Women and men have different discourse biases for pronoun interpretation

Jennifer E. Arnold
Department of Psychology, University of North Carolina at Chapel Hill

Address all correspondence to: Jennifer Arnold, UNC Chapel Hill, Dept. of Psychology, Davie Hall #337B, CB #3270, Chapel Hill, NC 27599-3270, Email: jarnold@email.unc.edu
Women and men have different biases

Abstract

Two experiments examine how men and women interpret pronouns in discourse. Adults are known to show a strong “first-mention bias”: When two characters are mentioned (Michael played with William...), comprehenders tend to interpret subsequent pronouns as coreferential with the first of the two characters, and to find pronouns more natural than names for reference to the first character. However, this bias is not absolute. Experiment 1 demonstrates a stronger first-mention bias for women than men in their naturalness ratings for short stories. Experiment 2 monitors eye movements during story comprehension, and finds that women are more likely than men to consider the first-mentioned character as the pronoun referent. These findings reveal the first known gender difference in reference processing, and reinforce the view that reference processing is driven by more than the discourse context alone.
Women and men are clearly different, and it is popularly assumed that they use language differently. Yet much is still unknown about how the cognitive mechanisms underlying language use differ, particularly in the domain of discourse processing. This paper examines one of the core aspects of language: how we refer and understand references. I focus on third-person pronouns (he, she), which require people to draw inferences about their meaning based on the discourse and real-world context. We know that adults rapidly interpret pronouns, based on a variety of cues from the discourse context. The question examined in this paper is whether men and women differ in their biases about what makes a pronoun appropriate for a particular referent, and how these biases affect comprehension of pronominal references.

Pronouns are often ambiguous, even in context. Consider this story: Aurora had lunch with Fiona. She made a big mess. Both female characters in the story are potential referents for the pronoun she. This requires listeners and readers to draw on information from the context, plausibility, and other information to draw the most likely interpretation. The study of pronoun comprehension therefore provides a window onto the processes of understanding language in context. This study examines the hypothesis that individuals and groups differ in the degree to which they utilize one aspect of the discourse context (the first-mention bias), in two experiments. As a first step toward understanding why individual differences occur, I compare the discourse hypothesis with an alternate hypothesis that men and women simply differ on linguistic skill, or attention paid to the task. A model of pronoun interpretation is presented in the general discussion, and used to narrow the range of possible explanatory mechanisms.
**Contextual effects on pronoun interpretation**

There is strong evidence that comprehenders tend to assume that pronouns refer to information that is highly salient, or accessible in the context. For example, pronouns typically refer to information that has been recently mentioned, or is predictable (inter alia, Brennan, 1995; Gordon, Grosz, & Gliom, 1993; Jarvikivi, van Gompel, Hyona, & Bertram, 2005; Kehler & Rohde, 2013; Koornneef & van Berkum, 2006). When two characters have been mentioned, as in the *Aurora/Fiona* example, listeners often assume that the pronoun is coreferential with the first character (Aurora), which in English is usually also the grammatical subject. Similarly, speakers will tend to use a pronoun more for the first than second character (e.g., Brennan, Friedman, & Pollard, 1987; Chafe, 1976; Gernsbacher & Hargreaves, 1992; Gundel, Hedberg, and Zacharski, 1994; for a review see Arnold, 2010). This tendency is called the **first-mentioned bias**. Since first-mentioned/subject referents tend to be topical in a story (Givon, 1983), and are likely to be re-mentioned, findings like these support the idea that the first-mentioned/subject position makes the referent salient and accessible. This referential accessibility is argued to be the necessary condition for a pronoun to be an acceptable form of reference (Ariel, 1990, 2001, 2004; Chafe, 1994; Gundel et al., 1993).

But importantly, discourse biases like the first-mention bias are not absolute rules. For example, when a friend said to me *I took my daughter to the doctor. She said her arm was broken*, I needed to use my knowledge of real-world plausibility to infer that *She* referred to the doctor, even though the daughter was more salient in the conversation. Thus, the job of comprehenders is to find the interpretation that most plausibly matches the speaker or writer’s intended referent. Discourse biases provide a tool, but do not fully specify the correct interpretation.
Indeed, there is substantial evidence that the first-mentioned bias can be modulated by other factors. For example, the semantics of some verbs may lead comprehenders to expect reference to the nonsubject, and facilitate pronoun comprehension (e.g., *Katherine threw the ball to Alessandra, and she made a shot*; Garvey & Caramazza, 1974; Kehler, et al. 2008; McDonald & MacWhinney, 1995; Rohde, Kehler, & Elman, 2007; Stevenson, Crawley, & Kleinman, 1993; Koornneef & van Berkum, 2006). Listeners also attend to evidence about the intended referent, like the speaker’s eyegaze or pointing. For example, Nappa and Arnold (2014) presented participants with videos in which speakers told short stories like *Panda Bear is playing with Puppy. He wants the ball*, where both characters were the same gender (here, both male). Listeners tended to identify the referent of the ambiguous pronoun as the first-mentioned character when no cues were provided; but gazing at Puppy weakened this bias, and pointing overturned it (see also Goodrich Smith & Hudson Kam, 2012).

Thus, despite the systematicity of discourse biases, the mechanisms of pronoun comprehension cannot be reduced to a simple rule or algorithm. If psycholinguists want to understand the mechanisms of pronoun comprehension, we need to understand variation in how discourse biases are utilized. In many cases, the final interpretation of a pronoun is redundantly determined by multiple sources, including the accessibility of discourse referents and the most plausible interpretation of the discourse as a whole.

In other words, the mechanisms of interpretation may depend on the linguistic context to a greater or lesser degree, and still yield the same interpretation. This raises questions about whether the contribution of the linguistic context varies across situations or individuals. If it does, it suggests that pronoun comprehension is not simply a function of linguistic heuristics like “recent and first-mentioned referents are prioritized as pronoun referents”, but rather is the result
Women and men have different biases of how the listener uses discourse accessibility and other sources of information to efficiently arrive at the most plausible interpretation. In support of this, Nieuwland and van Berkum (2006) demonstrate that readers with higher working memory capacity are better able to maintain multiple interpretations of an ambiguous pronoun. Even the participant’s mood can have substantial effects: Van Berkum et al. (2013) found that participants in a bad mood were less able to use referential expectation during on-line language processing.

The current study examines individual variation in the utilization of the discourse context, specifically how men and women as groups respond differently to the first-mention/subjeckhood status. This variability is interesting because it helps identify the mechanisms of language use. In some sentences, like the examples above, the pronoun may be truly ambiguous. In other sentences, the pronoun may be disambiguated by the following input, for example: *Lucas rolled the ball to Peter, but he didn’t realize that Peter wasn’t looking*. In these cases, the discourse context still guides the listeners’ expectations about upcoming references, and affects the speed with which they identify the correct referent “online”, as the sentence unfolds in real time (Almann & Kamide, 1999; Kaiser & Trueswell, 2004; Kehler & Rohde, in press). Thus, individuals may vary in their representations of the conditions for what makes a pronoun an appropriate form of reference (as opposed to a repeated name, for example), or they may vary in how quickly or frequently they use this information on-line.

**Individual Variation in the first-mentioned bias**

Recent evidence suggests that the first-mentioned bias in pronoun comprehension varies across both individuals and situations. Arnold and Lao (under review) report two experiments that examined how pronoun comprehension is influenced by the processing that occurs as the
characters are introduced to the story. We used a variant of the eyetracking task developed by Arnold, Eisenband, Brown-Schmidt, & Trueswell (2000), in which participants heard a story about two characters, and were asked to decide if it matched the pictured scene (see Figure 1).

Example story:

Doggy picked apples with Birdy near the farmhouse. He was wearing a hat to protect himself from the sun.

The story always mentioned two characters of the same gender in the first sentence, and the second sentence began with an ambiguous pronoun, e.g. Doggy picked apples with Birdy near the farmhouse. He was wearing a hat to protect himself from the sun. The pictured scene determined the target (the referent of the pronoun) by manipulating the critical property, e.g. who was wearing a hat. The pronoun referred to the first-mentioned character (N1) half the time, and to the second-mentioned character (N2) half the time. We examined the participants’ fixations on each character immediately after the pronoun, which revealed their biases to consider that character as a potential referent.

