Flexible Expectations of Speaker Informativeness Shape Pragmatic Inference

Sarah Fairchild and Anna Papafragou

1 Introduction

Human communication relies on shared expectations between speakers and hearers (Kuperberg and Jaeger 2016). One source of such expectations is the Cooperative Principle (CP), which consists of a set of maxims that state that interlocutors form their utterances such that they are as true (Maxim of Quality), informative (Maxim of Quantity), relevant (Maxim of Relevance), and clear (Maxim of Manner) as possible (Grice 1975). Because people expect speakers to follow these maxims, they will often pragmatically enrich the literal semantic meaning of an utterance that appears to be in violation of the CP, making an inference about what the speaker intended. For example, a sentence like "Some of my dogs bark" appears to violate the Maxim of Quantity: it is under-informative, because the speaker used the weaker term in a logical scale (*some*) when s/he could have used the stronger, more informative term (*all*). When comprehending this utterance, the listener typically assumes that the speaker did not intend the literal semantic meaning ("At least one (and possibly all) of my dogs bark"). Instead, the listener is likely to derive a scalar implicature (SI), inferring that the speaker intended to convey "Not all of my dogs bark."

Properties of the speaker are known to affect whether listeners compute SIs. For instance, individuals are less likely to derive an implicature from under-informative 'some' statements if they are led to believe, through prior linguistic context, that the speaker is not knowledgeable of the situation at hand (Bergen and Grodner 2012). For example, if a speaker were to say "At my client's request, I skimmed the investment reports" followed by "Some of the real estate investments lost money," the listener would be less likely to infer that "not all of the real estate investments lost money" for that speaker as compared to a speaker who meticulously compiled the reports. There is also some evidence that listeners can adapt to speaker-specific use of *some* and *many* to refer to various quantities of objects (Yildirim et al. 2015). For instance, three may be "some" for a particular individual, but another speaker may use "some" only for five or more objects. Critically, both of these examples involve listeners adjusting to a single speaker in some sort of adaptation period (reading context sentences or completing training trials). What is unclear is whether listeners also form expectations about speaker meaning based on speaker group identity (e.g., age, native language), in such a way that is stable across situational contexts and does not require an adaptation period. This would require overriding strong Gricean expectations on a regular basis, or having different sets of expectations for various groups of speakers. On the other hand, it may be the case that the expectations in the CP are only temporarily suspended in cases where the specific context at hand requires it.

We make the first investigation into speaker-group expectations in SI computation by comparing under-informative utterances spoken by native speakers to those spoken by non-native speakers, who may be assumed to be less pragmatically competent. We take our inspiration from research demonstrating that foreign-accented speech affects neural responses to syntactic violations, with P600 responses attenuated in response to violations produced by a non-native speaker (Hanulikova et al. 2012). If the prior expectations that guide syntactic processing also extend to higher-order language processing, under-informativeness may be accepted to a greater extent (with SIs possibly being derived less frequently) for non-native speakers. While the comprehender knows that there are better, more felicitous, ways of conveying the technically true information in the sentence, they may also assume that a non-native speaker is not fully aware of this. In Experiment 1, we investigate how the accent of a speaker affects SI computation, and in Experiment 2 we ask whether non-native speech alters the final interpretation of an under-informative utterance.

2 Experiment 1: Picture-Sentence Verification Task

In Experiment 1, we investigate how native speakers of English interpret pragmatically infelicitous sentences spoken by a native speaker of English and a native speaker of Mandarin Chinese using a picture-sentence verification task. Participants must respond as quickly as possible to whether a spoken description matches a previously presented image (e.g., "Some of these circles are green"

— an array of five green and three gray circles). Reaction times are measured. If listeners adjust their expectations to a speaker's identity and expect non-native speakers to produce pragmatically infelicitous descriptions (e.g., "Some of these circles are green" — an array of eight green circles), participants should respond that an infelicitous sentence is a bad description more slowly when it is spoken by a non-native speaker than when it is spoken by a native speaker.

