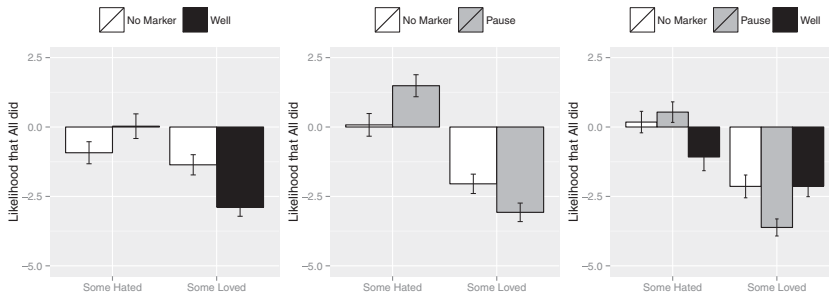

Figure 1. Main results of Experiments 1 (left), 2 (centre), and 3 (right). The dependent variable is the agreement with the interpretation that all loved (or hated), when told that some loved (or hated). Higher values thus correspond to rejection of the classic scalar inference. Pauses amplify the effect of valence (scalar inferences are more likely to be rejected for "hate" than for "love"), whereas the effect of Well is inconsistent across experiments.

10.5, $p = .001$, $\eta_p^2 = .06$. The presence of Well did not have any detectable main effect, $F < 1$, $p > .45$.

Follow-up analyses showed that the presence of Well decreased the perceived likelihood that "all loved" (from $M = -1.4$, mean standard error (MSE) = 0.4 to $M = -2.9$, MSE = 0.3; $t(89) = -3.1$, $p = .003$). The presence of Well did not significantly impact the perceived likelihood that "all hated" (from $M = -0.9$, MSE = 0.4 to $M = +0.1$, MSE = 0.4; $t(73) = 1.6$, $p = .11$). In other words, the presence of Well encouraged scalar inferences from positive statements ("some people loved your idea"), without discouraging scalar inferences from negative statements ("some people hated your idea").

Before we attempt to interpret these results, we will investigate the effect of Pauses in Experiment 2, and then directly compare the effects of Wells and Pauses in Experiment 3.

EXPERIMENT 2

Method

Participants (92 men and 68 women, age range 18–64, $M = 29.4$) were recruited in the USA through the Mechanical Turk platform, and compensated \$0.10 for their participation. They were randomly assigned to one group of a 2 × 2 between-participant design manipulating the valence of the target statement (Love vs. Hate) and the presence or absence of a Pause before the target statement. They read a vignette featuring the target statement and were asked for their agreement with the scalar inference based on this statement. For example, participants in the Hate–Pause group read the following vignette and question (emphasis added):

> Yesterday, you pitched an idea to a group of five persons. Today, you ask Bob (who was in the group) what people thought of your idea. Bob stays silent for a few seconds. Then he replies: "Some people hated your idea." How likely is it that everyone in the room hated your idea?

In the Love condition, both instances of "hated" were replaced with "loved". In the No Marker condition, the emphasised text was replaced with "Bob replies". Participants responded on a scale from −5 totally unlikely, to +5 totally likely. Higher ratings thus denote the rejection of the standard scalar inference, and lower ratings denote the acceptance of the scalar inference.

Results

Figure 1 (central panel) displays the mean perceived likelihood of "all loved/hated" in each experimental group. We conducted an analysis of variance in which the predictors were the valence of the target statement and the presence or absence of the Pause. This analysis detected a main effect of valence, F (1, 156) = 80.8, $p < .001$, $\eta_p^2 = .34$; and an interaction between valence and the presence of a Pause, F (1, 156) = 10.7, $p = .001$, $\eta_p^2 = .06$. The presence of a Pause did not have any detectable main effect, $F < 1$, $p > .60$.