As expected, participants’ eye movements revealed strong effects of the discourse context: immediately following the pronoun, they looked at N1 more than N2, and were more likely to say that the story and picture matched when the target was N1. The critical question was whether pronoun comprehension would also be influenced by the participant’s attention at the onset of the story. We used the listener’s eye movements during the first second of the story as a metric of the degree to which they happened to be attending to the target. In the first experiment,
we manipulated their trial-initial attention with a visual capture cue. In the second experiment the capture cue did not appear until later, and thus did not affect trial-initial attention. In both experiments, we found that if the participant attended to the target character at the onset of the story (even though there was nothing to identify it as the target yet, since the pronoun had not occurred), once the pronoun did occur participants were faster to look at the target. This effect was specific to the pronoun, and not a general viewing bias. This finding supports the idea that the accessibility of discourse referents is not uniquely determined by the linguistic and visual context. Instead, it is modulated by the participant’s attention to discourse stimuli.

This study also raised the possibility that there may be systematic differences between individuals in the way they use the discourse context. Although this study was not designed to look at participant gender effects, a posthoc exploration revealed that the N1 bias was stronger for the 43 female participants than it was for the 9 male participants. When the gender by condition interaction was added to the model for experiment 1, it was a nearly significant predictor of target looks (modeled as the empirical logit of the ratio of target-to-competitor looks: $\beta=0.51$, $t=1.92$, $p=.0551$). However, although a similar trend was seen in experiment 2, it was not significant ($\beta = .15$, $t=.49$, $p=.63$).

This finding was intriguing, since almost nothing is known about how men and women differ in the mechanisms that they use for discourse processing. Where there are known gender differences in language use, they make few predictions about this question. Women have been reported to use pronouns more than men, especially first-person pronouns ($I$, $we$; e.g., Newman, Groom, Handleman & Pennebaker, 2007), but the discourse properties of the referents were not measured. Men use acoustically reduced forms more than women (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Since pronoun use and acoustic reduction are both linguistic
Women and men have different biases

mechanisms for marking predictable, given information (Arnold, 2008), Bell et al.’s finding might predict that men would produce more pronouns than women, holding the discourse context equal. On the other hand, men also are more disfluent than women (Bell, Jurafsky, Fosler-Lussier, Girand, Gregory, and Gildea, 2003), and disfluency is associated with a drop in pronoun use (Arnold, Bennetto, & Diehl, 2009). However, none of these findings suggest differential sensitivity to the discourse context. Likewise, other known linguistic differences between women and men make few predictions about discourse processing, such as differences in sociolinguistic style (Eckert & McConnell-Ginet, 2003; Tannen, 2001), phonological processing (Irwin, Whalen, & Fowler, 2006; Majeres, 1999; Wei et al., 2012) or language localization in the brain (e.g., Jaeger et al., 1998).

Neither is it clear that a stronger N1 bias for women is related to “better” language processing. Although pronouns are used more frequently for N1 referents, they can also occur quite naturally with N2 referents. In these cases, a strong or inflexible N1 bias might lead to the incorrect interpretation. In addition, there is almost no real evidence that women are better at language use than men. A trip to the neighborhood playground may introduce you to parents who are convinced that verbal skills are typical of girls, but there is little evidence to support this. Although some studies report female superiority for some specific processes (e.g., Daltrozzo, Wioland, & Kotchoubey, 2007; Irwin et al., 2006; Majeres, 1999; Martin & Hoover, 1987; Ullman et al., 2002), reviews suggest that actual gender differences in language ability tend to be small (Cameron, 2007). Thus, differences in discourse processing are more likely to reflect a difference in the way information is used, not how well it is used.
Do women and men process pronouns differently?

The current study directly tests whether women and men have different biases during pronoun comprehension, focusing on the known tendency for first-mentioned (N1) referents to be perceived as more accessible than second-mentioned (N2) referents. Any such differences are expected to emerge as a small shift in the strength of the first-mention bias, holding other aspects of the discourse and task constant.

Experiment 1 uses a rating task to examine men and women’s judgments about the naturalness of stories that use pronouns, compared with stories that use names. The first-mentioned bias predicts that pronouns should be more natural for N1, while names should be more natural for N2. This experiment also examines stories with only a single character, which is predicted to be accessible and thus lead to a high pronoun preference. Variation in the N1-bias across individuals should emerge as variation in the differentiation between conditions in their ratings. Experiment 2 uses the auditory narrative comprehension task from Arnold and Lao, described above. If women have a stronger N1 bias than men, it should lead them to look at the N1 character more often than the N2 character in the moments immediately following the pronoun, as they search for the appropriate interpretation. The question in both experiments is whether women and men show different strengths in their preference for the first-mentioned character.

EXPERIMENT 1: PRONOUN PREFERENCES

Experiment 1 tested participants’ sensitivity to the naturalness of referential patterns, using a rating task. Participants read pairs of sentences, like those shown in Table 1, and rated the story for naturalness. Previous work predicts that the ratings should be influenced by the
discourse context: pronouns should be more natural than names when the referent was N1, or the only character in the sentence, compared to when the referent was N2. The critical question is whether this pattern differs for men and women.

Method

Participants

The analysis included 80 participants (40 female). An additional 12 women and 12 men were excluded. Of the women, 1 was a non-native speaker, 2 missed criterion on both the comprehension questions and fillers, 3 missed criterion on the fillers only, and 1 had no variation in responses. Of the men, 1 reported a cognitive disorder, 4 missed criterion on the comprehension questions, 2 had both no variation and missed criterion on the comprehension questions, and 2 missed criterion on the fillers only. 7 participants were excluded to even out the lists (5 F, 3 M), always choosing the last participant(s) run on a particular list. All participants received course credit for their participation.

There was no evidence that the women and men in our sample differed on cognitive ability. Although I did not have access to data on other cognitive measures for these subjects, they were drawn from the population of psychology 101 students at UNC Chapel Hill, in the fall of 2008. Overall the men and women in this course did not differ in their final course grades ($t(190) = 1.48, p=.14$; Dr. Elizabeth Jordan, p. c.).

Table 1. Example stimuli for Experiment 1.

<table>
<thead>
<tr>
<th>EXPERIMENTAL STIMULI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>One-character (n=20): Elijah fried eggs for breakfast.</td>
<td></td>
</tr>
<tr>
<td>{Elijah/He} wrapped them in a tortilla with salsa.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Two-character (n=40)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N1 referent: Greg criticized Mark during gym class.</td>
<td></td>
</tr>
<tr>
<td>{He/Greg} apologized to Mark after class.</td>
<td></td>
</tr>
</tbody>
</table>
Women and men have different biases

N2 referent: Greg criticized Mark during gym class. 
\{He/Mark\} held a grudge against Greg for days.

CRITICAL FILLERS
Ungrammatical: Tiffany interrogated Vince about the crime. 
He successfully convinced her that he were not responsible.

Awkward: Diana and Evan made s'mores over the fire. 
The marshmallows kept being caught on fire by Evan.

OTHER FILLERS
Conjoined NP; N1 ref James and Stephanie rode their bikes at the beach. 
James rode over a nail and got a flat tire.

Conjoined NP; N2 ref Carlos and Sally ate lunch at the new pizza place. 
Sally was on a diet, so she had a salad instead of pizza.

2-character event; N1 ref Brian carried Emma to the car because she hurt her ankle. 
He then went to get Emma an ice pack to reduce the swelling.

2-character event; N2 ref Jim floated over to Heather in the pool. 
She hoped that he would ask her out on a date.

Other Justin brewed tea for Sarah because she was sick. 
The tea was soothing to her sore throat.