2.1 Method

2.1.1 Participants

Fifty-two English monolinguals aged 18-20 (M = 18.56, SD = .712) participated in Experiment 1 for course credit. All were undergraduates at the University of Delaware.

2.1.2 Materials

One hundred sixty arrays of objects were created for Experiment 1. Each array featured eight simple objects arranged in two horizontal lines (circles, squares, stars, triangles, hearts, diamonds, crosses, or moons). In half of the pictures, every object had the same fill (red, orange, yellow, green, blue, purple, gray, black, striped, or checkered). In the other half, three of the eight objects were gray.

Eighty sentences were created to describe the arrays, each following the pattern "Some of these [shape]s are [fill]" or "All of these [shape]s are [fill]." Each sentence was recorded twice: once by a female native speaker of English (Native Speaker condition), and once by a female native speaker of Mandarin Chinese (Non-Native Speaker condition). Sentences and pictures were paired such that four Sentence Types were created: Infelicitous, Felicitous, False, True. Conditions are shown in Figure 1, below.

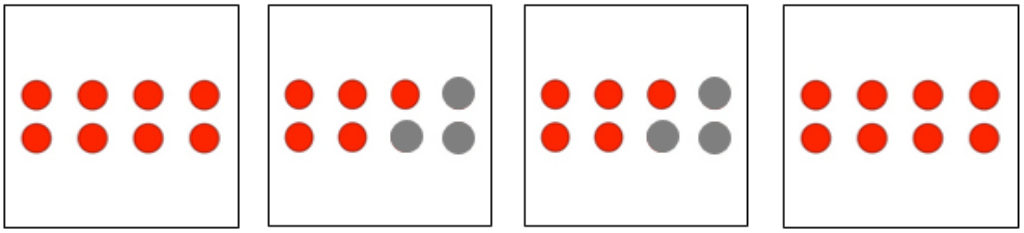

Figure 1: Sentence Types used in Experiment 1.

2.1.3 Procedure

Participants were tested individually or in pairs in a quiet room with the experimenter present. Each participant heard 160 sentences, 40 of each Sentence Type. Trials were presented in two blocks, one for each Speaker Type (counterbalanced across participants, stimuli fully rotated). Each trial began with a fixation for 500 milliseconds, followed by an array of objects for 500 milliseconds. After each array, a noise mask appeared for 200 milliseconds, and then a sentence played, accompanied by a fixation cross on the screen. The noise mask served to prevent an afterimage appearing when the participant was listening to the sentence. After the sentence, participants saw a screen with the words GOOD and BAD, and were asked to press a key as quickly as possible indicating whether the sentence was a good description of the picture, or a bad description. The next trial was presented after the participant made a response. If no response was detected within 3000 milliseconds, the experiment moved on to the next trial. Participants were given a break after the first block. The task was administered using Open Sesame presentation software (Mathôt et al. 2012).

The expected response for False trials was BAD, the expected response for True trials was GOOD, and the expected response for Felicitous trials was GOOD. If participants compute a scalar implicature, they should respond BAD to Infelicitous trials (e.g., all of the circles are red, not only

some). If, however, they take the logical interpretation they should respond GOOD to these trials (e.g., some of the circles are red, in fact they all are). If a speaker's accent affects online expectations about pragmatic felicity, participants should be faster to respond BAD to Infelicitous trials in the Native Speaker condition. Alternatively, if speaker identity does not affect the speed with which participants respond to Infelicitous trials may not differ across Speaker Types.

