Doherty, A. & Conklin, K. (2017). How gender expectancy affects the processing of “them”. *Quarterly Journal of Experimental Psychology, 70*(4), 718-735. [DOI: 10.1080/17470218.2016.1154582]

How gender-expectancy affects the processing of “*them*”

Alice Doherty[a] and Kathy Conklin[b]

[a]University of Derby Online Learning, Enterprise Centre, University of Derby, Derby, UK; [b]School of English Studies, University of Nottingham, Nottingham, UK

ABSTRACT

How sensitive is pronoun processing to expectancies based on real-world knowledge and language usage? The current study links research on the integration of gender stereotypes and number-mismatch to explore this question. It focuses on the use of them to refer to antecedents of different levels of gender-expectancy (low–*cyclist*, high–*mechanic*, known–*spokeswoman*). In a rating task, them is considered increasingly unnatural with greater gender-expectancy. However, participants might not be able to differentiate high-expectancy and gender-known antecedents online because they initially search for plural antecedents (e.g., Sanford & Filik), and they make all-or-nothing gender inferences. An eye-tracking study reveals early differences in the processing of *them* with antecedents of high gender-expectancy compared with gender-known antecedents. This suggests that participants have rapid access to the expected gender of the antecedent and the level of that expectancy.

KEYWORDS

Pronoun; Number agreement; Gender agreement; Stereotypical gender; Language usage

INTRODUCTION

Are *they, them,* and *their* plural pronouns? Certainly that is what we might have learnt at school. Indeed the American Psychological Association stipulate in their Publication Manual that; “Each pronoun should refer clearly to its antecedent and should agree with its antecedent in gender and number” (American Psychological Association, APA, 2011, p.79). However in natural language the use of *them* is far less constrained. The current research focuses on *them* used to refer to antecedents that are conceptually and grammatically singular (*singular-them*) as in “I stopped someone and asked *them* for the time”. Although this sentence may sound natural, it is grammatically unacceptable due to the mismatch in number between antecedent and pronoun. However, because *singular-them* occurs frequently in natural language (particularly in British English, Baranowski, 2002), native English speakers might have a usage-based expectancy that *someone-them* is acceptable. The current research explores how such usage-based expectancies might become integrated with conflicting linguistic cues.

Traditionally, *singular-them* is said to be used if the gender of the antecedent is unknown or irrelevant (Bodine, 1975; MacKay, 1980; Meyers, 1990). What constitutes unknown or irrelevant is itself an open empirical question. The current research investigates the use of *singular-them*: when the gender of the antecedent is unknown (*cyclist*), when there are real-world expectancies about the gender of the antecedent (*mechanic*), and when the gender is known (*spokes-woman*). There is evidence that, in establishing reference with antecedents like a mechanic, comprehenders retrieve real-world expectancies about the gender of the antecedent (e.g., Kreiner, Sturt, & Garrod, 2008). These expectancies could influence the way that the fit of *them* is assessed. Therefore the current research explores the integration of linguistic cues (number-match), *usage-based expectancies* (that *them* can be used as a generic pronoun), and real-world expectancies (gender stereotypes) in the processing of *singular-them*.

Although it has been informally observed that *singular-them* is typically used as a generic pronoun, it is an idiosyncratic nuance of natural language that has no formal constraints of use. Previous research on the processing of *singular-them* (e.g., Sanford & Filik, 2007) has focused on them used to refer to gender-neutral, singular antecedents (e.g., *someone*). This is of interest because, although generic *singular-them* may be considered natural by native speakers, it is ungrammatical. Thus there is a linguistic number cue that potentially competes with the usage-based expectancy that *them* could be co-referent with someone. Some models (e.g., Sanford, 1985; Sanford & Filik, 2007) assume that *them* is initially treated as a plural pronoun during online processing before usage-based expectancies can be taken into account. This account predicts that participants should experience a cost whilst they resolve the conflicting number information of the pronoun and its antecedent, even if the antecedent is generic. Other accounts (e.g., Sauerland, Anderssen, & Yatsushiro, 2005) assume that the meaning of the plural is weaker than the singular, so the use of singular pro-nouns is constrained to singular entities but the use of *them* is constrained to cases where a singular pronoun cannot be used. Thus given that generic antecedents cannot be clearly referred to by he/she/ it, there should not be a cost of using *them*. Findings on the processing of generic *singular-them* have so far proven contradictory. Foertsch and Gernsbacher (1997) found no evidence of a cost, but they used a self-paced reading design with large regions of interest (ROIs) that might not have been sensitive enough to detect it. Sanford and Filik (2007) used a more sensitive eye-tracking design and smaller ROIs. They found a cost of generic *singular-them*, in late reading measures. However recent research (Doherty, 2013) found no evidence of a processing cost even though it made use of the same experimental items as those of Sanford and Filik and a very similar design.

Existing findings on the processing of generic *singular-them* are problematic for models that assume that *them* is treated as a specifically plural pronoun and for models that assume that number is weaker for plural than it is for singular pronouns. The current research tests a third alternative account: that the prioritization of cues that are employed during pronoun processing is influenced by language usage (e.g., Arnold, 1998). Many findings from the studies above (and other studies on the processing of mismatching plural pronouns) implicate sensitivity to language use. Although Sanford and Filik (2007) argued that number-match was considered before participants' experience with natural language could be taken into account, they still allowed that this experience modulated the magnitude of the processing cost, and that any initial difficulty in resolving *singular-them* was straightforwardly revised. Additionally findings that have revealed a cost of generic *singular-them* have tended to find the effect at a delay compared to the effect of singular mismatch (e.g., Moxey, Sanford, Sturt, & Morrow, 2004) and, in explaining this result, authors have emphasized differences in the usage of plural and singular pronouns. Specifically it is claimed that experience will have shown that *them* less reliably points to a specific referent because *them* has a wider range of potential antecedents than him or her. These antecedents include: collective sets that could be conceptually singular or plural (e.g., the band; Carreiras & Gernsbacher, 1992), inexplicit institutions (e.g., the service was terrible. They didn't serve my food for an hour; Sanford, Filik, Emmott, & Morrow, 2008), and complex reference objects (e.g., John and Mary; Moxey, Filik, & Paterson, 2009). Because of this range, it is argued that plural pronouns can initially be left underspecified to allow the processing system more time to identify a suitable referent (Filik, Sanford, & Leuthold, 2008; Moxey et al., 2004; Sanford & Filik, 2007). Finally research by Doherty (2013) suggests that the cost associated with generic *singular-them* might be context bound, or changed by exposure to different types of ungrammaticality within an experiment, meaning that short-term language usage may influence online language processing behaviour (see also Kaschak & Glenberg, 2004; Kaschak, Kutta, & Schatschneider, 2011).

The findings above all implicate the role of usage-based expectancies in the processing of *singular-them*. The current research offers another way to address the influence of usage-based expectancies by focusing on the use of different types of antecedent. If the natural use of *singular-them* is constrained to generic antecedents, then using it to refer to antecedents of known gender (e.g., a spokeswoman) violates both usage-based expectancies and linguistic constraints and should lead to a processing cost. Assuming this to be the case, an interesting question arises about antecedents like mechanic that have high gender-expectancy. Formally an antecedent is generic wherever the gender of the antecedent is unknown. However, extensive research has shown that antecedents that have a strong gender stereotype might be formally generic but they are not treated as gender-neutral during online language processing (e.g., Esaulova, Reali, & von Stockhausen, 2013; Garnham, Oakhill, & Reynolds, 2002; Kreiner et al., 2008; Oakhill, Garnham, & Reynolds, 2005; Reynolds, Garnham, & Oakhill, 2006; Sturt, 2003). At an intuitive level it follows that there might be a scale of "generic-ness" based on the real-world likelihood, or probability that an antecedent has a specific gender, with low-expectancy antecedents (*a cyclist*) being treated as highly generic and high-expectancy antecedents (a mechanic) being treated as somewhat generic. This might mean that participants find *them* unacceptable for antecedents that have a known gender (e.g., *spokeswoma*n), and also, but potentially less, unacceptable for antecedents that have high gender-expectancy (e.g., *mechanic*). However, to the authors' knowledge, these predictions have not been empirically tested.

Assuming that participants do show offline sensitivity to gender-expectancy in the assessment of the fit of *singular-them*, it also is of interest to explore whether they show the same sensitivity during online processing. Such evidence could be problematic for the feature-matching approaches described above (e.g., Sanford & Filik, 2007), which assume that in the initial stages of processing *them*, the processing system can only take antecedent number into account. To date there has only been one study exploring the processing of *singular-them* that has manipulated the type of singular antecedent (Foertsch & Gernsbacher, 1997). This self-paced reading study did not pair *singular-them* with antecedents of known gender, but it did explore the role of gender-expectancy. The findings of this study showed that clauses that contained *them* and a high-expectancy antecedent were not significantly more difficult than those that contained a pronoun that matched the expected gender of the antecedent. However, as high-lighted above, their method might not have been fine-grained enough to detect small changes in reading behaviour. The current study, using eye tracking and smaller ROIs, should be more sensitive to transient effects in the reading record.