Experiment Design and Stimuli

Stimuli consisted of two-sentence stories, as shown in Table 1. There were 60 experimental stimuli. Of these, 20 were stories about a single character, and 40 were stories about two characters. In the two-character stimuli, the first sentence mentioned two human characters with the same gender (half female)\(^1\), and names that were matched by length in syllables. In all experimental stimuli, the second sentence began with a reference to either the single character (one-character items) or one of the two characters (two-character items), using

\(^1\) Although both male and female pronouns were used in this task, this variable cannot be analyzed because it was not systematically manipulated, and was confounded with other item-specific properties.
either a pronoun or a name. In the two-character items, the initial reference was always followed by a name reference to the other character, which made it clear who was intended by the ambiguous pronoun. Thus, the one-character stimuli were manipulated on only one dimension (pronoun vs. name), while the two-character stimuli were manipulated in a 2x2 design (N1 vs. N2; pronoun vs. name). These items were organized into four lists (with forwards and backwards orders) in a Latin square design.

For the two-character stimuli, I also manipulated verb type as a control variable. The first sentence described an event in which the verb was expected to create a bias toward either the first or second NP, based on previous research (e.g., Arnold, 2001; Kehler et al., 2008). The N1-biased verbs included 8 Goal-Source verbs like grabbed, accepted; 9 Stimulus-Experiencer verbs like inspired, amazed; and 8 Joint Action predicates like went sledding with, played basketball with). The N2-biased verbs included 8 Source-Goal verbs like brought to, sent to, and 7 Experiencer-Stimulus verbs like criticized, trusted). However, this measure did not significantly affect ratings, nor did it interact with reference form or discourse status, so it will not be discussed further.

In the 40 fillers, two different-gender characters were mentioned in the first sentence, either as a conjoined subject (e.g., Ethan and Rachel danced at the concert,) or in subject and nonsubject positions, using a variety of event types (e.g., Brittany introduced Ian to an old friend.) The second sentence in the fillers began with a name, a gender-disambiguated pronoun referring to N1 or N2, or a word referring to something else.

Participants were asked to rate each two-sentence story for naturalness: “Please read the following short stories. Think about whether the story is phrased in a way that sounds natural or
not. We aren't interested in how likely the events are, just how naturally the story is told. Rate the sentences on a scale of 1 to 7 where 1 is highly unnatural and 7 is completely natural.”

Two measures assessed whether participants followed instructions: 1) 10 critical fillers had pragmatically awkward continuations, and 3 had minor grammatical errors, for example replacing *was* with *were*. See Table 1 for examples. Participants were excluded if their ratings for these “bad” fillers were more than .05 points higher (on average) than their ratings for the “good” fillers.  2) 12 items were followed by a comprehension question, where participants judged whether a statement was true or false about the story. For example, following the story about Elijah (see Table 1), the statement said “Elijah ate sausages for breakfast” (false). In all cases the information queried had occurred in sentence 1, and was therefore unrelated to the manipulation in sentence 2. Of these statements, 5 were false and 7 were true. Of the false statements, two were referential ("The one who brought the sandwich was Charlotte", when Alice had brought a sandwich to Charlotte), one was correct except a new name was substituted for one of the two characters, and two had other details incorrect. Of the true statements, four were referential, and three were event paraphrases. Participants were excluded if they did not answer at least 8 of the comprehension questions correctly.

*Procedure*

The experiment was conducted using the Psychology Pool online software, such that participants could complete the experiment on the web at a location of their choosing.

*Analysis*

Participants’ ratings were evaluated in a mixed-effects linear regression model, using SAS proc mixed, including random effects for both participants and items. The dependent
variable was the participant’s rating score for each item, adjusted as a z-score based on the participant’s overall mean and standard deviation. This controlled for variation in how each participant used the rating scale. The models also included random slopes for the critical discourse and reference form predictors. Supporting analyses were performed on participant means using analyses of variance. The critical predictors (pronoun vs. name; n1 vs. n2 referent; gender) were centered.

Results and Discussion

1. Filler and descriptive results

The first question is whether men and women performed similarly on the task overall. Evidence from multiple sources suggested that they did. First, t-tests show that there were no gender differences on comprehension question accuracy ($M$ for female = 10.45; $M$ for male = 10.35; $t(78) = -.38, p = .7$). Second, both groups used a similar range of raw ratings (average 5.68 for both groups; $t(39) = .85, p = .39$) and had a similar degree of variation (as measured by the standard deviation). Table 2 reports performance for men and women on the filler items. T-tests confirm that there was no difference in the average rating of the “good” filler items across genders ($t(78) = -0.37, p = 0.7$)

There was, however, one indication that men and women differed on a measure besides reference form preferences. Recall that some fillers were either grammatically or pragmatically odd. These items are grouped together in Table 2 as “bad” fillers. An analysis of these fillers alone shows that men gave marginally higher ratings on average to bad fillers than women ($t(78) = 1.79, p = .08$).
In sum, both groups of participants performed the task with a similar level of accuracy, and used the rating range in a similar way. The only trend towards gender differences in these control analyses was that men were somewhat less critical of the awkward or ungrammatical fillers than women.

Table 2. Overall performance in Experiment 1 (raw rating scores)

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average rating overall</td>
<td>5.68 (St. Dev = .76)</td>
<td>5.68 (St. Dev = .62)</td>
</tr>
<tr>
<td>Average filler rating</td>
<td>4.56 (St. Dev = .75)</td>
<td>4.66 (St. Dev = .81)</td>
</tr>
<tr>
<td>Range of ratings used for all items</td>
<td>1-7 (n=32)</td>
<td>1-7 (n=30)</td>
</tr>
<tr>
<td></td>
<td>1-6 (n=4)</td>
<td>1-6 (n=3)</td>
</tr>
<tr>
<td></td>
<td>other (n=4)</td>
<td>other (n=7)</td>
</tr>
<tr>
<td>Average rating of “good” fillers (27 items)</td>
<td>5.41 (St. Dev = 0.86)</td>
<td>5.33 (St. Dev = 0.95)</td>
</tr>
<tr>
<td>Average rating of “bad” fillers (13 items)</td>
<td>2.79 (St. Dev = 1.12)</td>
<td>3.26 (St. Dev = 1.24)</td>
</tr>
</tbody>
</table>

2. Critical results

The critical question was how men and women would respond to the pronoun and name continuations in different discourse contexts. The most general finding was that both men and women distinguished between pronoun and name continuations. For both groups, pronouns were preferred over names in the one-character conditions, where the single character was highly accessible, with no competitors (see Figure 2). In the two-character conditions, both groups demonstrated a slight preference for pronouns over names when the referent was N1, and a larger preference for names over pronouns when the referent was N2.

In addition, there was a significant gender difference: women’s ratings for pronoun and name conditions were more extreme than men’s. Two conditions received relatively low ratings: name references in the one-character condition, and pronoun references to N2 in the two-character condition. Although both men and women dispreferred these conditions, men’s ratings were not as low as women’s.
Women and men have different biases

Figure 2. Average adjusted (z-score) ratings for critical stimuli in Experiment 1.

To assess these biases, I examined the one-character and two-character items separately. For each, I built a statistical model in three steps. First, I tested the contribution of the control variables (list, list forward vs. backward, item order, sentence length); only those predictors that contributed at $t >= 1.5$ were retained. Second, I tested the contribution of discourse factors and their interactions, again only retaining those predictors that contributed at $t >= 1.5$. Finally, I added participant gender to the model.
Women and men have different biases

Table 3. Parameter estimates, t-statistics and p-values for critical and control predictors in the critical analyses for experiment 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>TWO-CHARACTER ITEMS</th>
<th>ONE-CHARACTER ITEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (error)</td>
<td>DF</td>
</tr>
<tr>
<td>CRITICAL PREDICTORS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition (N1 vs. N2)</td>
<td>0 (0.06)</td>
<td>37</td>
</tr>
<tr>
<td>Form (Pronoun vs. Name)</td>
<td>-0.26 (0.04)</td>
<td>36</td>
</tr>
<tr>
<td>Gender (Women vs. Men)</td>
<td>0.02 (0.04)</td>
<td>38</td>
</tr>
<tr>
<td>Condition x Form</td>
<td>0.76 (0.06)</td>
<td>2790</td>
</tr>
<tr>
<td>Condition x Gender</td>
<td>-0.08 (0.1)</td>
<td>2790</td>
</tr>
<tr>
<td>Form x Gender</td>
<td>-0.09 (0.07)</td>
<td>2790</td>
</tr>
<tr>
<td>Condition x Form x Gender</td>
<td>0.471 (0.11)</td>
<td>2790</td>
</tr>
<tr>
<td>CONTROL PREDICTOR</td>
<td>Item order</td>
<td>Na</td>
</tr>
</tbody>
</table>

The results of the final models are presented in Table 3. For two-character items, there was a critical three-way interaction between gender, reference form, and discourse condition, reflecting the fact that compared with men, women provided a greater preference for N1 pronouns over names, and a greater preference for N2 names over pronouns. There was also a main effect of condition, reflecting higher ratings for N1 than N2, and an interaction between condition and form, due to the general preference for N1 pronouns over names and for N2 names over pronouns. For one-character items, there was a main effect of form (pronouns preferred over names). There was also a critical interaction between form and gender, in that women revealed a greater dispreference for names and a greater preference for pronouns than men did.