2.2 Results and Discussion

A repeated-measures ANOVA on the proportion of GOOD responses was performed with Sentence Type (Infelicitous, Felicitous, False, True) and Speaker Type (Native Speaker, Non-Native Speaker) as within-subjects factors. The main effect of Sentence Type was significant, $F(3, 153) = 325.374$, $p < .001$, but neither the main effect of Speaker Type nor the interaction reached significance. Follow-up tests (Bonferroni-corrected for multiple comparisons) revealed that False trials ($M = .04$, $SD = .05$) had a significantly lower proportion of GOOD responses than True ($M = .96$, $SD = .06$), Infelicitous ($M = .33$, $SD = .08$), and Felicitous ($M = .90$, $SD = .08$) trials (all p's $< .001$). True and Felicitous trials, which did not differ ($p > .1$), had a significantly higher proportion of GOOD responses than Infelicitous trials (both p's $< .001$).

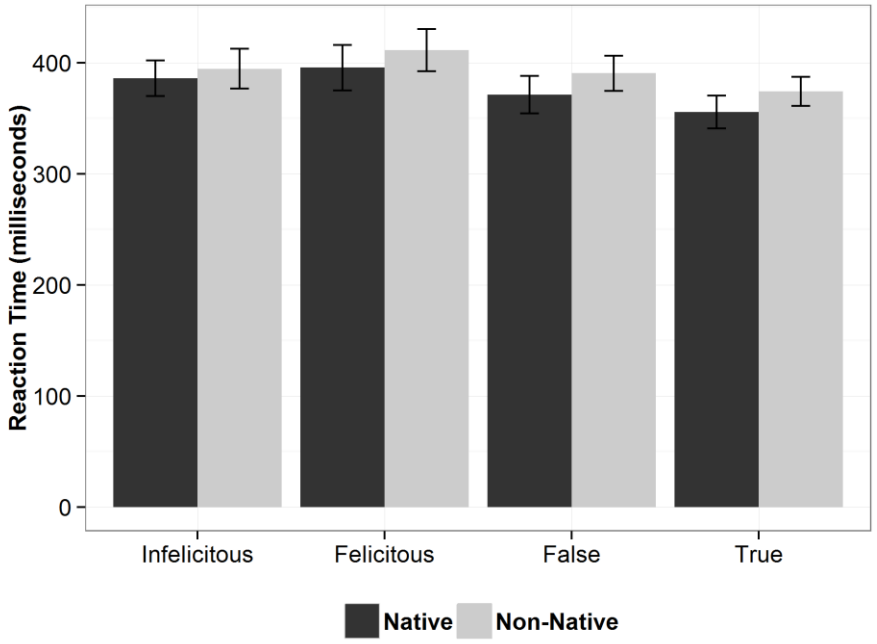

Figure 2: Reaction times for Experiment 1. Error bars represent +/- 1 S.E.M.

A complementary repeated-measures ANOVA on reaction times was then performed. Only "correct" responses were analyzed (Infelicitous — BAD, Felicitous — GOOD, False — BAD, True — GOOD). Although in the critical Infelicitous condition, there is no objectively correct answer, we analyzed BAD responses because we were interested in the speech of SI computation. Because 14 participants were logical interpreters who nearly always responded GOOD to Infelicitous trials, these individuals could not be included in the analysis, bringing our total number of participants to 38. Responses that exceeded 2500 milliseconds or 2.5 standard deviations above a participant's average response time were excluded. The ANOVA revealed a main effect of Sentence Type, $F(3, 111) = 5.774$, $p = .001$, but neither the main effect of Speaker Type nor the interaction reached significance. Post-hoc tests indicated that Felicitous ($M = 403.54$, $SD = 120.91$) and Infelicitous ($M = 390.31$, $SD = 103.52$) trials were responded to more slowly than True ($M = 364.91$, $SD = 86.23$) trials ($p < .001$ and $p = .035$, respectively. False trials ($M = 380.99$, $SD = 101.02$) did not differ from Felicitous, Infelicitous, or True trials. No other comparisons were significant. Results are depicted

in Figure 2.