Whilst the influence of gender-expectancy on the processing of plural pronouns (*mechanic–them*) is unclear, its influence in the processing of gender-marked pronouns (*mechanic–him/her*) has been clearly established. Consistent evidence has shown a cost associated with violations of gender-expectancy (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996). Although the antecedent could be of either gender, participants seem to make a commitment to the expected gender value. There remains significant debate about precisely how gender-expectancy is rep-resented in terms of both where those expectancies are represented (level of representation) and what the nature of *them* is (i.e., whether they are all-or-nothing or probabilistic). Lexical-based accounts (Duffy & Keir, 2004; Osterhout, Bersick, & McLaughlin, 1997; Sturt, 2003) assume that gender-expectancy is encoded as part of the lexical representation of the word, and is all-or-nothing (probability is not taken into account). Mental models accounts (Carreiras et al., 1996; Garnham, 2001; Garnham et al., 2002; Kreiner et al., 2008) assume that spokeswoman has gender encoded at the lexical level but mechanic does not. When a word like mechanic is encountered, and unless there is evidence to the contrary, then an elaborative inference is made that the referent is (probably) male. This information is stored at the dis-course or situation level (for similar claims see Esaulova et al., 2013; Siyanova-Chanturia, Pesciarelli, & Cacciari, 2012). Some researchers (e.g., Carreiras et al., 1996; Kreiner et al., 2008) have claimed that the language processing system makes a firm commitment that the mechanic is male, yet others have allowed that a probabilistic inference might be made (Garnham, 2001; Garnham et al., 2002).

Similar to some mental models approaches, Kennison and Trofe (2003) have proposed that representations of gender-expectancy include information about the frequency with which a particular word is used to refer to males and females. This account is consistent with constraint-based models of reference resolution (e.g., Arnold, 1998), which assume that cues are weighted probabilistically according to their reliability (see also, McRae, Spivey-Knowlton, & Tanenhaus, 1998; Trueswell, Tanenhaus, & Garnsey, 1994). Mechanics are more likely to be male than female, which increases the likelihood that the referent will be referred to with a particular marked pronoun (him). Thus an antecedent with an expected gender value has a higher weighting and is integrated more readily with the pronoun. This can account for the observed costs associated with gender-expectancy violations.

There are two approaches to the representation of gender-expectancy (all-or-nothing accounts and probabilistic accounts). Importantly, these accounts make different predictions about the influence of gender-expectancy on the processing of *singular-them*. Because an all-or-nothing-account cannot differentiate between the gender value of high-expectancy antecedents and antecedents with known gender, neither antecedent should be treated as being any more or less generic. Thus, neither spokeswoman nor mechanic would be an acceptable antecedent for *them*. In contrast, according to probabilistic accounts, spokeswoman should set up a stronger requirement for her than mechanic does for him. In this case, *them* should incur more of a processing cost when the antecedent is spokeswoman than when it is mechanic.

The research reported here used offline (Experiment 1) and online (Experiment 2) measures to explore how native English speakers accommodate gender- and number-mismatch during pronoun resolution with high-expectancy antecedents (*mechanic*) and antecedents with known gender (*spokeswoman*). To follow up on previous studies, the research also explored the on- and offline cost associated with *singular-them* when it is used to refer to low-expectancy antecedents (*cyclist*). In Experiment 1, an acceptability rating task explored whether participants find *them* used to refer antecedents of high gender-expectancy more natural than when it is used to refer to antecedents of known gender. An eye-tracking experiment (Experiment 2) then explored the influence of gender-expectancy and known gender on the online processing of gender-marked pronouns (him, her) and number-mismatching *them*.

Experiment 1

It has been argued that the use of *singular-them* is only acceptable when an antecedent has an unknown or an irrelevant gender (e.g., Bodine, 1975), but there has been little empirical evidence to support this claim. Importantly, if this is the case, then participants should find *them* used to refer to gender-known antecedents (spokeswoman) unacceptable, even if they find it natural when used to refer to low-expectancy antecedents. It is also not clear from this assertion how acceptable English speakers might find *them* when it is used to refer to antecedents that are formally generic but have a high gender-expectancy (mechanic). If participants treat such antecedents as being of a specific gender, then they might treat *them* as being just as unnatural for high-expectancy as for gender-known antecedents. However, if they take into account that mechanic is probably, but not always, male, then they might prefer *mechanic–them* to *spokeswoman–them*. To explore these questions, Experiment 1 employed an offline rating task. A secondary research question was to confirm that native speakers find the use of *singular–them* to refer to generic antecedents (cyclist) relatively natural. Before conducting Experiment 1, and to select antecedents to be employed in Experiments 1 and 2, a norming study was conducted to assess the gender-expectancy associated with a set of over 300 noun-phrases.

Method

Participants

Thirty-eight native English speakers participated for course credit as part of a second year U.K. undergraduate degree course in “English Studies”. Both male and female participants took part in the study.

Materials

Norming study. Eighteen participants who did not take part in Experiment 1 or 2 rated 306 role names (e.g., mechanic). Each name was rated on a 7-point Likert scale (1 = always male, 7 = always female). Sixty items were selected to be included as antecedents in Experiments 1 and 2 in one of three categories: (a) gender-known: 12 items with average ratings less than or equal to 1.5, 12 items with ratings greater than or equal to 6.5; (b) high-expectancy: 12 items rated between 1.5 and 3.5, 12 items rated between 4.5 and 6.5; (c) low-expectancy: 12 items rated between 3.5 and 4.5. To ensure that there was a good level of agreement between participants, items were not included if the standard deviation of ratings was greater than 1 (for a full list of the antecedents and their ratings see Appendix).

Naturalness rating task. Using the antecedents selected from the norming study, 60 experimental items were created. Twenty-four contained an antecedent with high gender-expectancy, as in Example 1, and 24 contained a gender-known antecedent, as in Example 2. Twelve items were included that contained an antecedent with low gender-expectancy, like in Example 3. Fewer low-expectancy items were included to avoid making the experiment too long and because these items were designed to simply confirm a secondary, well-established research question. In the experimental items the first line set the context (1a, 2a, 3a), the first half of the second line (1b, 2b, 3b) introduced the antecedent, and the second half of the second line used a pronoun (PRO) that, based on pragmatics and structure, pointed to that antecedent. A third line was included to conclude the discourse (1c, 2c, 3c). The texts were designed so that they fitted across three lines of text and so that the antecedent and three areas of interest all fitted on Line b. In most cases a single sentence was included per line. In some cases (see, for example, Sentences 3a and 3b) this involved including part of the second sentence on the same line as the first sentence.

1. High-expectancy: gender-match (i), gender-mismatch (ii), them (iii)
    a. Patrick took his car to the local garage for an MOT.
    b. He saw a mechanic there and
        i. |asked him| $_{\text{verb + PRO}}$ |warily| $_{\text{adverb}}$ | if there were any problems.| $_{\text{Final}}$ |
        ii. |asked her| $_{\text{verb + PRO}}$ |warily| $_{\text{adverb}}$ | if there were any problems.| $_{\text{Final}}$ |
        iii. |asked them| $_{\text{verb + PRO}}$ |warily| $_{\text{adverb}}$ | if there were any problems.| $_{\text{Final}}$ |
    c. Later that week he scrapped the car.

2. Gender-known: gender-match (i), gender-mismatch (ii), them (iii)
    a. John was a manager of a company when one of his staff noticed discrepancies in the finances.
    b. He called a spokeswoman and
        i. |instructed her| $_{\text{verb + PRO}}$ |hurriedly| $_{\text{adverb}}$ |to call a press conference.| $_{\text{Final}}$ |
        ii. |instructed him| $_{\text{verb + PRO}}$ |hurriedly| $_{\text{adverb}}$ | to call a press conference.| $_{\text{Final}}$ |
        iii. |instructed them| $_{\text{verb + PRO}}$ |hurriedly| $_{\text{adverb}}$ | to call a press conference.| $_{\text{Final}}$ |
    c. It was important to get to the press, before they found out about the story elsewhere.

3. Low-expectancy: gender-marked (i and ii), them (iii)
    a. Adam recently had an accident in his car. He was emerging from a junction when
    b. he hit a cyclist and
        i. |knocked him| $_{\text{verb + PRO}}$ |straight| $_{\text{adverb}}$ | off the bike.| $_{\text{Final}}$ |
        ii. |knocked her| $_{\text{verb + PRO}}$ |straight| $_{\text{adverb}}$ | off the bike.| $_{\text{Final}}$ |
        iii. |knocked them| $_{\text{verb + PRO}}$ |straight| $_{\text{adverb}}$ | off the bike.| $_{\text{Final}}$ |
    c. Fortunately the cyclist was not badly hurt.

As can be seen in the example sentences, each experimental item had three versions including one of three pronouns (*him/her/them*), an equal number of which were included in each list. The classification of these pro-nouns for analysis differed according to the antecedent. The high-expectancy and gender-known antecedents were included in the same analysis (as separate levels), and the three versions of the sentences corresponded to one of three pronoun conditions: gender-matching (1i and 2i), gender-mismatching (1ii and 2ii), *them* (1iii and 2iii). Because low-expectancy items cannot be gender-matching or mismatching (i.e., *cyclist–him* is just as good as *cyclist–her*), they were included in a separate analysis. Participants still saw one of three versions of an item (*cyclist–him/her/them*), which corresponded to one of two pronoun conditions, specifically gender-marked (3i and 3ii) and *them* (3iii). For low-expectancy generic antecedents there were eight gender-marked items and four items that included *them*.