Follow-up analyses showed that the presence of a Pause decreased the perceived likelihood that "all loved" (from $M = -2.0$, MSE = 0.3 to $M = -3.1$, MSE = 0.3; $t(82) = -2.1$, $p = .04$). In contrast, the presence of a Pause increased the perceived likelihood that "all hated" (from $M = -0.1$, MSE = 0.4 to $M = +1.5$, MSE = 0.4; $t(74) = 2.5$, $p = .02$). In other words, the presence of a Pause amplified the effect of valence by encouraging scalar inferences from positive statements, and discouraging scalar inferences from negative statements.

EXPERIMENT 3

Method

Experiment 3 served both as a replication of Experiments 1 and 2, and as an opportunity to directly compare the effects of Wells and Pauses within a single experiment. Participants (125 men, 107 women, 2 undisclosed, age range 18−75, $M = 30.4$) were recruited in the USA through the Mechanical Turk platform, and compensated $0.10 for their participation.

Participants were randomly assigned to one group of a 2 × 3 between-participant design manipulating the valence of the target statement (Love vs. Hate) and the presence of a marker before the target statement (Well vs.

Pause vs. No Marker). The material and procedure were similar to that already introduced in the descriptions of Experiments 1 and 2.

Results

Figure 1 (right panel) displays the mean perceived likelihood of "all loved/hated" in each experimental group. We first conducted an overall analysis of variance using the full 2 × 3 design. This analysis detected the familiar pattern of a main effect of valence,[1] $F\,(1, 228) = 58.8$, $p < .001$, $\eta_p^2 = .20$; no main effect of marker condition, $F\,(2, 228) = 1.3$, $p > .28$, $\eta_p^2 = .001$; and an interaction between valence and marker condition, $F\,(2, 228) = 46.4$, $p < .001$, $\eta_p^2 = .06$.

To better understand this interaction, we ran separate analyses of variance on subsets of the data excluding in turn the Well group, the Pause group, and the No Marker group. The first sub-analysis (focusing on the Pause and No Marker groups) replicated the results of Experiment 2: Main effect of valence, $F\,(1, 152) = 71.8$, $p < .001$, $\eta_p^2 = .32$; no main effect of Pause, $F\,(1, 152) = 1.5$, $p > .21$, $\eta_p^2 = .001$; and an interaction between valence and Pause, $F\,(1, 152) = 5.9$, $p < .02$, $\eta_p^2 = .04$. Follow-up analyses showed that the presence of a Pause decreased the perceived likelihood that "all loved" (from $M = -2.1$, MSE = 0.4 to $M = -3.6$, MSE = 0.3; $t(68) = -2.9$, $p < .001$). In contrast, the presence of a Pause non-significantly increased the perceived likelihood that "all hated" (from $M = 0.2$, MSE = 0.4 to $M = 0.5$, MSE = 0.4; $t(84) = 0.7$, $p = .50$).

The second sub-analysis (focusing on the Well and No Marker groups) failed to replicate the results of Experiment 1. It only detected a main effect of valence, $F\,(1, 155) = 17.1$, $p < .001$, $\eta_p^2 = .09$; but no main effect of Well,

[1] A reader suggested that this valence effect might be due to a simple response bias in favour of "everyone hated", which would have nothing to do with the presence of the scalar statement. To investigate the existence of such a response bias, we conducted an experiment in which we changed our vignette to completely exclude the "Bob replies" sentence. Participants (29 men and 25 women, age range 20–54, $M = 29.4$) were recruited as in our other experiments. Participants in the Hate (resp., Love) group read the following vignette and question:

> Yesterday, you pitched an idea to a group of five persons. Today, you ask Bob (who was in the group) what people thought of your idea. How likely is it that everyone in the room hated (resp., loved) your idea?

Participants responded on an 11-point scale anchored at *totally unlikely* and *totally likely*. Participants' ratings were higher in the Love condition ($M = +0.7$, MSE = 0.4) than in the Hate condition ($M = -1.6$, MSE = 0.4), $t(52) = 4.4$, $p < .001$. This is evidence for a response bias in favour of "everyone loved", rather than "everyone hated". Accordingly, higher ratings in the Hate conditions of our other experiments cannot be explained by a response bias in favour of "everyone hated".