In sum, women responded more strongly to the discourse appropriateness of names and pronouns than men did. In particular, men and women varied in their willingness to give a poor rating to items that were pragmatically odd. There were two conditions that appeared to be dispreferred: names in the one-character condition (e.g., Neil drank a cup of coffee. Neil had to
then make another pot for his co-workers), and N2 pronouns in the two-character condition (e.g., Andrew went shopping with Isaac at the new mall. He bought a shirt, but Andrew didn’t have enough money to buy anything.) In both cases, men’s ratings were slightly higher than women’s. Similarly, men gave somewhat higher ratings to the filler items that were pragmatically or grammatically infelicitous.

Discussion

The rating results demonstrated that men and women differed in how strongly they reacted to variation between pronouns and names. Women strongly dispreferred names for N1, and pronouns for N2, while men were relatively more likely to accept both. What explains this result?

Several explanations can be ruled out. One possibility is that, contrary to our hypothesis, women and men share the same biases, but translated them into rating scores differently. Yet men and women used the rating scale similarly, providing a similar range of ratings. Most importantly, our findings are the same for both z-scores (reported here) or raw ratings.

A related hypothesis is that women showed greater sensitivity to reference form because they were more careful in this task. Yet this is inconsistent with the fact that men and women performed equally well on the comprehension questions. Moreover, individual differences in accuracy did not consistently predict pronoun preferences across the genders. Figure 3 illustrates the average “pronoun bias” (pronoun – name ratings) for participants at each level of comprehension accuracy. For women, accuracy scores and pronoun bias were correlated positively in the “single” (R=.17) and N1 conditions (R=.23), and negatively in the N2 condition.
Women and men have different biases (R=-.47). For men, however, accuracy was not correlated with pronoun bias in any condition (ranging from R= -0.02 to R= -0.05).

Figure 3. Average of each subject’s “pronoun preference”, displayed by condition and comprehension question accuracy. The number of participants at each level of accuracy was 12, 7, 10, 11 (female) and 9, 14, 10, 7 (male).

Thus, even the most careful men were tolerant of N2 pronouns, suggesting that the gender difference reflects an actual difference in preferences, rather than a difference in care with the task.
Women and men have different biases

Figure 4. Average of each subject’s “pronoun preference”, displayed by condition and filler sensitivity (binned into four levels). The number of participants in each filler sensitivity bin was 7, 6, 18, 9 (female) and 14, 11, 6, 9 (male).

These findings raise a question about how general the observed gender difference is. Do men and women differ on pronoun preferences in particular, or on other aspects of linguistic appropriateness? The filler ratings provided one indication that women may be more sensitive to other variations in linguistic form as well, in that women gave slightly lower ratings to the “bad” fillers than men. I therefore tested whether there was a correlation between an individual’s pronoun bias in the experimental stimuli and a metric of “filler sensitivity” for each individual: The average rating for the “good” fillers minus the average rating for the “bad” fillers. An examination of Figure 4 suggests that “filler sensitivity” only correlated with reference form preferences in the N2 condition, where individuals with greater sensitivity also demonstrated a greater name bias. Moreover, the gender difference persisted in each condition.
Women and men have different biases

I tested the reliability of these patterns by adding filler sensitivity to the models of the adjusted (z-score) rating. The full model revealed a four-way interaction between filler sensitivity, reference form, condition, and gender (see Appendix A for details). To better understand this, I examined each condition separately. The critical finding was that in the single and N1 conditions, there was an interaction between gender and reference form, again showing that women had a stronger pronoun bias than men, even when filler sensitivity was controlled. There were no effects of filler sensitivity itself, or interactions between it and gender or reference form. In the N2 condition, there was a 3-way interaction among filler sensitivity, gender and reference form. This reflected the fact that the correlation between filler sensitivity and pronoun bias was stronger for women (R = -62) than men (R = -54). Notably, the most sensitive women displayed a stronger name bias for N2 references than the most sensitive men. This post-hoc analysis suggests that there is indeed a relationship between overall sensitivity to linguistic form and preference for reference form. However, this did not completely account for the gender differences in the stimuli ratings.

The results from experiment 1 thus tentatively suggest that there are gender differences in how participants react to the discourse appropriateness of pronouns and names, over and above any more general gender differences in sensitivity to linguistic form variation. The next question is: what aspect of language processing does this finding reflect? The rating task taps participants’ feelings of “naturalness”, which is likely to result from at least two sources: 1) the participant’s representation of the conditions for appropriate use of a pronoun, and 2) the degree to which on-line processing was easy. These two processes are likely related, in that pragmatic conditions are also used during on-line processing.
Women and men have different biases

On the other hand, it is possible that the gender differences observed here resulted from the strategic application of rules like “use a pronoun for the first character”, or other metalinguistic processes. Experiment 2 further examined differences in how men and women react to pronouns, using an on-line task of narrative comprehension. If the rating differences were purely the result of post-hoc processing and strategic judgments, we should not see gender differences in Experiment 2.

EXPERIMENT 2: ON-LINE PRONOUN COMPREHENSION

Experiment 2 investigated women and men’s biases during on-line pronoun comprehension, using a narrative visual world paradigm (Arnold, et al., 2000; Arnold & Lao, under review). Participants’ eye movements were monitored as they heard a story and decided if it matched a picture. The stories were about four characters: Doggy (male), Birdy (male), Bunny (female), and Kitty (female). In critical items, the first sentence mentioned two characters of the same gender, e.g. Birdy picked apples with Doggy near the farmhouse. The second (target) sentence in the story was always linguistically ambiguous, e.g. he was wearing a hat to protect himself from the sun. Both characters were consistent with the verb (i.e., both were wearing something), but only one matched the critical characteristic (wearing a hat). Therefore, participants’ direction of gaze immediately following the pronoun revealed their initial preferences to consider each character as the referent. Final responses indicated their willingness to accept the pronoun with the intended referent.

Method
Participants

Fifty-one undergraduates participated for course credit. Two were excluded because of track loss, one for poor accuracy on critical filler trials, one because she participated in Experiment 1, and one who was familiar with my research. The analysis included 23 women and 23 men, none of whom participated in Experiment 1.

Participant Matching Tasks

Two measures ensured that the samples of males and females were similar in terms of participants’ general cognitive abilities:

1) Participants were asked to report on a questionnaire their SAT scores, since SAT scores correlate with other measures of working memory (Engle, Kane, & Tuholski, 1999). There was no difference between men and women’s report of their total SAT score for the 44 participants who reported it (F(1,42)=.055, p=.82), or on Verbal SAT score for the 42 participants who reported it (F(1,40) = 1.18, p=.28).

2) Working memory was measured with the kiosk-ready version of the operation span task (Unsworth, Heitz, Schrock, & Engle, 2005). Participants read math problems out loud (e.g., \((1*2) + 1 = ?\)) and mentally solved them, while simultaneously remembering a letter (e.g., L). This program provides two working memory scores: the Absolute score (the sum of all perfectly recalled sets) and the Total score (the total of all correctly remembered letters). Men had slightly higher working memory scores on both measurements (Absolute: \(M=51\); Total: \(M=64\)) than women (Absolute: \(M=45\); Total: \(M=61\)). However, analyses of variance revealed that only the

\footnote{A standardized aptitude test required for application to college in the U.S.}
Absolute score approached significance (F(1,44) = 3.2, p = .079), and the Total score difference was not significant (F=2.1, p = .16).