The same experimental items were employed in Experiments 1 and 2. However, in Experiment 1, only the first two sentences (a and b) were presented. Twenty filler items were included, nine of which included other ungrammaticalities and expectancy violations, including violations of verb agreement (teacher–were) and pronoun agreement (wine–her) so that participants did not identify the experimental items on the basis of ungrammaticality alone. Three counterbalanced questionnaires were constructed in E-Prime® (Psychology Software Tools Inc.). Items were randomized for each participant.

Procedure

Participants were given one questionnaire. They were asked to rate each item for naturalness on a 5-point scale (1 = impossible, 5 = natural). They were given the following examples to explain the task. “She played goodly” would score 1 because it sounds bad, and it is grammatically wrong. “She played bad” would score 3 because we might use it, but it is ungrammatical and therefore sounds a bit strange. “She played badly” would score 5 because it is grammatically correct, and it sounds natural.

Results

Because low-expectancy antecedents cannot have a gender-match/mismatch, they were treated separately from the high-expectancy and gender-known antecedents. For high-expectancy and gender-known antecedents, average ratings were calculated for three pronoun conditions (gender-match, gender-mismatch, *them*) by participant and by item. A 2 × 3 analysis of variance (ANOVA) was conducted based on average ratings by antecedent (within subjects and between items) and pronoun (within subjects and items). In the case of the low-expectancy antecedents, average ratings were calculated by participant and by item for two pronoun conditions (gender-marked or *them*), and comparisons were conducted using a paired-samples t-test. Further analyses employed a series of t-tests (paired-samples by participant, independent-samples by item) to examine the ratings of *them* according to the different levels of gender-expectancy (low-expectancy, high-expectancy, gender-known). Mean ratings for each of the antecedents in each of the conditions can be seen in Table 1.

Antecedents with high gender-expectancy and gender-known antecedents

There was a main effect of antecedent [F1(1, 32) = 109.87, p < .01; F2(1, 46) = 45.93, p < .01], pronoun [F1(2, 64) = 48.62, p < .01; F2(2, 92) = 94.87, p < .01], and a significant interaction [F1(2, 64) = 104.63, p < .01; F2(2, 92) = 45.93, p < .01]. To further explore this effect, separate analyses were conducted for each antecedent. These revealed significant effects of pronoun for gender-known antecedents [F1(2, 64) = 111.89, p < .01; F2(2, 46) = 206.16, p < .01] and for high-expectancy antecedents [F1(2, 64) = 12.75, p < .01; F2(2, 46) = 24.90, p < .01].

For gender-known antecedents, participants rated gender-matching more natural than gender-mismatching pronouns [t1(32) = 16.41, p < .01; t2(23) = 20.51, p < .01]. They rated *them* less natural than gender-matching pronouns [t1(32) = 7.78, p < .01; t2(23) = 11.36, p < .01], but more natural than mismatching pronouns [t_1(32) = 6.36, p < .01; t_2(23) = 8.44, p < .01].

For high-expectancy antecedents there was no difference in participants' ratings of gender-matching and gender-mismatching pronouns [t_1(32) = 1.35, p = .19; t_2(23) = 1.31, p = .20], and *them* was rated significantly less natural than both gender-matching [t_1(32) = 4.25, p < .01; t_2(23) = 6.69, p < .01] and gender-mismatching pronouns [t_1(32) = 3.45, p < .01; t_2(23) = 5.57, p < .01].

Table 1. Mean ratings of naturalness on a 5-point scale and SE of the mean by pronoun and antecedent type in Experiment 1

		Rating	
Antecedent	Pronoun	M	SE
Low-expectancy	Marked	4.0	0.1
	Them	3.3	0.2
High-expectancy	Match	3.8	0.1
	Mismatch	3.7	0.1
	Them	3.0	0.2
Gender-known	Match	4.1	0.1
	Mismatch	1.6	0.1
	Them	2.5	0.2

Note: Scale: 1 = impossible, and 5 = natural. The labels in the heading titles correspond to the pronoun conditions as follows; marked = gender-marked condition; match = gender-match condition; mismatch = gender-mismatch condition.

Antecedents with low gender-expectancy

Participants rated *them* less natural than a gender-marked pronoun [t_1(32) = 3.81, p < .01; t_2(11) = 3.91, p < .01].

The influence of gender-expectancy on ratings of them

Comparisons of the naturalness ratings of *them* for each of the three antecedents (gender-known, high-expectancy, and low-expectancy) revealed that participants rated *them* with a low-expectancy antecedent as significantly more natural than with a high-expectancy one [t_1(32) = 2.20, p < .05; t_2(34) = 5.19, p < .01], and with a gender-known antecedent [t_1(32) = 4.70, p < .01; t_2(34) = 8.14, p < .01]. Results also revealed that participants rated *them* more natural with a high-expectancy than with a gender-known antecedent [t_1(32) = 4.58, p < .01; t_2(46) = 3.11, p < .01].

Discussion

The results of Experiment 1 revealed that participants only find gender-mismatch unnatural for the gender-known antecedents (spokeswoman). Interestingly, this result indicates that participants knew that the antecedents in the high-expectancy condition (mechanic) could be referred to using him/her, and that the use of gender-mismatching pronouns (i.e., her) was therefore acceptable. It was also found that participants found *them* less natural than gender-matching pronouns for high-expectancy and gender-known antecedents, but natural for low-expectancy antecedents. In fact, *them* was found least natural for gender-known antecedents, which suggests that compared with low-and high-expectancy items, gender-known antecedents are treated as the least likely referents of *them*. The findings that participants find *mechanic–them* more natural than *spokeswoman–them* but less natural than *cyclist–them* suggests that participants are aware that high-expectancy antecedents like mechanic are generic but due to gender-expectancy they find it less likely that such antecedents will be referred to with *them* than low-expectancy antecedents.

Experiment 2

The findings of Experiment 1 suggest that, offline, participants may be differentially sensitive to *them* used to refer to a high-expectancy antecedent than a gender-known antecedent. It is not clear whether the online processing system takes into account the gender of the antecedent when resolving *them* and whether it can employ more fine-grained representations of gender-expectancy than current all-or-nothing models allow (e.g., Duffy & Keir, 2004). There-fore, Experiment 2 explored the cost of *them* for high-expectancy and gender-known antecedents during online processing. Again for the purposes of comparison a set of low-expectancy antecedents were included. Consistent with Experiment 1, these items were treated separately in the analyses.

Method

Participants

Thirty-eight native English speakers participated for course credit. None of these participants had taken part in Experiment 1 or in the norming study, although they were all recruited from the same cohort of students as those that took part in Experiment 1.

Materials

Experimental items were the same as those from Experiment 1, but included a final sentence (in examples above, Lines 1c, 2c, 3c). For the eye-tracking analyses, Line b contained three ROIs (verb + PRO, adverb, and final), which all appeared on the same line of text. The ROIs are demarcated in the examples, but not to the participants. The average length of each ROI was closely controlled for each antecedent.[1] In addition to the fillers in Experiment 1, a further 40 fillers were included. Of those, 20 included other inconsistencies and ungrammaticalities like those in Experiment 1. A comprehension question followed 50% of items. Error rates were low (M = 12%, SE = 2).

Apparatus and procedure

Participants read sentences and responded to comprehension questions. Texts were presented double spaced in size 16 Courier New Font. All three lines were presented simultaneously, and participants were instructed to read naturally for comprehension. Eye-movements were monitored using an Eye Link I® eye-tracker (SR research, 250 Hz). Before the experiment started, participants were given instructions, and the eye-tracker was calibrated using a 9-point calibration. It was re-calibrated at least four times throughout the experiment. Gaze direction was drift corrected before each trial.

Analysis and measures

Prior to analysis, the eye-movement data from each participant were subject to the following procedures. Fixations less than 80 ms and within 0.5% visual angle of another fixation and fixations less than 40 ms and within 1.25% visual angle of another fixation were merged with those fixations. Any fixations less than 80 ms or greater than 1200 ms were then deleted.

For each ROI, three reading-measures were calculated (in ms); first-pass reading time (FPRT), the total duration of all fixations that occur from the point at which the region is first fixated until the point at which the region is first exited (either left or right); regression path reading time (RPRT), the total duration of fixations that occur from the point of first fixation, including regressions towards earlier regions of text, until the region is exited to the right; and finally total reading time (TRT), which is the total duration of all fixations that are directed at a ROI. To account for length differences in different regions (a) across the adverb and final regions for high-expectancy versus gender-known antecedents and (b) across the verb + PRO region for *them* versus gender match and gender mismatch, a residual reading measure was calculated for each region and measure. Residual reading scores were used in preference to character length corrections that can bias data that use small ROIs with only small differences in length: That is, the residual reading score is least likely to introduce artefacts (see

Trueswell et al., 1994). The residual reading time was calculated as follows: An expected reading time was calculated for each participant, item, region, and measure based on the length of the ROI.[2] The residual reading score was calculated by subtracting the expected reading time from the actual reading time.