F (1, 155) = 2.4, $p > .12$, $\eta_p^2 < .02$; and most importantly no interaction between valence and Well, $F(1, 155) = 2.3$, $p > .13$, $\eta_p^2 < .02$.

The third and final sub-analysis (focusing on the Pause and Well groups) confirmed that the relation between the Pause and Well groups was comparable to that between the Pause and No Marker groups. It detected a main effect of valence, F (1, 149) = 41.5, $p < .001$, $\eta_p^2 = .22$; no main effect of marker, $F < 1$, $p > .77$, $\eta_p^2 = .001$; and an interaction between valence and marker, $F(1, 155) = 15.1$, $p < .001$, $\eta_p^2 = .09$.

In sum, Experiment 3 confirmed that Pauses amplified the valence effect in the interpretation of scalar inferences—but it also cast further doubts that Wells could have a similar effect. Contrary to our expectation that Wells and Pauses, both being dispreferred turn markers, would have similar effects, we have found evidence that only Pauses are effective for disambiguating scalar expressions in polite contexts. We will now consider a possible explanation for this difference, before moving on to a second series of experiments linked to this explanation.

INTERIM DISCUSSION

Although Wells and Pauses can both be used as dispreferred turn markers, subtle differences in their pragmatic functions might account for the better performance of Pauses with respect to the disambiguation of scalar expressions in polite contexts. Indeed, previous research (reviewed below) suggests that the two markers may orient interpretation through different cognitive pathways. In both pathways, the marker functions as a signpost that orients processing. The two pathways, though, differ in terms of what signal the marker sends to listeners: Well encourages further interpreting effort of the utterance, whereas Pauses signal an unexpected completion of the utterance. We now consider the implications of this distinction, how it accounts for the results of Experiments 1–3, and the additional predictions it affords, which will be tested in Experiments 4a and 4b and 5a–5c.

We first consider the cognitive pathway through which Well produces its interpretative effects. Various authors suggested that the effect of Well was to orient the processing of the following utterance toward what amounts to further interpreting effort. For Jucker (1993), Well signals that the most relevant interpretation of what follows is probably not the correct one—and therefore, that additional effort must be engaged to reach an apparently irrelevant interpretation. For Blakemore (2002), Well signals that an apparently irrelevant interpretation of the upcoming utterance is actually relevant—and therefore, that additional effort must be engaged to find this interpretation. Finally, for Bronwen (2010), Well signals to the listener that the interpretation of the upcoming utterance will require further inferencing (and accordingly, further processing effort) than usual.

In contrast, Pauses (among other disfluent hesitations such as "um" or "uh") can signal that the next part of the message is unexpected (Arnold, Altmann, Fagnano, & Tanenhaus, 2004; Corley & Stewart, 2008). That is, Pauses can prepare listeners for an atypical, low-probability utterance. Evidence for this claim comes from neurophysiological studies that presented statements ending with predictable or unpredictable words, and analysed the N400 difference between these conditions as a function of whether the statement included a disfluent hesitation (Corley, MacGregor, & Donaldson, 2007; MacGregor, Corley, & Donaldson, 2010). Statements that included Pauses attenuated the N400 effect, suggesting that participants were prepared for an unexpected completion. Now, our hypothesis is that all other things being equal, and barring special circumstances, negative feedback is less expected than positive feedback due to its social awkwardness. Accordingly, in the context of valenced feedback, a Pause (which signals an unexpected completion) may prepare listeners to negative feedback, and therefore prepare them to adopt the least favourable interpretation of the following utterance.

Let us consider in this light the expected impact of Wells and Pauses on the interpretation of positive scalar statements (7-a), neutral scalar statements (7-b), and negative scalar statements (7-c):

(7)

a. Some people loved your idea;
b. Some people bought tickets;
c. Some people hated your idea.