**Experiment Design and Stimuli**

*Eyetracking Task.* Participants’ eye movements were monitored while they viewed a picture and heard a story (Figure 2). Using a paradigm adapted from Arnold, et al. (2000), each story introduced 2 same-gender characters in the first sentence, and referred to one with a pronoun in the second sentence. The visual stimulus manipulated whether the pronoun referred to the first- or second-mentioned (N1 vs. N2); e.g. in Figure 2 only one character has a hat.

There were 20 experimental items in two conditions (N1 vs. N2 target). Eight of the verbs were N1-bias (joint action); twelve were N2-bias (source-goal). These were pseudorandomly combined with 8 fillers in two lists, with a forwards and backwards version. Right/left location of target (i.e., the referent of the pronoun) and N1 were counterbalanced across items.

Participants judged whether the story matched the picture, responding by clicking on a button on screen after the story was over. All experimental stimuli matched the picture under the intended referential assignment. Of the 8 fillers, 6 were considered “critical fillers”. These had stories that mismatched the picture for non-pronoun reasons (e.g., *Doggy went boating one summer*..., when the picture showed Doggy snorkeling). These critical fillers provided a measure of each participant’s accuracy and attention to the task; participants with low accuracy were excluded from analysis. There were two practice items: one that matched and one that did not. The practice and filler items had 1-3 characters, different structures and no ambiguous pronouns.
Apparatus and Procedure

An Eyelink II monitored participants’ eye movements; the visual and auditory stimuli were presented on a PC running the ExBuilder software (Longhurst, 2006). Fixations were sampled every 4 msec and later collapsed into 16-msec samples (McMurray, 2002). Participants then completed a verbal questionnaire about the study, and the operation span task.

Eye gaze analysis procedure.

Eye movement data are presented in terms of looks, where a look is defined as a fixation grouped together with the prior saccade. Trials were excluded if there was more than 33% track loss during the critical region (1.7% of the data). I am particularly interested in the proportion of time spent looking at the target and competitor characters immediately following the pronoun (200 msec following pronoun onset until 300 msec following the disambiguation point). I therefore calculated an empirical logit of the proportion target looks (following Barr, 2009), out of all target and competitor looks: \( \log_e \left( \frac{\text{# samples target looks} + .5}{\text{# samples competitor looks} + .5} \right) \).

This target-elogit is used in the analyses reported here, and it can be thought of as an indicator of the “target bias”. This metric is ideal for tracking the discourse manipulation. When the target is N1 (based on the visual context), the target-elogit should be high (i.e., more target than competitor looks). When the target is N2, the target-elogit should be low.

Analysis procedure and control variables. The effects of experimental conditions were evaluated in a mixed-effects linear regression model, using SAS proc mixed, including random intercepts for both participants and items, as well as a random slope for the condition predictor (n1 vs. n2) with respect to both participants and items, and a random slope for participant gender by item (Barr, Levy, Scheepers, & Tily, 2013). For each analysis reported, I first built a control model to
Women and men have different biases

assess the effects of potential control variables (item order; verbtype; List, forward vs. backward order; target location (left/right); whether the participant was fixating the target at the onset of the pronoun (yes/no); whether the participant was fixating the competitor at the onset of the pronoun (yes/no). Following Arnold and Lao (under review), I also included Trial-initial attention as a control predictor. Any contributions that were significant at a level of $t=1.5$ or greater were retained for the final analysis. The final analysis added the predictors of interest: referent (first vs. second mention) and gender, which were also centered. This permitted me to assess the effect of participant gender against the backdrop of other well-understood predictors.

Response analysis procedure. The task that participants performed was to decide if the story matched the picture, which provided some information about listeners’ final interpretation of the story. For Mismatch responses, they provided a verbal explanation for why the story and picture didn’t match. I only counted Mismatch responses that were related to the pronoun (e.g., *Doggy was wearing the hat, not Birdy*). Mismatch responses that were irrelevant to pronoun resolution (e.g., *It’s not sunny in the picture*) were counted as Matches.

Because match responses were binary, I analyzed them in a multilevel logistic regression, using SAS proc glimmix with a Penalized Quasi-Likelihood (PQL) estimator, with a binary distribution and a logit link. The same procedure for including control variables was used as for the eye movement analyses.

Results and Discussion

Eyetracking Task. Figure 5 illustrates the proportion looks over time to both target and competitor characters, immediately following the pronoun. First, there was strong evidence for an N1 advantage for both women and men. There were more and earlier looks to the target when
it was N1 than N2; similarly, there were more and earlier looks to the competitor when it was N1 than N2. Second, and critically, the N1 advantage was stronger for women than men.
Women and men have different biases

Figure 5. Looks to target and competitor characters. The onset of the pronoun is at 0. The lines mark the average analysis region (200 after pronoun onset until 300 after the disambiguation point). The black lines correspond to the trials on which N1 was the target, and the red lines to trials on which N2 was the target. The solid lines represent looks to the target (pronoun referent).

These observations were supported by a model of the empirical logit of target looks during the ambiguous region following the pronoun. Condition means are shown in Table 4, and statistical results in Table 5. As Table 4 shows, there was a greater proportion of target looks when the target was N1, and a greater proportion of competitor looks when the target was N2 (and the competitor was N1). This led to a positive target e-logit metric (i.e., target bias) for the
Women and men have different biases

n1 target condition, and a negative target e-logit for the n2 condition. In addition, this contrast was greater for women than men. As shown in Table 5, there was a significant effect of discourse condition, and a significant interaction between condition and gender. This reflects the fact that women had a stronger preference to look at N1 than men did. Thus, when N1 was the target, women looked at the target more than men, but when N2 was the target, men looked at the target more than women did.

Table 4. Average target looks, competitor looks, and the empirical logit of target looks (i.e., target bias out of all target and competitor looks in the ambiguous region).

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N1 target condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>target looks</td>
<td>0.59</td>
<td>0.46</td>
</tr>
<tr>
<td>competitor looks</td>
<td>0.24</td>
<td>0.31</td>
</tr>
<tr>
<td>TARGET ELOGIT</td>
<td>1.32</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>N2 target condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>target looks</td>
<td>0.26</td>
<td>0.31</td>
</tr>
<tr>
<td>competitor looks</td>
<td>0.49</td>
<td>0.46</td>
</tr>
<tr>
<td>TARGET ELOGIT</td>
<td>-1.17</td>
<td>-0.76</td>
</tr>
</tbody>
</table>

Table 5. Parameter estimates, t-statistics and p-values for critical and control predictors in the eyetracking and response analyses for experiment 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>EYE MOVEMENTS</th>
<th>RESPONSES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (error)</td>
<td>DF</td>
</tr>
<tr>
<td><strong>CRITICAL PREDICTORS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition: N1 vs. N2 pronoun referent</td>
<td>1.893(0.335)</td>
<td>1,895</td>
</tr>
<tr>
<td>Gender (Women vs. men)</td>
<td>0.043(0.182)</td>
<td>1,895</td>
</tr>
<tr>
<td>Condition x Gender</td>
<td>0.99(0.406)</td>
<td>1,895</td>
</tr>
</tbody>
</table>
The gender difference occurred on top of the effects of several control predictors. There was a greater target bias when the competitor had not been fixated at pronoun onset than when it had. There was also a greater target bias overall in the source-goal items than the joint action items; however additional explorations revealed that this did not interact with condition, and was therefore most likely an effect of the target salience in those items. There was also an increase in the target bias when the target had been attended at trial onset; this effect is discussed further below.