Results

Consistent with Experiment 1, high-expectancy and gender-known antecedents were analysed separately from antecedents with low gender-expectancy. Mean residual reading times for the antecedents followed by each of the pronouns can be seen in Table 2.

Table 2. Mean reading times with SE of the mean by pronoun, antecedent, and ROI in Experiment 2

		Antecedent							
		Low-expectancy		High-expectancy			Gender-known		
Measure		Marked	Them	Match	Mismatch	Them	Match	Mismatch	Them
Verb + PRO region									
FPRT residual	M	−7	−4	2	−8	19	−6	−8	−19
	SE	7	13	9	6	7	6	8	7
RPRT residual	M	−12	−32	0	−14	22	23	16	−50
	SE	27	25	21	17	11	15	26	18
TRT residual	M	−47	−37	−8	40	21	−45	83	−63
	SE	11	24	12	14	15	10	16	13
Adverb region									
FPRT residual	M	−11	4	1	5	−15	−10	16	−12
	SE	6	11	9	9	6	8	9	6
RPRT residual	M	−17	−36	−21	38	−22	−61	112	−47
	SE	20	24	16	28	14	14	23	14
TRT residual	M	−18	−22	−8	34	−26	−40	88	−50
	SE	11	15	15	16	10	12	18	10
Final region									
FPRT residual	M	−27	−30	−5	26	−3	−38	−5	16
	SE	12	29	15	21	18	14	19	14
RPRT residual	M	−117	−100	−25	121	−74	-177	258	−16
	SE	33	42	39	41	28	28	62	28
TRT residual	M	−55	−67	−14	49	−19	−78	96	−10
	SE	18	31	21	21	20	18	31	14

Note: The labels in the heading titles correspond to the pronoun conditions as follows; marked = gender-marked condition; match = gender-match condition; mismatch = gender-mismatch condition. Reading time in milliseconds. ROI = region of interest; PRO = pronoun; FPRT = first-pass reading time; RPRT = regression path reading time; TRT = total reading time.

For high-expectancy and gender-known antecedents, averages were calculated by participant and by item for each reading measure at each ROI and analysed using a 2 (antecedent; high-expectancy, gender-known, within-subjects and between items) by 3 (pronoun; gender-match, gender-mismatch, them, within-subjects and items) ANOVA (Table 3). Further comparisons were conducted wherever there was an effect of pronoun, or an interaction.

Table 3. Results of ANOVA by ROI for each reported reading measure

	Antecedent		Pronoun		Antecedent × Pronoun	
Measure	$F_1(1, 37)$	$F_2(1, 46)$	$F_1(2, 74)$	$F_2(2, 92)$	$F_1(2, 74)$	$F_2(2, 92)$
Verb + PRO region						
FPRT residual	8.00**	2.44	0.60	0.31	3.34*	2.52†
RPRT residual	0.20	0.06	0.89	0.80	3.81*	4.99**
TRT residual	7.99**	1.5	21.61**	22.89**	10.04**	8.50**
Adverb region						
FPRT residual	0.04	0.01	4.72*	3.06†	0.90	0.65
RPRT residual	0.03	0.00	18.59**	20.81**	5.31**	4.13*
TRT residual	0.00	0.00	24.54**	25.34**	6.17**	4.52*
Final region						
FPRT residual	0.71	0.32	1.83	2.41†	1.92	1.70
RPRT residual	0.29	0.09	21.49**	25.69**	6.64**	6.02**
TRT residual	0.03	0.02	10.63**	18.72**	4.25*	3.91*

Note: ANOVA = analysis of variance; ROI = region of interest; PRO = pronoun; FPRT = first-pass reading time; RPRT = regression path reading time; TRT = total reading time.
†p < .10. *p < .05. **p < .01.

When an interaction was found, further analyses looked separately at the effect of pronoun for the two antecedents (high-expectancy and gender-known) using a one-way ANOVA. If the effect of pronoun was significant for either antecedent then follow-up paired-samples t-tests were conducted for that antecedent to compare the averages for each pronoun. Where such comparisons are conducted then we also report whether the mean residual score was positive (+ve) or negative (−ve). A positive mean residual score suggests that reading times were comparatively higher than average for that region, and a negative residual score suggests that reading times were comparatively lower than average for that region. Thus the direction (positive or negative) of these average scores is informative in understanding the overall pattern of results.

If there was no interaction, but an effect of pronoun, three follow-up t-tests were conducted to compare the effect of pronoun, collapsed by antecedent. There were only two levels of antecedent so it was not necessary to conduct any follow-up analyses if an effect of antecedent was found.

For the low-expectancy antecedents, average residual reading times were calculated by participant and by item for each pronoun (gender-marked, them), reading measure, and ROI. Paired-samples t-tests were conducted to compare the averages by pronoun. No further follow-up analyses were necessary.

Consistent with Experiment 1, analyses to compare reading times for sentences that contained *them* based on the three different antecedents (low-expectancy, high-expectancy, gender-known) were also conducted.

Verb + PRO region

Antecedents with high gender-expectancy and gender-known antecedents. For the verb + PRO region, residual first-pass reading times, residual regression path reading times, and residual total reading times all revealed very different patterns of results. For this reason we report all three measures separately.

As can be seen in Table 3, the results of residual FPRT revealed a significant effect of antecedent (by participant only), no effect of pronoun, and an interaction that was significant by participant but marginal by item. To explore the interaction, two ANOVAs were conducted to examine the effect of pronoun for the two antecedent types. They revealed an effect of pronoun for high-expectancy antecedents by participant only [$F_1(2, 74) = 3.48$, $p < .05$; $F_2(2, 46) = 1.88$, $p = .16$], but no effect of pronoun for the gender-known antecedents (all Fs < 1).

To examine the effect of pronoun that was observed for the high-expectancy antecedents, further t-tests were conducted. These follow-up comparisons indicated increased reading times for *them* (+ve mean residual score) compared with gender-mismatch (−ve mean residual score); however, the comparison was marginal by item [$t_1(37) = 3.59$, $p < .01$; $t_2(23) = 1.85$, $p = .08$]. Comparisons for gender-match versus mismatch (all ts < 1) and for *them* versus gender-match [$t_1(37) = 1.38$, $p = .17$; $t_2(23) = 1.15$, $p = .26$] revealed no significant effects.

Table 3 shows that for residual RPRT the effects of antecedent and pronoun were not significant. The interaction was significant. Looking at the effect of pronoun by antecedent, no difference was observed for the high-expectancy antecedents. There was a significant effect for the gender-known antecedents in the by-item comparisons with a marginal effect in the by-participant comparisons [F_1(2, 74) = 2.99, p = .06; F_2(2, 46) = 5.56, p < .01].

Further comparisons to explore the effect of pronoun for gender-known antecedents indicated that there was an advantage for *them* (−ve mean residual) compared with match and mismatch (+ve mean residual): *them* versus gender-match [t1(37) = 2.6, p < .05; t2(23) = 4.19, p < .01]; *them* versus gender-mismatch (significant by item only) [t1(37) = 1.62, p = .11; t2(23) = 2.26, p < .05]. For gender-match versus mismatch, all ts < 1.

For residual TRT both the effects of antecedent and pronoun were significant (see Table 3). There was also a significant interaction. When the effect of pronoun was explored separately by antecedent the results revealed a significant effect for gender-known antecedents only [F_1(2, 74) = 28.07, p < .01; F_2(2, 46) = 40.14, p < .01]. For high-expectancy antecedents the effect of pronoun was not significant in either measure [F_1(2, 74) = 2.89, p = .06; F_2(2, 46) = 2.34, p = .11].

The further comparisons for gender-known antecedents revealed a cost of gender-mismatch (+ve mean residual) versus match (−ve mean residual) [t_1(37) = 6.08, p < .01; t_2(23) = 6.53, p < .01]. There was no difference between *them* (−ve mean residual) versus match [t_1(37) = 1.09, p = .28; t_2(23) = 1.42, p = .17] and a cost of gender-mismatch compared with *them* [t_1(37) = 5.79, p < .01; t_2(23) = 7.11, p < .01].

Low-expectancy antecedent. There was no effect of pronoun by any measure, all ts < 1.

The influence of gender-expectancy on reading times where PRO = them. There was a significant effect of antecedent for all measures: residual FPRT [F_1(2, 74) = 3.21, p < .05; F $_2$(2, 57) = 4.06, p < .05], residual RPRT [F_1(2, 74) = 4.01, p < .05; F_2(2, 57) = 4.99, p < .05], residual TRT [F_1(2, 74) = 5.00, p < .01; F_2(2, 57) = 5.86, p < .01]. The pattern of results was the same each time (see Figure 1). Although numerically the residual reading scores for low-expectancy antecedents (−ve mean residual) fell between the high-expectancy (−ve mean residual) and gender-known antecedents (+ve mean residual), there was no evidence of a difference between low- and high-expectancy antecedents for each measure—residual FPRT [t_1(37) = −1.27, p = .21; t_2(34) = −1.16, p = .25], residual RPRT [t_1(37) = −1.77, p = .08; t_2(34) = −1.71, p = .10], residual TRT [t_1(37) = −1.77, p = .09; t_2(34) = −1.66, p = .11]—and no evidence of a difference between low-expectancy and gender-known antecedents for each measure—residual FPRT [t1(37) = 0.96, p = .34; t2(34) = 1.39, p = .17], residual RPRT [t1(37) = 0.70, p = .49; t2(34) = 0.74, p = .46], residual TRT [t1(37) = 1.01, p = .32; t2(34) = 1.06, p = .30]. However, there was a significant difference between gender-known and high-expectancy antecedents in all cases: residual FPRT [t1(37) = 3.63, p < .01; t2(46) = 2.71, p < .01], residual RPRT [t1(37) = 3.13, p < .01; t2(46) = 2.93, p < .01], residual TRT [t1(37) = 3.84, p < .01; t2(46) = 3.29, p < .01].

