RESEARCH ARTICLE

Rapid prediction of verbs based on pronoun interpretation is modulated by individual differences in pronoun processing

Jeffrey J. Green

Department of Linguistics, Brigham Young University, Provo, UT, USA

ARTICLE HISTORY

Compiled June 26, 2023

The Version of Record of this manuscript has been published and is available in *Language, Cognition and Neuroscience* (2023), https://www.tandfonline.com/doi/full/10.1080/23273798.2023.2226266

CONTACT Jeffrey Green. Email: jjgreen@byu.edu

ABSTRACT

How quickly can pronoun interpretation affect the prediction of a following verb? Readers were presented with implicit causality contexts in which a specific pronoun and following verb were predictable. N400 and reaction time results indicated that predictable verbs were facilitated relative to unpredictable verbs when following predicted pronouns, suggesting that verbal predictions were rapidly updated based on pronoun interpretation. There was also some evidence for rapid updating of verb predictions after unexpected pronouns, but this was modulated by individual differences. Some readers appear to have placed higher weight on top–down implicit causality predictions to interpret unexpected pronouns, and others on bottom–up information from the gender on the pronoun. These differences in turn affected the N400 response to expected and unexpected verbs. The results together demonstrate that pronouns can be interpreted quickly enough to affect predictions about the following word.

KEYWORDS

Prediction; pronoun resolution; N400; implicit causality; individual differences

1. Introduction

Beyond the integration of bottom–up information, sentence comprehension is affected by predictions readers and listeners make about upcoming lexical items (DeLong et al., 2014; Federmeier & Kutas, 1999; Frade et al., 2022; Lau et al., 2016; Sun, 2022), syntactic structure (Arai & Keller, 2013; Gibson, 2006; Green et al., 2020; Kazanina et al., 2007), reference (Grüter et al., 2018; Kehler & Rohde, 2013; Kush & Dillon, 2021; Reuter et al., 2020; Sturt, 2013), coherence relations (Cozijn et al., 2011; Kehler & Rohde, 2013; Kehler et al., 2008), and so on. One contextual property that reliably impacts readers' expectations is implicit causality. In these contexts, readers and listeners make predictions about likely next referents based on prior context and expected coherence relations. For example, readers expect Draco to be mentioned as the subject of the subsequent clause in the sentence "Harry hated Draco because…". A large body of research has demonstrated that when implicit-causality predictions are not met, processing difficulty ensues (e.g. Van Berkum et al., 2007). However, what is not clear is what cognitive processes underlie that difficulty, and how subsequent predictions are affected.

Within the research on linguistic prediction more generally, a question of increasing interest is about the speed of the chain-reaction effect on later predictions when additional information is encountered that affects word likelihood. Although many studies have shown that predictions can be updated quite rapidly (e.g. Chow & Chen, 2020), several recent studies have highlighted contexts where this is not the case (e.g. Chow et al., 2018; Liao & Lau, 2020). In addition, individuals may differ in how much they rely on top–down predictions versus bottom–up information during sentence processing (Farmer et al., 2005; Hersch & Andrews, 2012; Kaan, 2014).

The current study adds to the literature in these areas by investigating how quickly readers are able to update lexical predictions about verbs following pronouns that are either predicted or not predicted based on implicit causality, and whether that is affected by individual differences in the cognitive processing of unexpected pronouns.

1.1. Implicit causality and pronoun interpretation

The role of implicit-causality predictions in reference resolution has been well documented (e.g. Cozijn et al., 2011; Kehler et al., 2008). The focus of attention on specific event participants in comprehenders' mental models allows them to predict who or what is likely to be mentioned next in the upcoming discourse (Grüter et al., 2018). For example, readers expect a sentence such as (1) to continue with some statement of what Ron did to make Hermione angry, such as (1a). When this prediction is not met, as is the case in (1b), which talks about Hermione instead of Ron, readers exhibit reading slowdowns (Koornneef & Van Berkum, 2006; Stewart et al., 2000) and show evidence of increased processing difficulty in their event-related potential (ERP) brain responses (Van Berkum et al., 2007). This demonstrates that readers use implicit causality to predict upcoming referents, and potentially even specific pronouns.

(1) Hermione got angry with Ron when...

a. he spilled pumpkin juice on her homework.

b. she spilled pumpkin juice on her homework.