A marker whose effect is to encourage cognitive effort would increase the rate of scalar inferences from positive and neutral statements such as (7-a). This prediction derives from previous findings which showed that scalar inferences in neutral (Bott & Noveck, 2004; De Neys & Schaeken, 2007) or positive context (Bonnefon, De Neys, & Feeney, 2011) were linked to greater cognitive effort. It is not as clear, though, whether cognitive effort will increase or decrease the rate of scalar inferences from negative statements such as (7-c). Indeed, Bonnefon et al. (2011) suggested that in the case of negative statement, the effect of cognitive effort is likely to be non-monotonic. That is, cognitive effort in small measure may increase the rate of scalar inferences, whereas a large measure of cognitive effort may decrease the rate of scalar inferences.

A marker whose effect is to orient listeners toward the least favourable interpretation of the following utterance would increase the rate of scalar inferences from positive statements such as (7-a), decrease the rate of scalar inferences from negative statements such as (7-c), but would not have any effect on the rate of scalar inferences from neutral statements such as

(7-b). Indeed, the least favourable interpretation of (7-a) is that not all loved the idea (a scalar inference), whereas the least favourable interpretation of (7-c) is that all hated the idea (no scalar inference). Because the statement (7-b) is neutral, its scalar interpretation is neither more nor less favourable than its non-scalar interpretation. Finally, a marker that prepares listeners for an unfavourable utterance should prompt them to expect a negative, rather than positive statement. Consider an incomplete statement such as

(8) Some people . . . your idea

Prefacing this statement with a marker that prepares the listener for an unfavourable utterance, should prompt the listener to fill (8) with "hated", rather than "loved". A marker whose effect is to encourage cognitive effort would not have a similar impact. Armed with this analysis of the two cognitive pathways, we are in a position to account for the results of Experiments 1–3, and to make additional predictions for a second series of experiments (see Table 1 for a summary of all predictions). We concluded that a marker encouraging cognitive effort, such as Well, would:

- Increase the rate of scalar inferences from positive statements;
- Have inconsistent effects on scalar inferences from negative statements;
- Increase the rate of scalar inferences from neutral statements;
- Have no effect on the completion of sentences such as "some . . . your idea".

So far the first two predictions were grounded in the data of Experiments 1–3. The third prediction will be tested in Experiment 4a, and the fourth prediction will be tested in Experiments 5b and 5c.

We also concluded that a marker preparing for an unfavourable utterance, such as a Pause, would:

TABLE 1
Summary of all predictions derived from the interim discussion

	Pause	*Well*
Positive statements	More scalar inferences	More scalar inferences
Neutral statements	No effect	More scalar inferences
Negative statements	Less scalar inferences	Inconsistent effects
Incomplete statements	Preferred negative completion	No preferred completion

- Increase the rate of scalar inferences from positive statements;
- Decrease the rate of scalar inferences from negative statements;
- Have no effect on the rate of scalar inferences from neutral statements;
- Encourage the "hated" completion of sentences such as "some ... your idea".

So far the first two predictions were broadly validated by Experiments 1−3. The third prediction will be tested in Experiment 4b, and the fourth prediction will be tested in Experiment 5a.

EXPERIMENT 4a

Method

Participants (53 men and 32 women, age range 19−54, $M = 27.1$) were recruited in the USA through the Mechanical Turk platform, and compensated \$0.10 for their participation. They were randomly assigned to the Marker or No Marker group. Participants in the Marker group read the following vignette and question (emphasis added):

> Yesterday an announcement was made at your work that tickets for the company picnic would be on sale that afternoon. The next day you ask Bob if people went to purchase tickets. Bob replies: "Well, some people went to purchase tickets." How likely is it that everyone went to purchase tickets?

In the No Marker group, the word "Well" was omitted. Participants responded on a scale from −5 totally unlikely, to +5 totally likely. Higher ratings thus denote the rejection of the standard scalar inference.

Results

The presence of the marker Well decreased the perceived likelihood that everyone purchased tickets (from $M = -1.9$, MSE = 0.3 to $M = -3.0$, MSE = 0.3; $t(83) = -2.2$, $p = .03$). Accordingly and as expected, the marker Well encouraged scalar inferences from a neutral content, which was neither pleasant nor unpleasant to the listener.