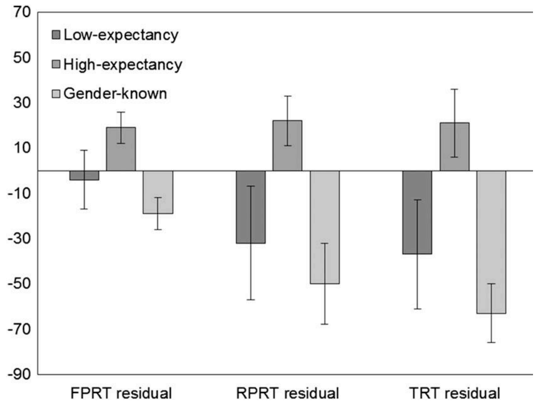

Figure 1. A summary of mean residual reading times (error bars represent ±1 standard error of the mean) for *them* by antecedent (high, low, gender-known) for first-pass, regression path, and total reading times in the verb + PRO region. PRO = pronoun; FPRT = first-pass reading time; RPRT = regression path reading time; TRT = total reading time.

Summary: verb+PRO region. At this region the results revealed some interesting differences in the effect of pronoun for the high-expectancy compared with the gender-known antecedents. These differences were supported by an antecedent by pronoun interaction for all measures and are summarized in Figure 2.

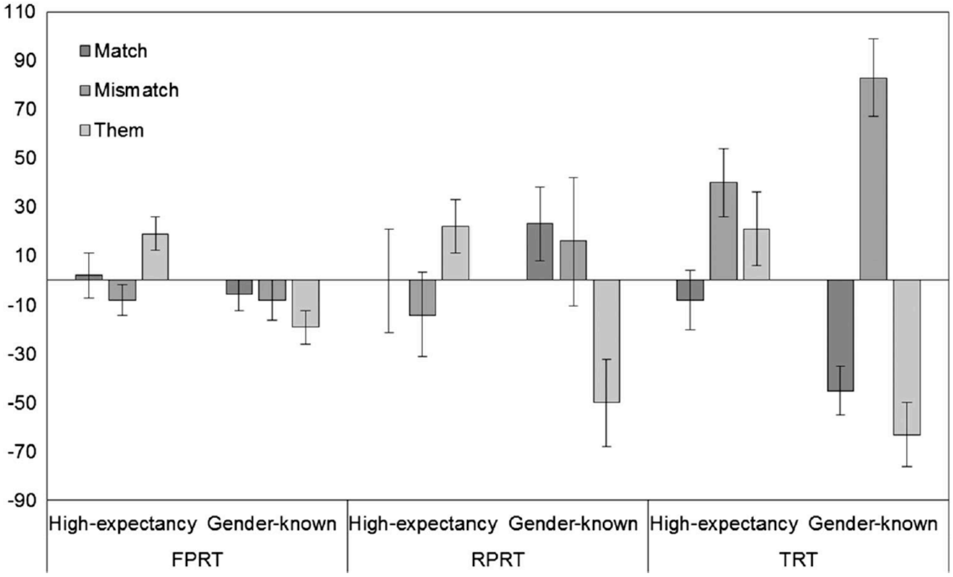

Figure 2. A summary of mean residual reading times (error bars represent ±1 standard error of the mean) by antecedent (high, gender-known) and pronoun type for first-pass, regression path, and total reading times in the verb + PRO region. PRO = pronoun; FPRT = first-pass reading time; RPRT = regression path reading time; TRT = total reading time.

Overall the findings at the verb + PRO region suggest an advantage of *them* for gender-known antecedents and a disadvantage of *them* for high-expectancy antecedents. Late measures (specifically residual TRT) also revealed an emergent cost associated with gender-mismatch for gender-known antecedents. Importantly, the comparative disadvantage of *them* for the high-expectancy compared with the gender-known antecedents is supported statistically in all three measures. In residual FPRT the disadvantage of *them* for high-expectancy antecedents is further supported by a statistically significant difference between gender-mismatching pronouns and *them* (but not gender-match vs. *them*). In residual RPRT the advantage for *them* is supported by significant differences for *them* compared with both gender-match and gender-mismatch. So there is clear evidence of a processing difference associated with *them* for high-expectancy compared with gender-known antecedents. This difference emerges very early in the reading record (in FPRT). Reading times for *them* used to refer to low-expectancy antecedents did not differ significantly from either high-expectancy or gender-known antecedents.

Adverb region

Antecedents with high gender-expectancy and gender-known antecedents. For this region, as shown in Table 3, residual FPRT patterned differently from residual RPRT and residual TRT, which both showed a much more similar pattern of results. For this reason we report residual FPRT separately, and residual RPRT and residual TRT together.

For residual FPRT there was no significant effect of antecedent or interaction, but a significant effect of pronoun (although this was marginal by item). Comparisons collapsed by antecedent revealed a significant cost of gender-mismatch versus gender-match, which was marginal by participant (gender-match, M = −5, SE = 5; gender-mismatch, M = 10, SE = 6) [$t_1(37) = 1.79$, $p = .08$; $t_2(47) = 2.07$, $p < .05$]. No difference was found between gender-match and *them* (them M = −14, SE = 4) [$t_1(37) = 1.20$, $p = .24$; $t_2(47) = 0.67$, $p = .50$]. Third, we observed a cost of gender-mismatch versus *them* [$t_1(37) = 3.25$,

$p < .01$; $t_2(47) = 2.24$, $p < .05$]. Thus these results indicated a cost of gender-mismatch compared with both gender-match and *them*.

For residual RPRT and residual TRT the effect of antecedent was not significant, but there was a significant effect of pronoun and a significant interaction (see Table 3). For high-expectancy antecedents there was an effect of pronoun that was marginally significant by participant in residual RPRT: residual RPRT [$F_1(2, 74) = 2.57$, $p = .08$; $F_2(2, 46) = 3.49$, $p < .05$], residual TRT [$F_1(2, 74) = 4.64$, $p < .05$; $F_2(2, 46) = 6.24$, $p < .01$]. For gender-known antecedents the effect of pronoun was highly significant: residual RPRT [$F_1(2, 74) = 26.72$, $p < .01$; $F_2(2, 46) = 20.15$, $p < .01$], residual TRT [$F_1(2, 74) = 28.71$, $p < .01$; $F_2(2, 46) = 19.97$, $p < .01$].

There was a cost of gender-mismatch versus match for both antecedent types (although note that this effect was marginal by participant for high-expectancy antecedents): high-expectancy antecedents, residual RPRT [t1(37) = 1.76, p = .09; t2(23) = 2.29, p < .05], residual TRT [t1(37) = 1.80, p = .08; t2(23) = 2.97, p < .01]; gender-known antecedents, residual RPRT [t1(37) = 5.79, p < .01; t2(23) = 5.6, p < .01], residual TRT [t1(37) = 5.54, p < .01; t2(23) = 7.62, p < .01].

There was no difference between gender-match and *them* at any measure for either the high-expectancy or the gender-known antecedents (all ts ≤ 1.16).

For high-expectancy antecedents there was a cost of mismatch versus *them* in residual TRT only: residual RPRT [$t_1(37) = 1.75$, $p = .09$; $t_2(23) = 1.92$, $p = .07$], residual TRT [$t_1(37) = 3.39$, $p < .01$; $t_2(23) = 3.01$, $p < .01$]. For the gender-known antecedents this cost was significant in both measures: residual RPRT [$t_1(37) = 5.27$, $p < .01$; $t_2(23) = 4.33$, $p < .01$], residual TRT [$t_1(37) = 6.51$, $p < .01$; $t_2(23) = 4.45$, $p < .01$].

Low-expectancy antecedents. There was no effect of pronoun by any measure: residual FPRT [$t_1(11) = 1.61$, $p = .12$; $t_2 = 1.15$, $p = .27$], residual RPRT [$t_1(11) < 1$; $t_2 = 1.18$, $p = .26$], residual TRT (all ts < 1).

The influence of gender-expectancy on reading times where PRO = *them*. There was no effect of antecedent for any of the three measures: residual FPRT [$F_1(2, 74) = 1.48$, $p = .23$; $F_2(2, 57) = 0.74$, $p = .48$], residual RPRT (all Fs < 1), residual TRT [$F_1(2, 74) = 1.7$, $p = .19$; $F_2(2, 57) = 0.58$, $p = .56$].