Van Berkum et al. (2007) found that participants exhibited a P600 response to

pronouns that violated the gender of the referent expected under implicit causality. The P600 is a posterior positivity that appears roughly 500–1000 ms after the onset of a word and is thought to index revision/repair strategies initiated after a perceived morphosyntactic error or other violation (Kaan & Swaab, 2003; Kaan et al., 2000; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995). Van Berkum et al. argued that the P600 seen in their study was due to a perceived morphosyntactic error, presumably because participants prioritised their implicit causality predictions over the given gender on the pronoun.

However, this is not the only potential source of processing difficulty for pronouns with unexpected gender, nor does a P600 in the grand average data mean that each participant showed the same brain response to unexpected pronouns. Tanner and Van Hell (2014) found that in response to morphosyntactic errors, the majority of participants exhibited an enhanced P600 response, but some instead showed an enhanced negativity to the errors. They argue that this reflects individual differences in the cognitive approach to dealing with perceived errors. The P600 seen in Van Berkum et al.'s grand averaged data may thus have obscured individual differences in the processing of unexpected pronouns.

Indeed, Canal et al. (2015) found that in response to pronouns violating the stereotypical gender of their antecedents, some participants showed a P600, in line with Van Berkum et al.'s results, but others showed an Nref effect, a sustained anterior negativity that is sensitive to referential ambiguity (Van Berkum, 2009). This suggests that some participants did not view the unexpected gender on the pronoun as a morphosyntactic violation, but simply had more difficulty finding a referent for the pronoun. Similarly, Nieuwland and Van Berkum (2008) found that in response to full noun phrases with ambiguous reference, some participants showed a strong Nref response, but others exhibited a P600. There is therefore also more than one possible reason for pronouns violating implicit causality to cause processing difficulty.

Although the previous literature on implicit causality has demonstrated that individuals consistently make referential predictions in implicit causality contexts and that there are strong processing costs for violations of those predictions, it remains unclear what the cognitive source of that cost is, as well as how that may differ across individu-

als. This is especially important given recent findings that individuals may differ in the extent to which they rely on top–down predictions versus bottom–up information in implicit causality contexts (Johnson & Arnold, 2021) as well as in sentence processing more generally (Farmer et al., 2005; Hersch & Andrews, 2012; Kaan, 2014).

1.2. Lexical prediction

In addition to referential predictions, a large body of research has demonstrated that readers and listeners actively predict upcoming lexical items in a probabilistic fashion (e.g. Frade et al., 2022; Lau et al., 2016). This prediction may involve the preactivation of words in the lexicon that can spread to not only the predicted word, but also words semantically related to it (Federmeier & Kutas, 1999). Predictability is often measured using the N400 ERP component. This component is a negative-going wave that usually peaks 300–500 ms after the presentation of a word (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984), and the amplitude of the N400 is inversely related to the predictability of a word in context (Lau et al., 2008; Thornhill & Van Petten, 2012).

One question that has received recent attention in the prediction literature is how quickly we are able to update our predictions based on ongoing linguistic input, and what kinds of information lead to rapid updating of our predictions. Prediction in sentence processing can sometimes be quite rapid. In one recent study, Szewczyk and Wodniecka (2020) demonstrated using N400 amplitude that readers can immediately update their predictions about nouns in Polish based on the gender of a preceding adjective, regardless of the plausibility of that noun in the context. In another study using visual-world eyetracking, Chow and Chen (2020) demonstrated that when Mandarin speakers hear a classifier that is incompatible with a predicted noun, they are able to rapidly update their predictions based on the new information. However, as the authors acknowledge, because this study presented only a small set of images on the screen, the potential candidates for upcoming nouns was severely limited, and it is unclear whether listeners would be as quick to update their predictions in a less constrained context.

Other studies have shown that prediction based on other types of information such as argument structure is not always fast (Chow et al., 2016). For example, Chow

et al. (2018) demonstrate that individuals' are unable to use argument-role information to make verb predictions (indexed by N400 amplitude) at short latencies (600 ms), but they are able to at longer a longer latency (1800 ms; see also Liao et al., 2022). Similarly, Liao and Lau (2020) examined how quickly readers were able to update predictions about the direct object of the verb based on the computation of complex verb-argument relations, again using N400 amplitude to index predictability. Their results demonstrated that participants were slower to use information from resultative constructions than from coordinate structures to update their predictions. However, given sufficient time, even verb-argument information from resultatives affected prediction.