The results of Experiment 1 suggest that a speaker's accent has no effect on the speed at which SIs are computed; reaction times for the critical Infelicitous condition did not differ by Speaker Type. Nevertheless, participants may have been engaged in another type of strategy during the task. Based on previous research (Bott and Noveck 2004), one would expect slower response times to the Infelicitous condition overall, reflecting the cost of computing a SI. Because reaction times reflect an aggregate of cognitive processes that occur in response to a particular stimulus, a purer test of speaker identity is to use a more time-sensitive measure like ERPs. Moreover, listeners may indeed form different expectations about pragmatic felicity based on native speaker status, but these expectations may not be integrated during the earliest moments of processing, and in a situation with visual context present. Because the task demands of Experiment 1 (timing, visual context) provided an extremely strict test of pragmatic expectations, we conducted Experiment 2 to determine whether listeners might adjust their offline interpretations of pragmatically infelicitous sentences depending on the speaker.

3 Experiment 2: Sentence Rating Task

In Experiment 2, we ask whether offline ratings of pragmatically infelicitous sentences ("Some giraffes have long necks") differ depending on whether the speaker is a native or a non-native speaker of English. If listeners expect non-native speakers to be more likely to produce infelicitous statements (in a situation devoid of visual context), they should rate these infelicitous sentences as more acceptable in the non-native speaker condition as compared to the native speaker condition. As a secondary goal, we sought to understand how individual characteristics of the listener might affect the extent to which they adjust to non-native speech. To address this issue, participants also completed measures of social-communicative ability (known to affect SI computation; Nieuwland et al. 2010) and language processing ability (known to affect non-native speech comprehension; Lev-Ari et al. 2016).

3.1 Method

3.1.1 Participants

Sixty English monolinguals aged 18-21 (M = 18.57, SD = .81) participated in Experiment 2 for course credit. All were undergraduate students at the University of Delaware.

3.1.2 Materials

Eighty sentences were created for the Sentence Ratings task. Sentences were evenly distributed across four conditions: False ("All women are doctors who went to medical school"), True ("All snow is cold and can melt into water"), Felicitous ("Some people have dogs as pets in the house"), and Infelicitous ("Some people have noses with two nostrils"). For the critical Infelicitous condition, sentences were created such that they were technically true but pragmatically infelicitous if one computes a SI. The other three conditions were either straightforwardly true (True, Felicitous) or false (False). Sentences in all four conditions were based on general knowledge subject matter, such that the truth value of the utterance is easily discernable by the typical college student. The four conditions did not differ from one another in sentence length in words or sentence length in syllables (all p's > .1). Each sentence was recorded twice: once by a female native speaker of English (Native Speaker condition) and once by a female native speaker of Mandarin Chinese (Non-Native Speaker condition). Recordings were made such that emphasis was not placed on any particular word.

The Communicative Subscale of the Autism-Quotient Questionnaire (AQ-COMM; Baron-Cohen Wheelwright Skinner Martin and Clubley 2001) was administered to all participants. It consists of 10 statements designed to probe social communication skills, such as "I am often the last to understand the point of a joke" and "I know how to tell if someone listening to me is getting bored." For each statement, participants indicate how true it is of themselves. The standard scoring method was used, calculating a total score out of 10 of the number of autistic traits the person possesses.

Participants also completed a Lexical Decision task as a measure of English comprehension

ability. Performance on this task has been correlated with various measures of language processing ability (Harrington 2006). Twenty English words (e.g., EDUCATION) and twenty pseudowords (e.g., EMUCATION) were presented in a random order for 500 milliseconds each. Participants were asked to indicate by keypress, as quickly as possible, whether or not each stimulus was a real English word. Both accuracy and reaction times were collected. For the analysis of reaction times, incorrect responses and responses slower than 3000 milliseconds were excluded.

3.1.3 Procedure

Participants were tested individually or in pairs in a quiet room with the experimenter present. The experimental session lasted 30 minutes, and consisted of the main Sentence Rating task followed by a Language History questionnaire to confirm that all participants were native speakers, the AQ-COMM, and the Lexical Decision task.