Summary: adverb region. At the adverb region the results revealed a cost of gender mismatch. For most measures this contrast emerged in comparisons between both gender match and them, the exception being residual RPRT for high-expectancy antecedents where the cost was observed in comparisons between gender-mismatch and match but by item only and not at all for gender mismatch versus them. There was an interaction between antecedent and pronoun, which was observed in residual regression path and total reading times. However, the overall pattern of results was consistent for both measures, and comparisons by antecedent based on reading times for them did not show any significant effects, which means that this interaction most likely reflected a greater cost of mismatch for the gender-known antecedents than for the high-expectancy antecedents. None of the comparisons that were conducted supported a cost or advantage of them compared with match, and there was no difference associated with the reading times of them based on the antecedent type.

Final region

Antecedents with high gender-expectancy and gender-known antecedents. Similar to the adverb region, we describe the pattern of results for the residual FPRT separately from residual RPRT and residual TRT, which are reported together. Table 3 provides a full summary of the results of the comparisons.

For residual FPRT there were no significant effects or interactions.

For residual RPRT and residual TRT there was no effect of antecedent. There was a significant effect of pronoun and a significant interaction.

For both antecedent types there was an effect of pronoun that was significant: high-expectancy antecedents, residual RPRT [$F_1(2, 74) = 6.85$, $p < .01$; $F_2(2, 46) = 5.83$, $p < .01$], residual TRT [$F_1(2, 74) = 3.33$, $p < .05$; $F_2(2, 46) = 3.46$, $p < .05$]; gender-known antecedents, residual RPRT [$F_1(2, 74) = 20.33$, $p < .01$; $F_2(2, 46) = 24.65$, $p < .01$], residual TRT [$F_1(2, 74) = 11.87$, $p < .01$; $F_2(2, 46) = 20.41$, $p < .01$].

There was a significant cost of match versus mis-match across all of the reported measures for both high-expectancy and gender-known antecedents. This effect was marginal for residual TRT by participant for the high-expectancy items: residual RPRT [$t_1(37) = 2.23$, $p < .05$; $t_2(23) = 2.55$, $p < .05$], residual TRT [$t_1(37) = 1.86$, $p = .07$; $t_2(23) = 2.40$, $p < .05$]. It was consistently significant for the gender-known items: residual RPRT [$t_1(37) = 5.29$, $p < .01$; $t_2(23) = 6.25$, $p < .01$], residual TRT [$t_1(37) = 3.88$, $p < .01$; $t_2(23) = 5.99$, $p < .01$].

The comparisons of match versus *them* differed by antecedent type. For high-expectancy antecedents

the results did not reveal any significant differences (all ts ≤ 1). For the gender-known antecedents, in contrast, the comparisons revealed a cost of *them* compared with gender match: residual RPRT [$t_1(37) = 4.77$, $p < .01$; $t_2(23) = 4.73$, $p < .01$], residual TRT [$t_1(37) = 3.6$, $p < .01$; $t_2(23) = 4.1$, $p < .01$].

The comparisons of mismatch versus *them* were broadly similar for the two antecedent types, with a cost of mismatch for high-expectancy antecedents— residual RPRT [$t_1(37) = 3.84$, $p < .01$; $t_2(23) = 3.02$, $p < .01$], residual TRT [$t_1(37) = 2.14$, $p < .05$; $t_2(23) = 2.59$, $p < .05$]—and gender-known antecedents—residual RPRT [$t_1(37) = 3.43$, $p < .01$; $t_2(23) = 3.61$, $p < .01$], residual TRT [$t_1(37) = 2.72$, $p < .01$; $t_2(23) = 3.10$, $p < .01$].

Low-expectancy antecedents. There was no effect of pronoun by any measure (all ts < 1).

The influence of gender-expectancy on reading times where PRO = them. There was no evidence of any effect of the gender-expectancy of the antecedent based on comparisons of reading times for *them*: residual FPRT [$F_1(2, 74) = 1.48$, $p = .23$; $F_2(2, 57) = 0.74$, $p = .48$], residual RPRT [$F_1(2, 74) = 1.81$, $p = .17$; $F_2(2, 57) = 0.91$, $p = .41$], residual TRT [$F_1(2, 74) = 2.18$, $p = .12$; $F_2(2, 57) = 0.93$, $p = .40$].

Summary: final region. For the final region residual RPRT and TRT revealed a cost of gender mismatch compared with both gender match and *them* for both gender-known and high-expectancy antecedents. For the high-expectancy antecedents some of these comparisons were marginal, which suggests that the cost of mismatch was weaker. The results also revealed a cost of *them* compared with gender-matching pronouns, but only for gender-known and not for high-expectancy antecedents. Interestingly, however, the comparisons conducted based on reading times for *them* did not reveal any difference by antecedent. Thus the results suggest that there is a greater advantage of gender-match for the gender-known than for the high-expectancy antecedents. The pattern of results, which highlights the differential effect of *them* and gender match for high-expectancy and gender-known antecedents, can be seen in Figure 3.

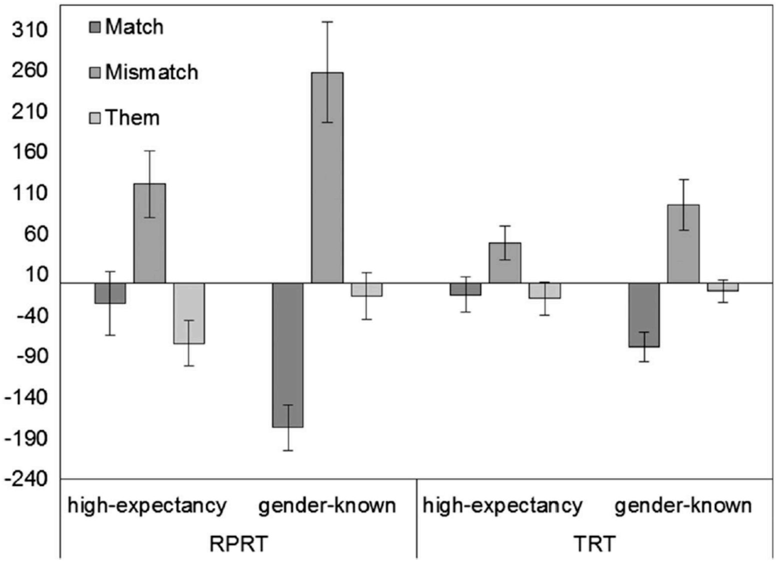

Figure 3. A summary of mean residual reading times (error bars represent ±1 standard error of the mean) by antecedent (high, gender-known) and pronoun type for regression path and total reading times in the final region. PRO = pronoun; RPRT = regression path reading time; TRT = total reading time.

Discussion

In summary, the results indicated that participants experienced a processing cost of gender-mismatch for both high-expectancy and gender-known antecedents. The data further suggested that the cost of gender-mismatch for gender-known antecedents was greater than that for high-expectancy antecedents, indicating that the magnitude of these costs is associated with level of gender-expectancy. At the verb + PRO region the data indicated that processing patterns for *them* differ based on the antecedent, such that for the high-expectancy antecedents there appeared to be some processing cost (in residual FPRT—compared with gender mismatch) while for the gender-known antecedents there was a clear processing advantage, particularly in residual RPRT. These differences in reading times for *them* by antecedent (high-expectancy vs. gender-known) emerged very early in the reading record. Participants also experienced a cost of *the*m for gender-known antecedents but this time, late in the reading record in residual RPRT and residual TRT at the final region compared to gender-match but not compared to *mechanic-them*. In the verb + PRO region (in residual RPRT) there was an apparent advantage of *them* compared with gender-match for gender-known antecedents. It therefore seems that, for both high-expectancy and gender-known antecedents, any cost associated with the use of *them* is much more transient and less robust than the cost of gender-mismatch. There was no evidence of a cost or advantage associated with *them* for antecedents with low gender-expectancy at any region or measure, nor was there any evidence of a cost/advantage of *them* for low-expectancy antecedents compared with gender-known or high-expectancy antecedents. All of these findings are considered in further detail in the General Discussion.

General discussion

The study has revealed a number of novel findings that have interesting implications in terms of models of pronoun resolution and the representation of gender. Most critically the results have indicated that participants are differentially sensitive to the use of *them* to refer to high-expectancy and gender-known antecedents in offline, but also online, measures. They also suggest that participants can easily integrate *them* when it is used to refer to low-expectancy antecedents. These findings and their implications are addressed in more detail.

Naturalness ratings revealed that participants find *them* increasingly less natural as gender-expectancy increases. This is in line with the argument that *them* is most likely to be used where the gender of the antecedent is unknown or is considered irrelevant (Bodine, 1975; MacKay, 1980; Meyers, 1990). Taking into account that mechanic is formally generic, then the finding that *someone–them* is treated as more natural than *mechanic–them* is challenging for under-specification models (e.g., Sauerland et al., 2005) that assume that *them* is acceptable wherever there is an antecedent that cannot be clearly referred to by he/ she/it. The eye-movement results suggest that, during online processing, participants are sensitive to the usage-based expectancies elucidated in the offline task. In particular, the finding that participants experience different processing patterns for *them* associated with gender-known and high-expectancy antecedents suggests that they take into account these different levels of gender-expectancy when establishing co-referential links with *them*.