These results indicate that some types of linguistic information can lead to rapid updating of predictions, while other types take more time to have an effect. However, further research is needed to understand why this is the case and exactly what types of information can rapidly affect predictions. The current study contributes to our understanding by examining whether the interpretation of predictable and unpredictable pronouns in implicit causality contexts has immediate consequences for verb prediction. On the one hand, successful implicit causality predictions may make further prediction easier. In (1), for example, not only will the reader predict reference to Ron, but also something that Ron did to upset Hermione. This may make them predict Ron as an agent as opposed to Ron as an experiencer or patient. If this argument role is part of the prediction, then readers may be more likely to be able to use the information from the expected pronoun *he* to predict likely verbs (see Garrod et al., 1994, for a similar claim). On the other hand, a richer prediction of pronoun and argument role information based on implicit causality may make updating predictions following pronouns that violate that rich prediction even more difficult. This may also depend on how different individuals treat the unexpected pronoun. If participants strongly rely on top–down implicit causality predictions to resolve the pronoun, they may make similar predictions about the verb regardless of the gender of the pronoun. If, on the other hand, they are less likely to rely on prediction to resolve reference, they may in turn be less likely to make predictions about the verb given an already unexpected pronoun.

1.3. The current study

The current study addresses two main research questions that are left unanswered by previous literature: (i) how quickly are individuals able to update verb predictions based on pronoun interpretation in implicit causality contexts, and (ii) what effect do individual differences in the processing of pronouns that are inconsistent with implicit causality predictions have on the prediction and processing of the following verb.

In a 2×2 design, participants were presented with mini stories such as (2), which contained either a predictable or unpredictable pronoun based on the context; the following verb was also either predictable or unpredictable based on the expected pronoun.

(2) The street Juliet had to cross to get to school was dangerous and busy; however with the help of Isaiah, Juliet was able to get to the other side of the street. Crossing Guard Isaiah assisted Preschooler Juliet in front of the school while...

a. **expected pronoun, expected verb**
she crossed the busy intersection.

b. **expected pronoun, unexpected verb**
she stopped traffic for everyone to cross.

c. **unexpected pronoun, expected verb**
he crossed the busy intersection.

d. **unexpected pronoun, unexpected verb**
he stopped traffic for everyone to cross.

The effects of reading the expected and unexpected pronouns and verbs were measured in two experiments. Experiment 1 measured reaction times in a maze task. This experiment addresses the first research question by measuring reaction times at the expected and unexpected verbs as modulated by which pronoun preceded them. However, the task is not sensitive enough to examine individual differences in how the unexpected pronouns may be processed. Experiment 2 addresses both research question by measuring ERP effects to the pronouns and verbs.

If participants successfully make implicit causality predictions, they should show

evidence of processing difficulty in response to the unexpected pronouns in both experiments. The type of difficulty seen may differ across individuals, which would lead to differences in the ERP response to the unexpected pronoun seen in Experiment 2. Furthermore, if readers are able to rapidly update their verb predictions based on pronoun interpretation, the expected verb should be facilitated following the expected pronoun, but not following the unexpected pronoun. However, if there are individual differences in the response to unexpected pronouns, then this may also affect the prediction and processing of verbs following unexpected pronouns.

2. Experiment 1

Experiment 1 had two main purposes: to replicate the well-established finding that reading a pronoun whose gender contrasts with readers' implicit causality predictions causes slowdowns, and to determine what effect expected and unexpected pronouns have on the reading times for following expected and unexpected verbs.

Reaction times were measured using a maze task. This task is similar to self-paced reading, in which participants press a button to see each word flash on the screen one at a time. It differs in that participants are shown two words at a time: a target and a competitor. Participants are asked to choose which of the two words continues the sentence. Although this is not as natural a task as self-paced reading, previous studies have shown that the maze task elicits similar effects as self-paced reading, and in some cases may even be more effective. Boyce et al. (2020) directly compared self-paced reading to maze, and found that the maze task was much more sensitive than self-paced reading in detecting processing difficulties resulting from structural ambiguity resolution. The effects were also more localised to the source of that difficulty, since self-paced reading tends to have strong spill-over effects. In addition, the maze task is less vulnerable to inattentive participants, since attention can be monitored based on accuracy at each word pair, which is ideal for online data collection (the experiment was completed during the COVID-19 shutdown when in-person data collection was impossible).

If participants make implicit causality predictions when reading these items and

then successfully interpret the pronoun, they should show a slowdown at unexpected versus expected pronouns (Koornneef & Van Berkum, 2006; Stewart et al., 2000). We also expect a slowdown for the unexpected verb, but if processing the pronouns has an immediate effect on either the predictions readers make or their processing of the following verb in general, the slowdown for unexpected verbs may be reduced or eliminated following unexpected pronouns.