In the sentence rating task, participants listened to all 80 sentences divided into two blocks of 40 sentences each. In one block, participants heard sentences in the Native Speaker condition, and in another block, participants heard sentences in the Non-Native Speaker condition. Blocks were counterbalanced across participants and the sentences were fully rotated through Speaker Type. Sentences were presented in a random order in each block. On each trial, the sentence was presented auditorily through headphones. Then, participants were instructed to rate how "Good" or "Bad" the sentence was on a 5-point scale (1 = very bad, 5 = very good). Participants were asked to rate the sentences based on whether or not they made sense. After rating the sentence, participants pressed a button to move on to the next sentence. There was no time limit; participants could take as long as they needed to make a response. The sentence rating task and both questionnaires were administered using Qualtrics survey software. The Lexical Decision task was administered from the PEBL test battery (Mueller and Piper 2014).

3.2 Results and Discussion

A repeated-measures ANOVA was performed on the sentence ratings with Sentence Type (Infelicitous, Felicitous, True, False) and Speaker Type (Native, Non-Native) as within-subjects factors. The main effect of Sentence Type was significant, $F(3, 177) = 163.137$, $p < .001$. Post-hoc tests (Bonferroni corrected for multiple comparison) revealed that False sentences ($M = 2.13$, $SD = .99$) were rated lower than True ($M = 3.95$, $SD = .64$), Felicitous ($M = 4.11$, $SD = .62$) and Infelicitous ($M = 2.59$, $SD = .99$) sentences (all p's $< .001$). Infelicitous sentences were rated lower than True and Felicitous sentences (both p's $< .001$). True and Felicitous sentence ratings did not differ significantly from one another ($p > .05$). The main effect of Speaker was also significant, $F(1, 59) = 5.641$, $p = .021$, such that Non-Native sentences ($M = 3.14$, $SD = 1.12$) were rated lower than Native sentences ($M = 3.25$, $SD = 1.21$), $p = .009$. The interaction between Sentence Type and Speaker did not reach significance. Thus, a foreign accent affected sentence ratings overall, not specifically the interpretation of pragmatically Infelicitous statements.

Despite the lack of Speaker Type effect, visual inspection of the data revealed a great deal of variation in the extent to which participants rated sentences in the two Speaker Types as similar or different, even in the True, False, and Felicitous conditions which should yield similar ratings. Thus, an Accent Tolerance score was calculated by subtracting the average rating for Non-Native speech in the True, False, and Felicitous conditions from the average rating for Native speech in the same three conditions. Then, a median split was performed on the participants based on this rating. This resulted in an Accent Tolerant group who rated the two speakers as relatively similar and an Accent Intolerant group who rated the Native and Non-Native Speakers more differently in the unambiguous True, False, and Felicitous conditions.

Two repeated-measures ANOVAs were then performed separately for Accent Tolerant individuals ($N = 30$) and Accent Intolerant individuals ($N = 30$). The ANOVA for the Accent Intolerant participants revealed a main effect of Sentence Type, $F(3, 87) = 82.943$, $p < .001$, as well as a main effect of Speaker Type, $F(1, 29) = 43.070$, $p < .001$, with no significant interaction between the two. Specifically, post-hoc tests showed that False sentences were rated significantly lower than True ($p < .001$), Felicitous ($p < .001$), and Infelicitous ($p = .010$) sentences. Infelicitous sentences were rated significantly lower than both True and Felicitous sentences (both p's $< .001$), which did not differ

from one another ($p > .1$). Additionally, Non-Native speech was rated significantly lower than Native speech ($p < .001$).