EXPERIMENT 4b

Method

Participants (39 men and 34 women, age range 18−72, $M = 29.5$) were recruited in the USA through the Mechanical Turk platform, and

compensated $0.10 for their participation. They were randomly assigned to the Marker or No Marker group. Participants in the Marker group read the following vignette and question (emphasis added):

> Yesterday an announcement was made at your work that tickets for the company picnic would be on sale that afternoon. The next day you ask Bob if people went to purchase tickets. Bob stays silent for a few seconds. Then he replies: "Some people went to purchase tickets." How likely is it that everyone went to purchase tickets?

In the No Marker group, the emphasised text was replaced with "Bob replies". Participants responded on a scale from −5 totally unlikely, to +5 totally likely. Higher ratings thus denote the rejection of the standard scalar inference.

Results

The presence of a Pause made no difference to perceived likelihood that everyone purchased tickets ($M = -2.6$, MSE = 0.4 with a Pause, $M = -2.5$, MSE = 0.4 without a Pause; $t(71) = 0.2$, $p = .81$). In contrast to what we observed with the marker Well in Experiment 4a, the presence of a Pause did not encourage scalar inferences from a neutral content. The power of this experiment to detect an effect comparable to that observed in Experiment 4a was .67 (compared to .73 in Experiment 4a). Whereas the two experiments were not designed to detect an interaction effect, we tentatively pooled their data to run an analysis of variance (ANOVA) on the acceptance of scalar inference as a function of sample and presence of a marker. The interaction term fell short of significance at $p = .11$. Overall, we interpret these results as broadly confirming that the presence of a Pause was not detected to have a similar effect as that detected for Well.

EXPERIMENT 5a

Method

Participants (44 men and 32 women, age range 19−66, $M = 32.2$) were recruited in the USA through the Mechanical Turk platform, and compensated $0.10 for their participation. They were randomly assigned to the Marker or No Marker group. Participants in the Marker group read the following vignette (emphasis added):

> Yesterday, you pitched an idea to a group of 5 colleagues. Today, you ask Bob (who was in that group) what people thought of your idea. Bob stays silent for a few seconds. Then he replies: "Some people . . . your idea."

In the No Marker group, the emphasised text was replaced with "Bob replies". Participants were asked whether the most likely completion (to fill in the dots) was "loved" or "hated".

Results

In the No Marker condition, 15 participants out of 37 chose the "hated" completion. This proportion (41%) was not significantly different from chance (binomial, $p = .32$). In the Marker condition, though, 27 participants out of 39 chose the "hated" completion, and this proportion (69%) was significantly greater than chance (binomial, $p = .02$). The introduction of a Pause significantly increased the proportion of "hated" completions ($\chi^2 = 6.3, p = .01, F = .29$).

Data thus support our hypothesis that silent Pauses would shift expectations toward negative feedback. It remains to be tested, though, whether Wells can have a similar effect. Indeed, if this effect is to account for the difference between Wells and Pauses in Experiments 1–3, it should not be observed for Wells. Experiments 5b and 5c report two attempts to detect for Wells the effect that Experiment 5a detected for Pauses.

EXPERIMENT 5b

Method

Participants (47 men and 43 women, age range 19–64, $M = 32.4$) were recruited in the USA through the Mechanical Turk platform, and compensated \$0.10 for their participation. They were randomly assigned to the Marker or No Marker group. Participants in the Marker group read the following vignette (emphasis added):

> Yesterday, you pitched an idea to a group of 5 colleagues. Today, you ask Bob (who was in that group) what people thought of your idea. Bob replies: "Well, some people ... your idea."

In the No Marker group, the word "Well" was omitted. Participants were asked whether the most likely completion (to fill in the dots) was "loved" or "hated".