**Responses.** Both men and women provided Match responses to nearly all N1 target items ($M = 97\%$), and less often to N2 target items. Men were more likely to accept the N2 target (24\%) than women (15\%), but this numerical difference did not reach significance in the statistical model. As shown in Table 5, responses were heavily predicted by the discourse condition, where N1 targets elicited Match responses much more often than N2 targets. No other predictor or interaction contributed significantly. This finding confirms the well-known N1 bias. In this task, where participants were asked to find “errors” (i.e. cases where the story mismatched the text), the N1 bias emerged as a rejection of the text, and probably a higher rejection of N2 pronouns than in natural contexts.
Control Analysis. The primary finding of this experiment was that women showed a greater tendency to look at the first-mentioned character after hearing a pronoun. Why did this occur? The hypothesis of interest is that women and men respond differently to the discourse when processing pronouns. But before we can conclude that they do, we must consider an alternate hypotheses: the pronoun-processing difference was just a symptom of a more general tendency for women in our sample to either process the stories more thoroughly overall (not specifically in the pronoun region), or to move their eyes more in response to the stories. This might occur if the women in our sample had a higher cognitive ability, or a greater engagement with the task. If so, we would expect to see gender differences occur on other measures.

1. Tests of general cognitive and verbal ability. As reported above, there was no evidence that the women in our sample had higher verbal or cognitive ability than men. There was no difference on reported SAT scores. While there was a slight difference in working memory scores, it did not suggest that the women in our sample were more cognitively able – men scored slightly higher than women.

2. Task engagement. Participant’s broad accuracy with the task was measured by analyzing the “critical no” fillers, on which the story and picture clearly did not match, for a reason not related to pronoun resolution. Analyses of variance revealed that women and men did not differ in their ability to correctly reject these “critical no” fillers (female $M = 91\%$; male $M = 88\%$; $F(1,44) = 0.33$, $p = 0.57$; $F(1,5) = 2.14$, $p = .20$).

3. Metalinguistic awareness. There was also no evidence that metalinguistic awareness differed across groups. The postexperiment questionnaire probed their awareness of the manipulation by asking increasingly leading questions. On the first two general questions (What did you think the task was about? and Was there anything difficult about the task?), 17% of women and 25% of
Women and men have different biases

men spontaneously mentioned pronouns. This difference was not significant in a model of whether the participant mentioned pronouns spontaneously (yes or no), using SAS proc glimmix, with subjects as a random effect, and only gender as a predictor (F=0.48, p = .49).

4. Viewing patterns. Another possibility is that the differences observed between men and women had to do with differences in viewing patterns, and not with pronoun processing per se. For example, perhaps men rely more on their mental situation models than women, on average, and thus do not look at discourse-related images during the task. To examine this hypothesis, I analyzed gender differences on several other eyetracking measures, using the same procedure as outlined above for the primary analysis.

Looks to animate characters. I analyzed the total number of looks to both target and competitor, using the empirical logit of character looks out of all looks during the ambiguous region (\( \log_e ((\text{total target and competitor samples} + 0.5)/(1 – \text{total target and competitor samples} + 0.5)) \)). Total character looks were not different across groups (female \( M = 75\% \), male \( M = 78\% \); \( t = -1.46; p = .15 \)), or conditions (\( t = 0.08; p = .94 \)), and there was no interaction between gender and condition (\( t = -0.09, p = .92 \)). The only significant predictor was verbtype: there were more character looks overall in the source-goal (83\%) than joint action (71\%) items.

Total number of saccades. Do women simply move their eyes around more than men? Not in this task: the average number of total saccades per trial was 27.1 for both men and women; it did not differ by condition (\( t = -1.26, p = .21 \)), participant gender (\( t = -.11, p = .92 \)), or the interaction between the two (\( t = 0.054, p = .59 \)).

Early attention to characters. Following Arnold and Lao (under review), I measured participants’ attention in the first second of the trial, and used the proportion of target looks out of all character looks as a metric of the degree of Trial-initial attention to the target. Trial-initial
attention did correlate with an increase in the ratio of target looks (see Table 5), but it did not interact with gender: when the interaction term was added to the model, it was not significant (t=1.01, p = 0.31).

The importance of trial-initial attention raises the question of whether women and men allocated attention differently at the start of the trial. If so, this might modulate the accessibility of referents in the discourse representation. To analyze this, I used the Trial-initial Attention metric as a dependent measure, to ask what drives looks in the first second of the task. Arnold and Lao’s results predict that Trial-initial Attention should be driven by physical location (more looks to the left character), and discourse condition (more looks to the first-mentioned character), since the first name occurred 400 msec after the picture was displayed. I built a model including these predictors, participant gender, and interactions between gender and each predictor. As expected, there were more initial looks to both the left than the right character (t = 9.8, p < .001), and to N1 than N2 characters (t = 2.24, p = .025). However, there were no differences between women and men, nor were there any interactions between gender and either condition or target location (all t’s < 1.5, all p’s > 1.5).

5. Relationship between eyeegaze and response measures. For both men and women, the participant’s on-line viewing patterns are an excellent predictor of their offline choices. That is, looks to the target during the ambiguous region of the pronoun predict the Match/Mismatch response. To a certain extent, this measure is a duplication of the condition measure: when the target is N1, people look at the target more, and also tend to respond Match. Nevertheless, as Figure 6 shows, there is a gradient effect of online viewing time when we examine the N2 condition by itself. (By contrast, the N1 condition yields almost 100% Match responses, and there is no effect of online viewing time). Both men and women were more likely to give a
Women and men have different biases

Match response in the N2 condition when they had fixated the target during the ambiguous region. When target e-logit was added to the model of Match responses, it was a highly significant predictor ($t = -2.69, p < .001$), as was condition ($t = -10.10, p < .0001$). However, despite the numerical trend for men to give more N2-match responses, there was still no effect of participant gender, nor any interactions between it and target looks.

These analyses show that both men and women display a similar relationship between their on-line looking behavior, as they hear the pronoun, and their final decision about whether the story and picture match. However, even though groups differed in on-line behavior, our off-line measure did not elicit a lot of variation, and thus was not sensitive enough to identify group differences.

Figure 6. Proportion of Match responses for items in the N2 condition, as a function of target looks during the ambiguous region. Negative bins (< -4.5; and between -1.5 and -4.5) correspond to a greater proportion of competitor looks, positive bins (> 4.5 and between 1.5 and 4.5) correspond to a greater proportion of target looks, out of all character looks.
6. The effect of individual working memory differences.

Our samples of men and women were well matched in most measures, but the men in our sample had a slightly higher working memory capacity, as measured by the Absolute score in the operation span measure. I used this metric to test whether our findings are better explained by working memory differences than male/female differences.

One hypothesis\(^3\) is that individuals with higher working memory capacity should be better able to maintain alternatives during comprehension (Nieuwland & van Berkum, 2006). If so, we should expect an individual’s absolute operation-span score (OSPA) to predict responses: higher-span subjects should accept both N1 and N2 targets, whereas lower-span subjects should either show a general N1 bias, or accept neither. However, this prediction was not supported. Instead, OSPAN score interacted with gender. Since participants provided near-ceiling Match responses for N1 targets, I examined this question by analyzing Match responses for N2 targets, as a function of gender, OSPAN, and the gender x OSPAN interaction. I found that the interaction was significant (β = .09 (.05), t=2.07, p = .039), such that for women, higher OSPAN scores led to fewer Match responses, while for men, higher OSPAN scores led to more Match responses. This analysis thus highlights the contrast between men and women, as opposed to accounting for it.

Another way to test the effects of working memory is to select a sample for which the working memory scores are matched. In our current sample, there was only one man with an OSPAN lower than 38, while there were six women. An analysis without these seven participants revealed the same pattern of significance for eye movements (except the control predictors of trial-initial attention and verbtype were only marginal). Critically, the interaction

---

\(^3\) Thank you to an anonymous reviewer for pointing this out.
between condition and gender was still significant ($\beta = -0.71 \ (0.24), t= -2.94, p = .003$). With this subset of participants, the responses analysis again revealed a main effect of condition, and this time the interaction between gender and condition was marginally significant ($\beta = -1.75 \ (1.02), t= -1.73, p = .084$). Thus, the observations above cannot be explained purely in terms of individual working memory differences.

**GENERAL DISCUSSION**

In an on-line pronoun comprehension task (Experiment 2), women exhibited a greater preference for pronouns to refer to first-mentioned characters, and a greater dispreference for pronouns to refer to second-mentioned characters. This contrast echoed the results from a naturalness rating study (Experiment 1), in which women dispreferred pronouns with second-mentioned referents more than men. These findings suggest that while both men and women find N1 pronouns easier to understand than N2 pronouns, this bias is stronger for women. A series of control analyses suggests that these are real effects, in that they cannot be easily attributed to general cognitive or strategy-related differences across individuals.