The ANOVA for the Accent Tolerant individuals also revealed a main effect of Sentence Type, $F(3, 87) = 83.336$, $p < .001$, and a main effect of Speaker Type, $F(1, 29) = 13.295$, $p = .001$, as well as a significant interaction between the two, $F(3, 87) = 4.107$, $p = .009$. False sentences were rated significantly lower than True, Felicitous, and Infelicitous sentences (all p's $< .001$). Infelicitous sentences were rated lower than True and Felicitous sentences (both p's $< .001$), which did not differ from one another ($p > .1$). In contrast to Accent Intolerant participants, Non-Native speech was rated higher than Native speech ($p = .001$). Post-hoc analyses showed that specifically, Infelicitous sentences (but not any other type) were rated significantly higher in the Non-Native Speaker condition as compared to the Native Speaker condition ($p = .002$). Thus, the Accent Tolerant group was selectively tolerant to pragmatic infelicities produced by Non-Native speakers.

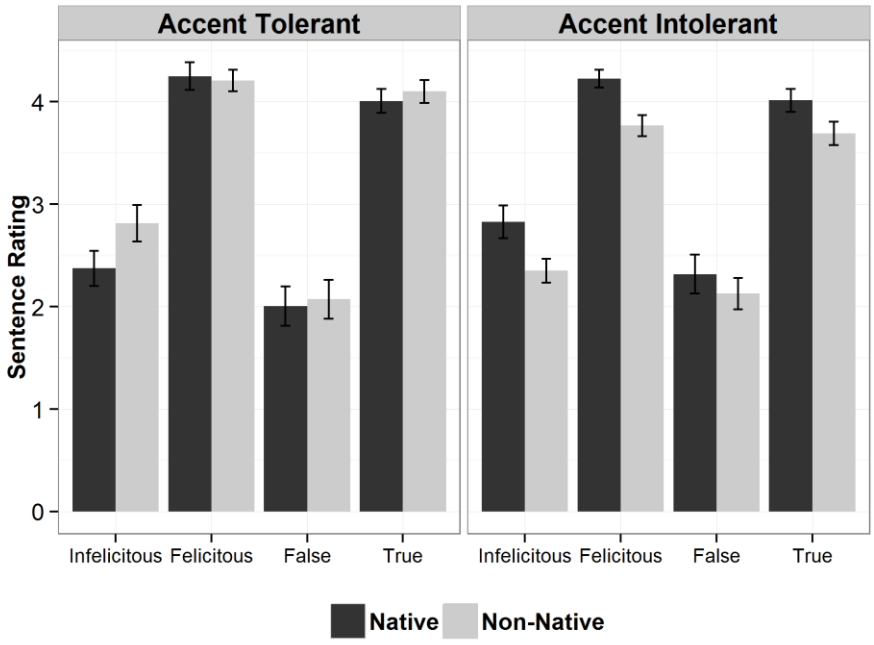

Figure 3: Results of Experiment 2, broken down by Accent Tolerant and Accent Intolerant participant groups. Error bars represent +/- 1 S.E.M.

To understand why the Accent Tolerant group showed a selective pragmatic tolerance but the Accent Intolerant group rated Non-Native speech lower overall, we turned to our measures of social-communicative ability and language processing ability. We performed two One-way ANOVAs with AQ-COMM score and Lexical Decision Task reaction times as the dependent variables, and Group (Accent Tolerant, Accent Intolerant) as the independent variable. The groups did not differ significantly in terms of AQ-COMM score, $F(1, 58) = .006$, $p = .938$, but the Accent Tolerant group had significantly faster Lexical Decision Task reaction times ($M = 730.8$, $SD = 147.7$) than the Accent Intolerant group ($M = 830.8$, $SD = 208.0$) , $F(1, 58) = 5.374$, $p = .024$. Thus, the Accent Intolerant participants with poorer language processing ability may have had more difficulty understanding the foreign-accented speech, therefore rating it lower overall. Figure 3 depicts the results of both groups.