2.1. Participants

Participants were recruited using Amazon Mechanical Turk and gave informed consent. In order to encourage careful completion of the task, participants were told in the recruitment form that their work would be rejected if it contained too many mistakes. Data was collected from a total of 50 participants. Of these, 11 participants with accuracy below 60%, or below 65% with more than 50% of their reaction times below 200 ms, were rejected and did not receive payment. The remaining 39 participants were paid $8 for their time. The data from 7 additional participants whose accuracy was below 80% was also excluded from analysis. This resulted in a final data set from 32 adult native English speaking participants (12 female, 14 male, 6 unknown; mean age = 42.5, SD = 13.8; demographic information from 6 participants was not collected due to a coding error).

The experiment protocol was approved by the Institutional Review Board at the University of Illinois at Urbana-Champaign.

2.2. Materials

Participants viewed stimuli such as those in Table 1. Each item had a preamble setting up an initial context. This was followed by a final sentence, presented in the maze format, including an implicit causality verb that would lead to prediction of reference to one of two characters using a specific pronoun (either *he* or *she*, based on the stereotypical gender associated with the characters' names).[1] Stimuli were counterbalanced with respect to whether implicit causality biased toward reference to the character corresponding to the subject or object of the main clause. The verb following the predicted

pronoun was also either predictable or unpredictable based on the preamble/sentence context and the expected pronoun. The predictability of the expected pronoun and verb were normed in a separate sentence-completion task. The expected pronoun had a cloze probability of 0.70 ± 0.13 (mean ± SD); the unexpected pronoun had a cloze probability of 0.10 ± 0.09. Crucially, the expected verb had a higher cloze probability than the unexpected verb only after the expected pronoun (0.34 ± 0.20 vs. 0.001 ± 0.007). After the unexpected pronoun, the difference in cloze probability between the expected and unexpected verbs was greatly reduced (0.04 ± 0.07 vs. 0.02 ± 0.07). The unexpected verb in a given item set was always an expected verb in a different item set. The items were distributed across 4 lists in a Latin square design such that each participant saw one version of each item set, and no participant saw the same verb for more than one item. Each list contained 36 critical items, as well as 54 fillers. The full set of critical items can be found in the online supplementary material.

Competitor words for the maze task were generated using A-maze (Boyce et al., 2020). This program uses a language model (Gulordava et al., 2018) to create distractor words that are generally ungrammatical in the context, but are at least a poor fit. Competitor words were matched with target words in length and overall log frequency. The same competitors were used for all conditions that had the same verb continuations.

2.3. Procedure

Data was collected using Ibex Farm (Alex Drummond: http://spellout.net/ibexfarm). After giving consent and reading instructions, participants were given three practice items, followed by the main experiment. For each item, the entire preamble appeared on the screen first. After participants read the preamble, they pressed a key to move onto the rest of the sentence, which was presented in the maze format. Target–distractor pairs appeared on the screen one at a time, with location of the target and competitor on the left or right randomised for each pair (see Figure 1). Participants used the 'e' and 'i' keys to select which word continued the sentence. In order for participants to successfully complete the task, they needed to identify each candidate word, determine whether or how easily each candidate fit the context, decide which candidate was

Table 1. Sample item set for Experiment 1. The critical pronoun and verb are in bold. Competitor words are in parentheses.

Preamble		
Known for being able to save even the riskiest pregnancies,		
Pre-critical region		
Midwife Crystal amazed Nervous Parent Jackson last night when		
(x-x-x) (Screwed) (theirs) (Custody) (Urgent) (Painful) (sale) (earth) (jack)		
Critical region		
Pron.	**Verb**	**Target clause**
exp.	exp.	**she** **delivered** a healthy baby girl.
		(seat) (depending) (sir) (granted) (gods) (laws.)
exp.	unexp.	**she** **answered** the phone call about his wife.
		(seat) (timeline) (gone) (shall) (cool) (ha) (grow) (exist.)
unexp.	exp.	**he** **delivered** a healthy baby girl.
		(seat) (depending) (sir) (granted) (gods) (laws.)
unexp.	unexp.	**he** **answered** the phone call about his wife.
		(seat) (timeline) (gone) (shall) (cool) (ha) (grow) (exist.)