This finding constitutes the first evidence that adult women and men process pronominal references differently in context. As such, it adds to the growing literature on gender differences in language processing. Despite the popular assumption that women and men use language differently, few language processing differences have been reported, especially for adults. In addition, most reported gender differences are relatively small (Cameron, 2007). Moreover, those behavioral processing differences that have been observed typically come from tasks where participants responded to words and sentences out of context. The current study goes beyond existing findings by demonstrating gender differences in discourse processing.
What mechanisms differ across individuals?

Our observation that women have a greater N1 bias than men raises a critical question: what is the mechanism underlying this effect? Discourse processing notably draws on multiple mechanisms. Before speculating about where these differences lie, I will sketch out a general model of pronoun comprehension.

A model of pronoun comprehension. Recent research suggests that understanding pronominal references involves the following processes and representations:

1) A discourse representation. All language processing requires participants to maintain a representation of discourse events and characters, commonly called a discourse model or situation model (Johnson-Laird, 1983; van Dijk & Kintsch, 1983; van Berkum et al., 2007; Zwaan & Radvansky, 1998). Some information in the model is relatively more salient and accessible, possibly due to attentional processes. This may be represented in terms of activation, where accessible entities have higher activation (Ariel, 2001, 2004; Arnold & Griffin, 2007).

The accessibility representation is influenced by the linguistic context, as described above, such that recently mentioned and prominent information (like first-mentioned entities) are accessible. This representation could be seen as the comprehender’s estimate of what is accessible to the speaker/writer, and functionally, what is likely to be most important to the ongoing story. Yet the text is an imperfect representation of the speaker’s intentions, so this representation is also guided by other things, such as visual salience, and whether the comprehender chooses to attend to N1 or N2 and process it more deeply (Arnold & Lao, under review).
Women and men have different biases

One possibility is that women utilize the linguistic context more strongly than men do for the purpose of establishing the accessibility of referents in their mental model, while men may use non-linguistic information more heavily. Alternatively, men may use the discourse context in a more coarse-grained way, for example considering all recently mentioned characters to be fairly accessible. The N2 condition is interesting, because it presents a case of ambiguous accessibility. The referent has been mentioned recently, which does make it accessible -- and in fact speakers often do use pronouns to refer to N2 referents, even when the pronoun is ambiguous. Yet it is still somewhat less accessible than N1. Thus, the current results are also consistent with the hypothesis that women may specifically pay more attention to either structural information or order of mention than men. This is potentially consistent with evidence that men and women process syntactic information differently (e.g., Osterhout, Bersick and Mobey, 1997). Another possibility is that women may have a richer or more robust discourse memory for the discourse context, consistent with the claim that females have an advantage in episodic memory over males (Block, Hancock, & Zakay, 2000; Hartshorne & Ullman, 2006; Ullman et al., 2002). Further work is needed to distinguish these possibilities.

2) Knowledge that pronouns are pragmatically specialized. Language users must represent the knowledge that referential forms are used appropriately in different contexts. Pronouns are specialized for highly accessible referents, while names and definite descriptions are more appropriate for less accessible, discourse-new, or contrastive information (see Arnold, 2010, for a review).

The findings reported here demonstrate that both men and women share knowledge that pronouns are specialized for accessible referents, to some extent. For example, in Experiment 1, the single-character items provide a situation where the character’s accessibility is strong and
unambiguous. Both groups showed a preference for pronoun stories in this condition. In Experiment 2, both groups exhibited the same N1 bias, even thought it was stronger for women. This means that the gender difference is unlikely to come from a wholesale difference in the underlying representation of pragmatic specialization. If anything, it may stem from a difference in the ability to access this knowledge during their search for a referent.

3) **Search for a referent.** When the pronoun is encountered, comprehenders instantiate a search for the most likely referent. This may involve the activation of multiple referents in parallel, depending on constraints such as the morphological properties of the pronoun (e.g., *she* refers to a female, singular, animate entity; Arnold et al., 2000), and the preference for accessible over less-accessible referents. This activation may even begin before the pronoun is encountered, as comprehenders anticipate reference and the most likely entity to be mentioned (Arnold, 1998, 2001; Kaiser & Trueswell, 2004; van Berkum et al., 2013).

Simultaneously, these potential referents are compared against their fit to the rest of the unfolding sentence, and the most plausible meaning of the story. E.g., in *Zach photographed Daniel and he smiled*, it is likely that Daniel is the preferred referent of *he* (Hobbs, 1979). When a visual context is available, it may also guide these inferences. These inferences reflect the listener’s primary goal of identifying the speaker’s intended meaning. As Hobbs (1979) proposed, this is heavily guided by inferences that make the discourse coherent (see also Kehler & Rohde, 2013), but they also include other cues to the speaker’s meaning, like eyegaze or pointing (Nappa & Arnold, 2014).

There are two areas in which the referential search may differ amongst individuals. One possibility is that women access accessibility information more quickly or easily than men when they encounter a pronoun. However, this seems unlikely, since there was no evidence that our
female participants had an advantage in working memory or processing. Another possibility is that women weight accessibility constraints more strongly in comparison with morphological and inferential considerations, compared with men. In experiment 1, pronoun items were disambiguated by the use of a name for the other character, after the pronoun. For example, consider the example *Thomas impressed Leo during the performance. He waited backstage to get Thomas's autograph*. A reader who relies on accessibility may immediately assigned the pronoun to N1 (Thomas) while reading, which would later result in confusion and a low rating. By contrast, a reader who relies more on inferential processing may maintain both potential interpretations until the following context is available. If men tended to use the latter processing strategy, they would have experienced less confusion from N2 pronouns. In experiment 2, participants who prioritized accessibility considerations would only have to look at N1 following the pronoun, to check that the image matched the linguistic input. By contrast, participants who utilized inferences from other sources of information would need to search the display, and look at N2 to see if it matched the input.

Thus, women may have weighted accessibility information more strongly, while men may have depended more strongly on other aspects of contextual processing, like whether the story made sense. A variant on this conclusion is that women are more generally sensitive to variations in linguistic form – that is, “how it’s said”, and not just “what is said”. This is consistent with our finding in Experiment 1 that women distinguished between good and bad fillers more than men, as well as between stories with pronouns and names. Although individual “filler sensitivity” did not account for our pronoun comprehension findings, further research is needed to understand the extent of gender differences in different areas of language comprehension.
In summary, this paper reports evidence that men and women differ in the way they use the discourse context to understand pronouns, and I have proposed several mechanisms that may account for this difference. While additional research is needed, I have argued against explanations based on verbal ability or working memory.

The current findings are relevant for models of reference comprehension and discourse representation. Two experiments demonstrated that there is substantial individual variation in the strength of the first-mentioned bias. While these biases may correlate with the participant’s gender, the deeper question is why such individual variation exists. Thus, models of pronoun comprehension need to go beyond simple rules or heuristics (see also Kehler & Rohde, 2013). Instead, an adequate model of pronoun comprehension must specify how individuals integrate various sources of accessibility with inferences about the speaker’s intentions, and the range of variation in these processes.
AUTHOR NOTE

I am grateful to Kellen Carpenter, Shin-Yi Lao, and Cathleen Sparks for their help designing stimuli and administering the tasks. Thanks to Sarah Brown-Schmidt for her help creating the visual stimuli for Experiment 2. I am also thankful for Chris Wiesen’s patient help with the statistical models. This research partially funded by NSF grant BCS-0745627.

REFERENCES


use of gender information: evidence of the time-course of pronoun resolution from eyetracking.

Arnold, J.E. & Lao, S. C. (Under review). Effects of non-shared attention on pronoun
comprehension. Ms., University of North Carolina at Chapel Hill.

Arnold, J.E., Bennetto, L., & Diehl, J. J. (2009). Reference Production in Young
Speakers with and without Autism: Effects of Discourse Status and Processing Constraints.
*Cognition, 110*, 131-146.