The finding that *them* is processed differently for high-expectancy and gender-known antecedents is in contrast to that of Foertsch and Gernsbacher (1997). This suggests that the current methodology was indeed more sensitive to the effect of number mismatch. Crucially, this finding challenges feature matching models that assume that plural pronouns are initially only matched for number (Sanford, 1985; Sanford & Filik, 2007) and therefore predict that gender-known and high-expectancy antecedents should all demonstrate the same pattern of processing cost. The findings have also shown that participants experienced less online difficulty with *them* than a gender-mismatching pronoun when referring to a high-expectancy antecedent and an antecedent of known gender. This suggests that participants find it easier to accommodate plural number-mismatch than gender-mismatch and violations of gender expectancy. This may be because plural pronouns have a wider range of potential referents and therefore less reliably point to a unique referent (Moxey et al., 2004; Sanford & Filik, 2007). However, it could also be that, because it has greater social implications (Bodine, 1975), referring to someone with a pronoun that is marked with the wrong gender is less expected and less probable than using a number mismatching pronoun.

The results of the current study support previous findings (e.g., Kreiner et al., 2008) indicating that violations of the known or highly expected gender of an antecedent (spokeswoman–him, mechanic–her) lead to robust online processing difficulties, even though the naturalness ratings indicate that participants know that they are acceptable in the case of high-expectancy antecedents. Thus for both the high-expectancy and the gender-known antecedents, it would seem that participants have activated the probable (or known) gender of the antecedent. In assessing the fit of *them*, participants seemed to be able to discriminate these two antecedents from very early in the reading record (in residual FPRT at the verb + PRO region). This poses a challenge to models that assume that gender-expectancy is encoded as binary or all-or-nothing (e.g., Duffy & Keir, 2004; Sturt, 2003) and therefore assume that antecedents that have a gender-expectancy are essentially treated as gender-known antecedents. Such models would not predict any difference between *spokeswoman–them* and *mechanic–them*.

The finding that participants can differentiate gender-known from high-expectancy antecedents in assessing the fit of *them* could be seen to be consistent with a constraint-based model of reference resolution (Arnold, 1998) that assumes that cues are probabilistically weighted according to their reliability (see also Kennison & Trofe, 2003). Because gender-known antecedents are the least generic, they are the least likely to be referred to using *them*. High-expectancy antecedents have a strong gender association, but they are formally generic and therefore might be referred to using *them*. Indeed the findings of Experiment 1 support that participants find *them* used to refer to gender-known antecedents less natural (or less expected) than when used to refer to high-expectancy antecedents.

Constraint-based models of pronoun resolution assume that cues to reference resolution are weighted probabilistically. Intuitively it follows that the cost of *them* used to refer to a gender-known antecedent is likely to be greater than when it is used to refer to a high-expectancy antecedent. However, this is not what the findings in Experiment 2 showed. In fact the results at the verb + PRO region (relatively early in the reading record) indicate an advantage (in residual RPRT) associated with *them* for gender-known antecedents. This advantage emerges based on comparisons with high-expectancy antecedents but also based on comparisons with gender-match and mismatch. It is only later in the reading record (in residual RPRT and TRT at the final region) that any cost associated with using *them* to refer to gender-known antecedents emerges, and, even then, this finding is based on comparisons with sentences that contain a gender-matching pronoun and not based on comparisons of reading times for *them* used to refer to high-expectancy antecedents (thus it could be that this shows a comparative advantage of known gender-match—rather than any specific disadvantage for *them*).

Why might participants find *them* used to refer to gender-known antecedents initially easier to process than when used to refer to high-expectancy antecedents? One possibility is to assume that, when it is used to refer to a gender-known antecedent, *them* is initially left unresolved because there is not sufficient evidence to support the referent becoming integrated with the pronoun, and usage-based experience has shown that *them* less reliably points to a specific referent than a gender-marked pronoun. A range of evidence suggests that under certain conditions a pronoun can be left unresolved (e.g., Rigalleau, Caplan, & Baudiffier, 2004). It has also been claimed that gender-marked singular pronouns place greater pressure on the processing system for immediate resolution than *them* (Filik et al., 2008; Moxey et al., 2004; Sanford & Filik, 2007). Under this account then, in the high-expectancy condition, because the antecedent is formally generic, *them* may be treated as an acceptable pronoun, just not the preferred one. This account can accommodate the early advantage of *them* that is observed for the gender-known antecedents. However, it is not clear why there are no significant differences in reading times at this region for *them* when used to refer to low-expectancy antecedents compared to either the gender-known or the high-expectancy antecedents.

We have thus far argued that the results challenge all-or-nothing accounts of gender representation on the basis that they cannot differentiate between high-expectancy antecedents and gender-known antecedents. The current results are particularly problematic for models (e.g., Duffy & Keir, 2004; Sturt, 2003) that assume that gender stereotypes are lexically encoded. However some “mental models” approaches (e.g., Kreiner et al., 2008) assume different levels of gender encoding for gender-known antecedents (both at the lexical and at the discourse level) and antecedents with high gender-expectancy (dis-course level only). Postulating two different levels of gender encoding in this way could explain the difference between the cost of *them* for gender-known antecedents and high-expectancy antecedents, whilst maintaining the assumption that gender-expectancy is encoded as an all-or-nothing cue. To do so it would need to be assumed that the assessment of the fit of *them* initially only takes into account the lexical level of gender representation. However, in this case it is not clear how the model might account for the evidence of an (early) cost of *them* associated with high-expectancy antecedents (which is significant in comparisons with *mechanic–her* and *spokeswoman–them*) in residual FPRT at the verb + PRO region. According to a constraint-based model, it could be argued that this cost emerged because *them* offered a "good" but not the "best" fit. So, in contrast with gender-known antecedents, for the high-expectancy antecedents *them* is immediately integrated but with some additional processing effort.

The main challenge posed for all-or-nothing accounts of gender-expectancy representation is that they cannot differentiate between the fact that high-expectancy antecedents are less likely to be of a specific gender category than gender-known antecedents, and therefore more likely to be referred to by *them*. According to the unrestricted race model of sentence parsing (van Gompel, Pickering, & Traxler, 2001) in cases of ambiguity the processing system makes an early commitment to a single interpretation that is determined by probabilistic information. The more likely a particular interpretation becomes, the more likely it is that the processing system will consistently make the same interpretation. Applying this model to the current study, and assuming that participants make one of three commitments (male, female, unspecified), then both high-expectancy and low-expectancy antecedents are gender ambiguous but participants are more likely to make a commit-ment to a gender value where an antecedent has high-expectancy. This account permits the processing system to differentiate high-expectancy and gender-known antecedents on the basis of their gender-expectancy because participants should select a specific gender value for gender-known antecedents while, for high-expectancy antecedents, participants might select the expected gender value or leave gender unspecified. However, assuming that an advantage of *them* only emerges when a gender value has been specified (as for the gender-known antecedents), then this model predicts that some participants might experience an advantage of *them* for high-expectancy antecedents whilst others might not. This model would therefore predict a modulation of the magnitude of the processing advantage, but there is no evidence to support this.

The results for the low-expectancy antecedents were somewhat inconclusive, and there was no evidence of cost associated with *them* for low-expectancy antecedents (a cyclist) when compared with gender-marked pronouns. However, using very similar items, a late cost of *someone–them* has been observed in other studies (Doherty, 2013; Sanford & Filik, 2007). Looking more carefully at the past research and the current study, it appears that under some conditions, a plural match constraint might have a higher weighting (i.e., when all of the filler stimuli are grammatical as in the Doherty, 2013, and the Sanford & Filik, 2007, studies), which could prevent a generic antecedent from becoming integrated with *them*. This is a question for further research.

The current study provides clear evidence that experience with language can be rapidly integrated during language processing. An issue that needs to be addressed is how such experience should be defined. One approach is to characterize it solely in terms of language usage—that is, the number of times that participants encounter the use of different pronouns to refer to a specific antecedent. An immediate question that arises from this argument is whether participants will have been exposed to enough specific cases for such learning to occur. If not, then a clear account that explains how participants learn to generalize from one case to another needs to be developed. Additionally, differences in the magnitude of the cost associated with violations of number and gender could be accounted for in terms of exposure, but violations of gender-match also have social implications that might influence the processing cost (Bodine, 1975). This could imply that participants are also integrating knowledge of the social constraints of language usage.

Finally, for ease of exposition, the current study has differentiated language usage and real-world expectancies. In fact the boundaries between these two types of knowledge are likely to be much fuzzier. For example, is it enough to argue that the constraints that lead to gender mismatch effects for antecedents with high gender-expectancy reflect real-world knowledge alone? Or do the differences in processing time between nurse–him and nurse–her simply reflect statistical differences in language usage? The current research cannot address these issues, but nonetheless they are important questions for consideration.