Results

In the No Marker condition, 14 participants out of 47 chose the "hated" completion. This proportion (30%) was significantly different from chance (binomial, $p = .008$). In the Marker condition, 15 participants out of 43

chose the "hated" completion, a proportion (32%) which was not significantly different from chance (binomial, $p = .07$). Overall, the introduction of the marker did not make any detectable difference to the proportion of "hated" completions ($\chi^2 = 0.6$, $p = .61$, $\Phi = .05$). Experiment 5b had a power of .98 to detect an effect comparable as that observed in Experiment 5a. However, in order to consolidate this null result, we conducted an exact replication of this experiment, which we report as Experiment 5c below.

EXPERIMENT 5c

Method

Experiment 5c is an exact replication of Experiment 5b, with a new sample of 40 men and 27 women, age range 18−55, $M = 28.5$.

Results

In the No Marker condition, 11 participants out of 32 chose the "hated" completion. This proportion (34%) was not significantly different from chance (binomial, $p = .11$). In the Marker condition, 14 participants out of 35 chose the "hated" completion, a proportion (40%) which was not significantly different from chance (binomial, $p = .31$). Overall, the introduction of the pause marker did not make any detectable difference to the proportion of "hated" completions ($\chi^2 = 0.6$, $p = .80$, $\Phi = .06$). Experiment 5c had a power of .92 to detect an effect comparable as that observed in Experiment 5a.

In sum, two independent studies confirmed that prefacing an incomplete scalar statement ("some people ... your idea") with Well did not shift expectation toward the negative completion "hated", in contrast to what we observed for silent pauses in Experiment 5a.

GENERAL DISCUSSION

Scalar inferences are robust, compelling, and a staple of efficient communication. In some contexts, though, people find it difficult to understand whether speakers intend them to derive scalar inferences. For example, people find it hard to interpret statements such as "some people hated your idea", when the scalar term some applies to face-threatening contents (e.g., criticism, imposition, bad news). In this case, people cannot very well perceive whether they ought to derive the scalar inference (i.e., some but not all hated their idea), or whether the speaker is politely conveying the stronger interpretation (i.e., some and possibly all hated their idea). As argued in Bonnefon, Feeney, and De Neys (2011), this uncertainty can lead to costly

misunderstandings, and it is important to identify the cues that people rely on to understand scalar terms in polite contexts. In this article, we considered two such cues: Making a long Pause before the statement, and prefacing the statement with Well.

A first series of three experiments tested whether two dispreferred turn markers (Well and a Pause) could facilitate the interpretation of scalars in face-threatening contexts. Only Pauses showed a reliable facilitation effect. We hypothesised that this specific impact of Pauses might be the result of shifting listeners' expectation toward negative feedback, whereas the effect of Well was to encourage further processing effort. A second series of five experiments confirmed these respective effects. Although the eight experiments give us a reasonably robust basis for deriving implications, we must acknowledge two limitations of this research programme. First, all our experiments were based on a single item, drawing on similar vignettes, which inevitably leads to concerns about potential material effects and generalisability. Second, we used written text to represent speech, and the presence or absence of a pause, in particular, was accomplished descriptively. Accordingly, the task remains to verify our results with auditory stimuli. Even without this verification, our findings still apply to the large amount of language that we encounter in written form (e.g., fictional or journalistic accounts of conversations). It is eminently desirable, though, for future research to investigate our pattern of results using a broader range of methods.

The practical and theoretical implications of our findings need to be addressed separately, starting with practical implications. Our results allowed us to conclude that a long Pause was an effective preface to a polite statement. That is, a long Pause signalled to listeners that an unpleasant statement was incoming, and prompted them to adopt the most unpleasant interpretation of this statement. This is a very useful thing to know when training agents to overcome the uncertainties of politeness, which is typically done in professional fields where clarity of expression is critical (e.g., airplane crews, emergency teams, nuclear operations personnel; Kanki, Helmreich, & Anca, 2010). Over and beyond the difficulties it poses for high-stake human-to-human communication, politeness was also identified as one of the challenges of human–computer interaction, especially when this interaction involves humanoid robots, or the three-dimensional virtual entities known as embodied conversational agents (Niewiadomski & Pelachaud, 2010; Nomura & Saeki, 2010; Rehm & André, 2007). In these fields too, it will be useful to know that a pause is a reliable way to prepare human users for a polite statement.