Barr, D. J. (2009). Analyzing 'visual world' eyetracking data using multilevel logistic

confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3),
255-278.

effects on durations of content and function words in conversational English. *Journal of Memory
and Language, 60*, 92-111.

Effects of disfluencies, predictability, and utterance position on word form variation in English


Women and men have different biases


doi:10.1177/0272431687071007
Women and men have different biases


Nappa, R., & Arnold, J. E. (2014). The road to understanding is paved with the speaker’s intentions: Cues to the speaker’s attention and intentions affect pronoun comprehension. *Cognitive Psychology*.


Women and men have different biases


APPENDIX A

The following tables report the results of a post-hoc analysis that examined the relationship between individual differences on their sensitivity to reference form variation, and “filler sensitivity” (represented by the difference between the average raw ratings for good and bad filler ratings).

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate (Error)</th>
<th>DF</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Women vs. Men)</td>
<td>-0.113(0.082)</td>
<td>1.38</td>
<td>-1.38</td>
<td>0.177</td>
<td></td>
</tr>
<tr>
<td>Condition (N1 vs. N2)</td>
<td>0.038(0.095)</td>
<td>1.37</td>
<td>0.4</td>
<td>0.6908</td>
<td></td>
</tr>
<tr>
<td>Form (Pronoun vs. Name)</td>
<td>0.063(0.069)</td>
<td>1.36</td>
<td>0.91</td>
<td>0.3668</td>
<td></td>
</tr>
<tr>
<td>Filler sensitivity</td>
<td>0.067(0.016)</td>
<td>1.38</td>
<td>4.06</td>
<td>0.0002 ***</td>
<td></td>
</tr>
<tr>
<td>Gender x Condition</td>
<td>0.071(0.181)</td>
<td>1.2749</td>
<td>0.39</td>
<td>0.6946</td>
<td></td>
</tr>
<tr>
<td>Gender x Form</td>
<td>0.16(0.134)</td>
<td>1.2749</td>
<td>1.2</td>
<td>0.2317</td>
<td></td>
</tr>
<tr>
<td>Gender x Filler sensitivity</td>
<td>0.042(0.03)</td>
<td>1.2749</td>
<td>1.4</td>
<td>0.1616</td>
<td></td>
</tr>
<tr>
<td>Condition x pronoun</td>
<td>-0.044(0.117)</td>
<td>1.2749</td>
<td>-0.38</td>
<td>0.7059</td>
<td></td>
</tr>
<tr>
<td>Condition x Filler sensitivity</td>
<td>-0.012(0.034)</td>
<td>1.2749</td>
<td>-0.35</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Form x Filler sensitivity</td>
<td>-0.132(0.025)</td>
<td>1.2749</td>
<td>-5.35</td>
<td>&lt;.0001 ***</td>
<td></td>
</tr>
<tr>
<td>Gender x Condition x pronoun</td>
<td>-0.125(0.232)</td>
<td>1.2749</td>
<td>-0.54</td>
<td>0.5893</td>
<td></td>
</tr>
<tr>
<td>Gender x Form x Filler sensitivity</td>
<td>-0.076(0.049)</td>
<td>1.2749</td>
<td>-1.56</td>
<td>0.1189</td>
<td></td>
</tr>
<tr>
<td>Gender x Condition x Filler sensitivity</td>
<td>-0.061(0.067)</td>
<td>1.2749</td>
<td>-0.92</td>
<td>0.3582</td>
<td></td>
</tr>
<tr>
<td>Condition x Form x Filler sensitivity</td>
<td>-0.332(0.043)</td>
<td>1.2749</td>
<td>7.72</td>
<td>&lt;.0001 ***</td>
<td></td>
</tr>
</tbody>
</table>

Table A.1. The effects of filler sensitivity on ratings for the two-character conditions. The model included random intercepts for participants and items, as well as random slopes for reference form and condition for both participants and items, and random slopes for gender and filler sensitivity for items.
Women and men have different biases

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate (Error)</th>
<th>DF</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Women vs. Men)</td>
<td>-0.094(0.089)</td>
<td>1.36</td>
<td>-1.05</td>
<td>0.3023</td>
<td></td>
</tr>
<tr>
<td>Condition (N1 vs. N2)</td>
<td>0.045(0.077)</td>
<td>1.36</td>
<td>0.58</td>
<td>0.5651</td>
<td></td>
</tr>
<tr>
<td>Form (Pronoun vs. Name)</td>
<td>0.06(0.021)</td>
<td>1.36</td>
<td>2.85</td>
<td>0.0071</td>
<td>**</td>
</tr>
<tr>
<td>Gender x Form</td>
<td>0.125(0.074)</td>
<td>1.1285</td>
<td>1.69</td>
<td>0.0906</td>
<td>(*)</td>
</tr>
<tr>
<td>Gender x Filler sensitivity</td>
<td>0.018(0.029)</td>
<td>1.1285</td>
<td>0.63</td>
<td>0.5299</td>
<td></td>
</tr>
<tr>
<td>Form x Filler sensitivity</td>
<td>0.034(0.028)</td>
<td>1.1285</td>
<td>1.21</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

Table A.2. The effects of filler sensitivity on ratings for the N1 condition only. The model included random intercepts for participants and items, as well as random slopes for reference form and for both participants and items, and random slopes for gender and filler sensitivity for items.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate (Error)</th>
<th>DF</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Women vs. Men)</td>
<td>-0.148(0.12)</td>
<td>1.36</td>
<td>-1.23</td>
<td>0.2253</td>
<td></td>
</tr>
<tr>
<td>Condition (N1 vs. N2)</td>
<td>0.089(0.126)</td>
<td>1.36</td>
<td>0.71</td>
<td>0.4846</td>
<td></td>
</tr>
<tr>
<td>Form (Pronoun vs. Name)</td>
<td>0.072(0.023)</td>
<td>1.36</td>
<td>3.09</td>
<td>0.0038</td>
<td>**</td>
</tr>
<tr>
<td>Gender x Form</td>
<td>0.225(0.248)</td>
<td>1.1288</td>
<td>0.91</td>
<td>0.3634</td>
<td></td>
</tr>
<tr>
<td>Gender x Filler sensitivity</td>
<td>0.073(0.044)</td>
<td>1.1288</td>
<td>1.65</td>
<td>0.099</td>
<td>(*)</td>
</tr>
<tr>
<td>Form x Filler sensitivity</td>
<td>-0.299(0.046)</td>
<td>1.1288</td>
<td>-6.55</td>
<td>&lt;.0001</td>
<td>***</td>
</tr>
</tbody>
</table>

Table A.3. The effects of filler sensitivity on ratings for the N2 condition only. The model included random intercepts for participants and items, as well as random slopes for reference form and for both participants and items, and random slopes for gender and filler sensitivity for items.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate (Error)</th>
<th>DF</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Women vs. Men)</td>
<td>-0.214(0.115)</td>
<td>1.19</td>
<td>-1.87</td>
<td>0.0773</td>
<td>(*)</td>
</tr>
<tr>
<td>Condition (N1 vs. N2)</td>
<td>0.332(0.111)</td>
<td>1.19</td>
<td>2.99</td>
<td>0.0075</td>
<td>**</td>
</tr>
<tr>
<td>Form (Pronoun vs. Name)</td>
<td>0.04(0.026)</td>
<td>1.19</td>
<td>1.58</td>
<td>0.1305</td>
<td></td>
</tr>
<tr>
<td>Gender x Form</td>
<td>0.29(0.096)</td>
<td>1.1358</td>
<td>3.02</td>
<td>0.0025</td>
<td>**</td>
</tr>
<tr>
<td>Gender x Filler sensitivity</td>
<td>0.006(0.04)</td>
<td>1.1358</td>
<td>0.14</td>
<td>0.8863</td>
<td></td>
</tr>
<tr>
<td>Form x Filler sensitivity</td>
<td>0.04(0.036)</td>
<td>1.1358</td>
<td>1.11</td>
<td>0.268</td>
<td></td>
</tr>
</tbody>
</table>

Table A.4. The effects of filler sensitivity on ratings for the single condition only. The model included random intercepts for participants and items, as well as random slopes for reference form and for both participants and items, and random slopes for gender and filler sensitivity for items.