In summary, some comprehenders have different expectations about use of nuanced pragmatic means of conveying information based on the group that the speaker belongs to (native vs. non-native, in this case). Pragmatically infelicitous sentences are deemed to be more acceptable when they are produced by a non-native speaker as compared to a native speaker. Critically, this pragmatic

tolerance is subject to individual differences in language processing ability, with less skilled comprehenders judging non-native speech as worse overall. Additionally, the increased tolerance towards non-native speakers does not appear to come online in the earliest moments of processing (at least in a case where there is visual context), but does affect untimed, offline ratings.

4 General Discussion

After being given additional context and/or a training period, listeners adjust to individual speakers when computing scalar implicatures and interpreting the meaning of quantifiers. In the present study, we asked whether listeners also adjust their pragmatic processing due to expectations about different *groups* of speakers (without an adaptation period). Specifically, we focused on interpretation of foreign-accented non-native speech. A hypothesis one could make is that because listeners adjust to characteristics of an individual speaker, they also adjust to characteristics of a group of speakers and are more likely to expect a non-native speaker to produce an under-informative utterance (and are therefore less likely to compute a SI). On the other hand, one might argue that listeners very strongly abide by the CP, and only shift their behavior when it is necessary, e.g., when they are given reason to believe that a single speaker will produce an under-informative utterance when describing a single context.

Across two experiments, we found that linguistically skilled listeners judge infelicitous sentences to be more acceptable when they are produced by a non-native speaker, but this adjustment did not appear for less linguistically skilled participants, or in an experimental paradigm that required speeded judgments and included visual context. It seems, then, that listeners can be sensitive to the language background of the speaker when engaging in pragmatic processing, but this sensitivity is highly subject to individual differences and task demands. The selective pragmatic tolerance we observed may be relatively weak because it requires overriding the maxims of the CP, which are a set of expectations we as communicators hold very strongly.

Because this is the first investigation into speaker identity effects at the group level for SI computation, there are still several alternative explanations of our findings left open. First, listeners may never use speaker group identity information in the earliest moments of SI computation, but may re-analyze a sentence and change their initial interpretation if necessary. Alternatively, listeners may be influenced by information about a speaker's language background early during sentence processing, but not in cases that include supporting visual context. Both of these potential explanations may be subject to individual differences in language processing ability or other listener characteristics. Furthermore, more research is needed in determine exactly what expectations listeners hold about non-native speakers (e.g., linguistic competence, world knowledge, cultural stereotypes), and how these expectations influence pragmatic inference.

References

Baron-Cohen, Simon, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. 2001. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* 31:5–17.

Bergen, Leon, and Daniel J. Grodner. 2012. Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38:1450–1460.

Bott, Lewis, and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51:437–457.

Grice, H. Paul. 1975. Logic and conversation. In *Syntax and Semantics*, ed. P. Cole and J.L. Morgan, 41–58. New York: Academic Press.

Hanulíková, Adriana, Petra M. Van Alphen, Merel M. Van Goch, and Andrea Weber. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience* 24:878–887.

Kuperberg, Gina R., and T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31:32–59.

Lev-Ari, Shiri, Marieke van Heugten, and Sharon Peperkamp. To appear. Relative difficulty of understanding foreign accents as a marker of proficiency. *Cognitive Science*.

Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44:314–324.

Mueller, Shane T., and Brian J. Piper. 2014. The psychology experiment building language (PEBL) and PEBL test battery. *Journal of Neuroscience Methods* 222:250–259.

Nieuwland, Mante S., Tali Ditman, and Gina R. Kuperberg. 2010. On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language* 63:324–346.

Yildirim, Ilker, Judith Degen, Michael K. Tanenhaus, and T. Florian Jaeger. 2015. Talker-specific adaptation in quantifier interpretation. *Journal of Memory and Language* 87:128–143.

Department of Psychological & Brain Sciences
University of Delaware
Newark, DE 19716
sfairchild@psych.udel.edu
apapafragou@psych.udel.edu