The pragmatic inference from "Some x are y" to "Some x are not y" is a pivotal element of all theories of syllogistic reasoning (Chater & Oaksford, 1999; Khemlani & Johnson-Laird, 2012; Politzer, 2011; Schmidt &

Thompson, 2008). More precisely, any theory of the syllogism must come to terms with the fact that reasoners inconsistently adopt the scalar interpretation of some, with consequences for the syllogisms they perceive as valid. Accordingly, theories of syllogistic reasoning directly benefit from results that clarify the circumstances under which the scalar interpretation of some is adopted. We already knew that content, and more precisely valence, affected the rate of scalar inferences from some (Bonnefon et al., 2009)—this result is now refined by considering how context (in the form of conversational markers) interacts with valence to produce nuanced effects.

Even more generally, the same logic applies to the thriving literature addressing valence effects in reasoning (e.g., Bonnefon, 2009, 2012; Bonnefon, Haigh, & Stewart, 2013; Bonnefon & Sloman, 2013; Corner, Hahn, & Oaksford, 2011; Egan & Byrne, 2012; Evans, Neilens, Handley, & Over, 2008; Thompson, Evans, & Handley, 2005). Valence has a direct relation with politeness: Almost by definition, negative contents are more likely to be threatening than positive contents, and more likely to trigger politeness-based interpretations as a result. Results that reveal moderators of politeness-based interpretations are accordingly beneficial to research that investigates valence effects on reasoning. Moderators such as that we investigated in this article are especially interesting for researchers wishing to explore reasoning in natural everyday contexts, in which premises do not come neatly packaged in syllogisms, but rather surrounded by conversational noise such as uh or well.

Aside from their practical implications and their interest for reasoning research, our results are of general interest for research on discourse markers. The corpus-based approach to discourse markers can sometimes lead to long lists of disparate and contradictory functions for a given marker (Fox Tree, 2010). This approach (a corpus-based identification of the functions of discourse markers) can be fruitfully complemented with experimental studies investigating the effects of discourse markers on interpretation. Indeed, experimental studies can help substantiate subtle hypotheses concerning the effect of discourse markers. In this article, the fact that two dispreferred turn markers had a different impact on the interpretation of scalars, led to further studies which supported a subtle distinction between the effects of these two markers on processing.

Our experimental work should still be complemented with corpus-based research. In particular, we could not find data on the respective frequencies of different dispreferred turn markers in the context of scalar expressions—which leaves open the possibility that the two markers had different effects because of their respective frequencies in our context of interest. Unlocking the core meaning of discourse markers indeed requires the careful

confrontation of experimental and corpus-based approaches. Experiments were contributed in this paper, but corpus data will have to await future research.

Manuscript received 24 March 2014
Revised manuscript received 29 August 2014
Revised manuscript accepted 10 September 2014
First published online 9 October 2014

REFERENCES

Arnold, J. E., Altmann, R., Fagnano, M., & Tanenhaus, M. K. (2004). The old and thee, uh, new. *Psychological Science, 15*, 578–582.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.

Bonnefon, J. F. (2009). A theory of utility conditionals: Paralogical reasoning from decision-theoretic leakage. *Psychological Review, 116*, 888–907.

Bonnefon, J. F. (2012). Utility conditionals as consequential arguments: A random sampling experiment. *Thinking & Reasoning, 18*, 379–393.

Bonnefon, J. F. (2014). Politeness and reasoning: Face, connectives, and quantifiers. In T. M. Holtgraves (Ed.), *Oxford handbook of language and social psychology*. New York: Oxford University Press.

Bonnefon, J. F., De Neys, W., & Feeney, A. (2011). Processing scalar inferences in face-threatening contexts. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3389–3394). Austin, TX: Cognitive Science Society.

Bonnefon, J. F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science, 20*, 321–324.

Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition, 112*, 249–258.

Bonnefon, J. F., Haigh, M., & Stewart, A. J. (2013). Utility templates for the interpretation of conditional statements. *Journal of Memory and Language, 68*, 350–361.