In sum, the current experiments explore how native English speakers employ real-world knowledge and usage-based expectancies during off- and online processing of matching and mismatching pronouns. It is seen that the processing system is differentially sensitive to the fit of *them* when used to refer to high-expectancy and gender-known antecedents, suggesting that *them* becomes harder to integrate as gender-expectancy increases. This finding has a number of important implications. First, it suggests that participants must have access to information about the gender-expectancy of potential antecedents when resolving *them*, and that this information is available very early during processing. It further suggests that genericness is based on the likelihood that an antecedent can be of a particular gender according to real-world expectancies. These findings challenge models of reference resolution that assume that initially the processing system only considers number in identifying acceptable referents for plural pronouns (Sanford, 1985; Sanford & Filik, 2007) and suggests that participants have acquired a usage-based expectancy that *them* can be used to refer to gender-neutral antecedents. Furthermore, these findings offer a significant challenge to models of gender representation that assume that, like antecedents of known gender, high-expectancy antecedents are treated as having a specific lexically encoded gender value (e.g., Duffy & Keir, 2004). The findings can be best encapsulated within a probabilistic approach on the basis that the likelihood that *them* is used in reference to a given antecedent decreases as the gender-expectancy of that antecedent increases.

References

American Psychological Association. (2011). Publication manual of the American Psychological Society (6th ed.). Washington, DC: Author.

Arnold, J. E. (1998). Reference form and discourse patterns (Doctoral dissertation). Retrieved from http://www.unc.edu/~jarnold/papers/diss/fulldissertation.doc

Baranowski, M. (2002). Current usage of the epicene pronoun in written English. Journal of Sociolinguistics, 6(3), 378–397.

Bodine, A. (1975). Androcentrism in prescriptive grammar: Singular “they”, sex-indefinite “he”, and “he or she”. Language in Society, 4(2), 129–146.

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. Quarterly Journal of Experimental Psychology, 49A(3), 639–663.

Carreiras, M., & Gernsbacher, M. A. (1992). Comprehending conceptual anaphors in Spanish. Language and Cognitive Processes, 7(3–4), 281–299.

Doherty, A. (2013). The use of gender and number cues in L2 pronoun processing (Unpublished doctoral dissertation). University of Nottingham, Nottingham.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. Memory & Cognition, 32(4), 551–559.

Esaulova, Y., Reali, C., & von Stockhausen, L. (2013). Influences of grammatical and stereotypical gender during reading: Eye movements in pronominal and noun phrase anaphor. Language and Cognitive Processes, 29, 1–23.

Filik, R., Sanford, A. J., & Leuthold, H. (2008). Processing pronouns without antecedents: Evidence from event-related brain potentials. Journal of Cognitive Neuroscience, 20(7), 1315–1326.

Foertsch, J., & Gernsbacher, M. A. (1997). In search of gender neutrality: Is singular they a cognitively efficient substitute for generic he? Psychological Science, 8(2), 106–111.

Garnham, A. (2001). Mental models and the interpretation of anaphora. Hove, UK: Psychology Press.

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters’ gender made elaboratively?. Memory & Cognition, 30(3), 439–446.

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. Journal of Experimental Psychology-General, 133(3), 450–467.

Kaschak, M. P., Kutta, T. J., & Schatschneider, C. (2011). Long-term cumulative structural priming persists for (at least) one week. Memory & Cognition, 39(3), 381–388.

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. Journal of Psycholinguistic Research, 32(3), 355–378.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. Journal of Memory and Language, 58
(2), 239–261.

MacKay, D. G. (1980). On the goals, principles, and procedures for prescriptive grammar: Singular they. Language in Society, 9, 349–367.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language, 38(3), 283–312.

Meyers, M. W. (1990). Current generic pronoun usage: An empirical study. American Speech, 65, 228–237.

Moxey, L. M., Filik, R., & Paterson, K. B. (2009). On line effects of what is expected on the resolution of plural pronouns. Language & Cognitive Processes, 24, 843–875.

Moxey, L. M., Sanford, A. J., Sturt, P., & Morrow, L. I. (2004). Constraints on the formation of plural reference objects: The influence of role, conjunction, and type of description. Journal of Memory and Language, 51(3), 346–364.

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. Memory & Cognition, 33(6), 972–983.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. Memory & Cognition, 25(3), 273–285.

Reynolds, D. J., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. Quarterly Journal of Experimental Psychology, 59 (5), 886–903.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. Quarterly Journal of Experimental Psychology, 57A(5), 893–933.

Sanford, A. J. (1985). Aspects of pronoun interpretation: Evaluation of search formulations of inference. In G. Rickheit & H. Strohner (Eds.), Inferences in text processing (pp. 183– 204). Amsterdam, The Netherlands: Elsevier Science.

Sanford, A. J., & Filik, R. (2007). “They” as a gender-unspecified singular pronoun: Eye tracking reveals a processing cost. The Quarterly Journal of Experimental Psychology, 60(2), 171–178.

Sanford, A. J., Filik, R., Emmott, C., & Morrow, L. (2008). They’re digging up the road again: The processing cost of Institutional They. The Quarterly Journal of Experimental Psychology, 61(3), 372–380.

Sauerland, U., Anderssen, J., & Yatsushiro, K. (2005). The plural is semantically unmarked. In S. Kepser & M. Reis (Eds.), Linguistic evidence: Empirical, theoretical & computational perspectives (pp. 413–434). The Hague: Mouton de Groutyer.

Siyanova-Chanturia, A., Pesciarelli, F., & Cacciari, C. (2012). The electrophysiological underpinnings of processing gender stereotypes in language. PLoS ONE, 7(12), e48712.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. Journal of Memory and Language, 48(3), 542–562.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing- use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language, 33(3), 285–318.

van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. Journal of Memory and Language, 45(2), 225–258.

Notes

[1] The average number of characters was as follows: For the verb in the verb + PRO region it was gender-known, M = 7, SE = 1, gender-neutral, M = 6, SE = 1, high-expectancy, M = 8, SE = 1. For the adverb region it was gender-known, M = 8, SE = 1, low-expectancy, M = 8, SE = 1, high-expectancy, M = 9, SE = 1. For the final region it was gender-known, M = 20, SE = 2, low-expectancy, M = 22, SE = 3, high-expectancy, M = 22, SE = 2.

[2] Expected reading time was calculated based on the intercept and slope of participants’ by-character reading times. Separate intercepts and slopes were calculated for the different ROIs and reading measures.

Appendix

Table A1. List of experimental antecedents by condition with their average gender ratings and standard deviation

Noun	M	SD	Condition	Noun	M	SD	Condition
Cashier	4.3	0.7	low-expectancy	Pilot	2.7	0.9	high-expectancy
Cyclist	3.5	0.6	low-expectancy	Primary school teacher	4.8	0.8	high-expectancy
Doctor	3.8	0.4	low-expectancy	Prostitute	6.0	0.7	high-expectancy
Jogger	4.0	0.2	low-expectancy	Sailor	2.3	0.8	high-expectancy
Lawyer	3.6	0.7	low-expectancy	Secretary	5.5	0.9	high-expectancy
Musician	3.9	0.4	low-expectancy	Supermodel	6.6	0.5	high-expectancy
Neighbour	4.0	0.2	low-expectancy	Actress	7.0	0.0	gender-known
News reporter	3.9	0.5	low-expectancy	Boy	1.0	0.0	gender-known
Pedestrian	4.0	0.4	low-expectancy	Boyfriend	1.0	0.0	gender-known
Sales assistant	4.4	0.6	low-expectancy	Businesswoman	7.0	0.0	gender-known
Social worker	4.5	0.7	low-expectancy	Cleaning lady	7.0	0.0	gender-known
Spectator	3.9	0.4	low-expectancy	Dinner lady	6.9	0.3	gender-known
Babysitter	5.5	0.7	high-expectancy	Dustman	1.5	0.7	gender-known
Beggar	3.2	0.9	high-expectancy	Fisherman	1.4	0.6	gender-known
Boxer	2.3	0.9	high-expectancy	Gentleman	1.1	0.4	gender-known
Cheerleader	6.0	0.7	high-expectancy	Lady	7.0	0.0	gender-known
Drummer	2.7	0.9	high-expectancy	Lady in waiting	7.0	0.0	gender-known
Explorer	2.7	0.9	high-expectancy	Lollipop lady	6.8	0.5	gender-known
Farmer	2.5	0.8	high-expectancy	Man	1.0	0.0	gender-known
Flight attendant	4.6	0.8	high-expectancy	Milkman	1.3	0.6	gender-known
Gardener	3.3	0.8	high-expectancy	Niece	7.0	0.0	gender-known
Golfer	3.0	0.7	high-expectancy	Nun	7.0	0.0	gender-known
Hairdresser	4.9	0.9	high-expectancy	Old man	1.0	0.0	gender-known
Housekeeper	5.6	0.8	high-expectancy	Pageboy	1.0	0.0	gender-known
Hunter	2.1	0.9	high-expectancy	Policeman	1.4	0.9	gender-known
Librarian	4.7	0.7	high-expectancy	Schoolboy	1.0	0.0	gender-known
Make-up artist	5.2	0.9	high-expectancy	Sportswoman	7.0	0.0	gender-known
Mechanic	2.5	0.6	high-expectancy	Teenage girl	7.0	0.0	gender-known
Miner	1.8	0.6	high-expectancy	Waitress	7.0	0.0	gender-known
Nurse	5.0	0.9	high-expectancy	Workman	1.4	0.7	gender-known