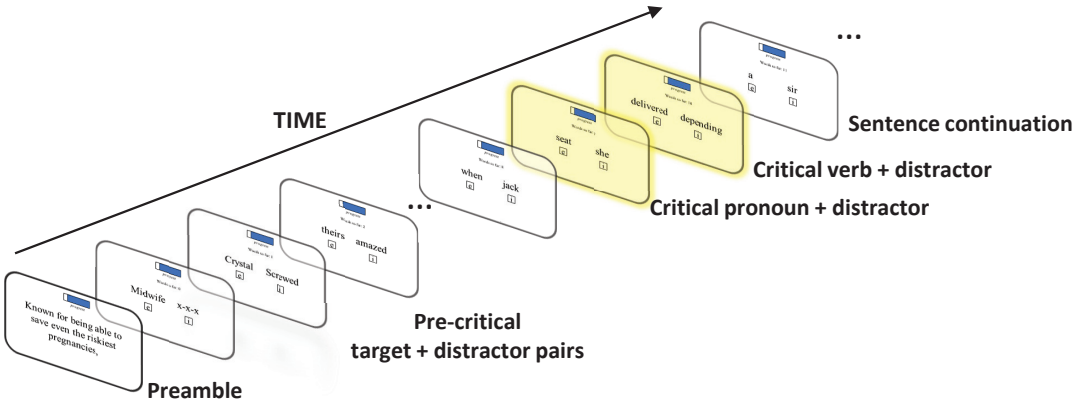

Figure 1. Stimulus presentation in Experiment 1.

correct, initiate or complete motor actions to press the corresponding key, and complete the integration of the chosen candidate into the growing context so that processing of the next candidate pair could continue.

If the correct word out of a pair was chosen, the next pair appeared. If an incorrect word was chosen, the participants saw "Incorrect! Please try again." written above the two words, and they were able to try again. When participants reached the end of the sentence, the next sentence preamble automatically appeared. In order to introduce a slight gamification of the task, a running total of correct words selected was displayed at the top of the screen throughout the experiment.

2.4. Data processing and analysis

Data analysis was completed in R (R Core Team, 2021; RStudio Team, 2021). For each word pair, the reaction time for participants to make a correct response was measured. Trials where the participant's initial response was incorrect were removed from analysis, resulting in a loss of 3.4% of the data.

Log-transformed RTs were analyzed in three different regions: at the critical pronoun, at the following verb, and at the word following the verb in order to account for possible spillover effects. Only one spill-over region was analyzed because the study was specifically investigating verb prediction. Haeuser and Kray (2022) found that predictability effects in self-paced reading appear directly on the predictable/unpredictable word, and that plausibility effects appear later, and Boyce et al. (2020) demonstrated

that effects are generally less delayed in the maze task. Husband (2022) also demonstrates that the maze task is immediately sensitive to word predictability.

For each region of interest, a linear mixed effects model was computed using the lmerTest package (Kuznetsova et al., 2017), which uses Satterwaithe approximations to calculate degrees of freedom. Fixed effects for all models included pronoun predictability (expected/unexpected), and for the verb and following region, the models also included fixed effects for verb predictability (expected/unexpected), and its interaction with pronoun predictability. Effects were sum coded using the car package (Fox & Weisberg, 2019). In addition, participants and items were crossed, starting with random intercepts and slopes, and removing one level of complexity until the model converged without singular fit, following the recommendations of Baayen et al. (2008) and Barr et al. (2013). The final model for the pronoun region was `logRT ~ Pronoun + (1 + Pronoun|Participant) + (1|Item)`, and for the verb and spillover regions, it was `logRT ~ Pronoun*Verb + (1 + Pronoun + Verb|Participant) + (1 + Pronoun + Verb|Item)`. Tukey-adjusted pairwise follow-up comparisons were computed for significant interactions using the emmeans package (Lenth, 2021). Main effects are reported only when interactions including those effects were not significant.

The full data and statistical analysis are available in the online supplementary material.

2.5. Results

Reading times for the regions of interest across the four conditions are given in Figure 2.

At the pronoun, there was an effect of pronoun predictability, $t = -4.09$, $p<0.001$, with unexpected pronouns read significantly more slowly than expected pronouns (mean±SD = 841±311 vs. 786±285).

At the verb, there was an interaction of pronoun and verb predictability, $t = -4.61$, $p<0.001$. Follow-up comparisons indicated that unexpected verbs were read more slowly than expected verbs after expected pronouns (mean±SD = 1198±430 vs. 994±346), t.ratio = -4.90, $p<0.001$, but not after unexpected pronouns (mean±SD = 1130±417 vs. 1106±432), t.ratio = -1.00, $p>0.1$).