Bonnefon, J. F., & Sloman, S. A. (2013). The causal structure of utility conditionals. *Cognitive Science, 37*, 193–209.

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science, 17*, 747–751.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Cognition, 51*, 437–457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434–463.

Bronwen, I. (2010). "Well, that's why I asked the question sir": Well as a discourse marker in court. *Language in Society, 39*, 95–117.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38*, 191–258.

Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology, 61*, 1741–1760.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*, 658–668.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistic Compass, 2*, 589–602.

Corner, A. J., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language, 64*, 153–170.

Davidson, J. (1984). Subsequent versions of invitations, offers, requests and proposals dealing with potential or actual rejection. In J. M. Atkinson & J. C. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 102–128). Cambridge: Cambridge University Press.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load – dual task impact on scalar implicatures. *Experimental Psychology, 54*, 128–133.

Demeure, V., Bonnefon, J. F., & Raufaste, E. (2009). Politeness and conditional reasoning: Interpersonal cues to the indirect suppression of deductive inferences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*, 260–266.

Egan, S. M., & Byrne, R. M. J. (2012). Inferences from counterfactual threats and promises. *Experimental Psychology, 59*, 227–235.

Evans, J. St. B. T., Neilens, H., Handley, S. J., & Over, D. E. (2008). When can we say if? *Cognition, 108*, 100–116.

Feeney, A., & Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology, 32*, 181–190.

Fox Tree, J. E. (2010). Discourse markers across speakers and settings. *Language and Linguistic Compass, 4*, 269–281.

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some", and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42–55.

Holtgraves, T. M. (2000). Preference organization and reply comprehension. *Discourse Processes, 30*, 87–106.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context* (pp. 11–42). Washington, DC: Georgetown University Press.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology, 58*, 376–415.

Juanchich, M., & Sirota, M. (2013). Do people really say it is "likely" when they believe it is only "possible"? Effect of politeness on risk communication. *Quarterly Journal of Experimental Psychology, 66*, 1268–1275.

Juanchich, M., Sirota, M., & Butler, C. L. (2012). Effect of the perceived functions linguistic risk quantifiers on risk perception, severity and decision making. *Organizational Behaviour and Human Decision Processes, 118*, 72–81.

Jucker, A. H. (1993). The discourse marker well: A relevance-theoretical account. *Journal of Pragmatics, 19*, 435–452.

Kanki, B. G., Helmreich, R. L., & Anca, J. (2010). *Crew resource management* (2nd ed.). San Diego, CA: Academic Press.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin, 138*, 427–457.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia, 48*, 3982–3992.

Niewiadomski, C., & Pelachaud, C. (2010). Affect expression in ECAs: Application to politeness displays. *International Journal of Human-Computer Studies, 68*, 851–871.

Nomura, T., & Saeki, K. (2010). Effects of polite behaviors expressed by robots: A psychological experiment in Japan. *International Journal of Synthetic Emotions, 1*, 38–52.

Pighin, S., & Bonnefon, J. F. (2011). Facework and uncertain reasoning in health communication. *Patient Education and Counseling, 85*, 169–172.

Politzer, G. (2011). Solving natural syllogisms. In K. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reason* (pp. 19–36). Hove, UK: Psychology Press.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn-shapes. In J. M. Atkinson & J. C. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge: Cambridge University Press.

Rehm, M., & André, E. (2007). More than just a friendly phrase: Multimodal aspects of polite behavior in agents. In T. Nishida (Ed.), *Conversational informatics: An engineering approach* (pp. 69–84). Chichester: Wiley.

Schmidt, J., & Thompson, V. A. (2008). 'At least one' problem with 'some' formal reasoning paradigms. *Memory and Cognition, 36*, 217–229.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*, 109–147.

Sirota, M., & Juanchich, M. (2012). To what extent do politeness expectations shape risk perception? Even numerical probabilities are under their spell! *Acta Psychologica, 141*, 391–399.

Thompson, V. A., Evans, J. St. B. T., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language, 53*, 238–257.