The region following the verb (verb+1) showed the same effects as at the verb,

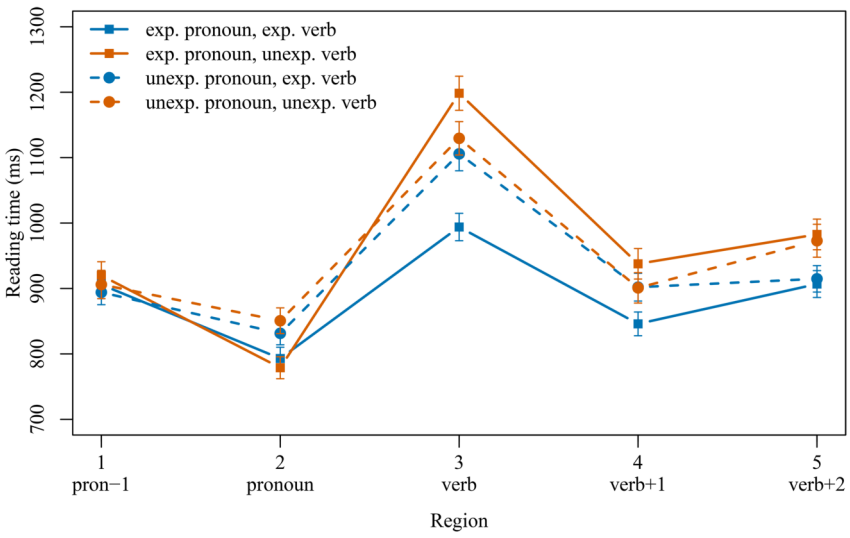

Figure 2. Mean reading times in Experiment 1. Error bars represent the standard error of the mean.

with a significant interaction between pronoun and verb predictability, t = -3.15, p=0.002. Follow-up comparisons again revealed an effect of verb predictability only following expected pronouns, with unexpected verbs causing a sustained slowdown compared to expected verbs (mean±SD = 938±390 vs. 846±299), t.ratio = -2.24, p = 0.030; there was again no effect of verb predictability following unexpected pronouns (mean±SD = 901±385 vs. 902±352), t.ratio = 0.13, p>0.1.

2.6. Discussion

This experiment investigated how quickly verb predictions can be updated based on pronoun resolution during sentence comprehension. If participants were able to quickly update their predictions based on the interpretation of the pronoun, then we should see a stronger effect of the unexpected verb when it followed an expected pronoun than when it followed an unexpected pronoun. The results for Experiment 1 aligned with this prediction. Reaction times to unexpected pronouns were significantly slower than to expected pronouns. At the verb, there was an interaction between pronoun and verb type; unexpected verbs resulted in slowdowns, but only when following the expected pronoun. After the unexpected pronoun, there was no significant difference in reaction times to unexpected and expected verbs.

2.6.1. Effect of unexpected pronouns

The slowdown seen at the unexpected pronoun was not surprising. The contexts of the sentences led participants to predict that a specific character would be referred to based on implicit causality using a specific pronoun—either *he* or *she*. When participants instead encountered a pronoun of the opposite gender, RTs were slowed. Although this effect was expected, RT data alone does not tell us the cognitive difficulty behind the slowdown.

There are at least four potential reasons participants may have slowed down at unexpected pronouns. First, participants were predicting that one of the characters would be mentioned next. When they read the unexpected pronoun, they may have interpreted it as referring to the other character, in violation of their implicit causality predictions. The effort required to revise their predicted situation model and to retrieve that unpredicted character from memory could have caused a slowdown. Second, participants may not have abandoned their implicit causality predictions at the unexpected pronoun, and instead may have assumed that the pronoun contained a gender agreement error or typo, but that it was still being used to refer to the predicted character. Processing this perceived agreement error would also cause a slowdown. Third, it is possible that upon reading the pronoun, participants did not treat it as a gender agreement error, but still did not want to abandon their implicit causality predictions. The slowdown may have come from participants reconsidering what pronouns the expected character prefers. Finally, participants may have simply been unsure who the pronoun was intended to refer to, resulting in a slowdown due to the effort needed to resolve the unclear reference. Experiment 2 examines these possibilities by measuring ERP responses to unexpected pronouns in these types of items.

2.6.2. Effect of unexpected verbs

As for whether reading the pronoun affected the processing of the following verb, the results reveal that the unexpected verb caused a slowdown in reading times only after the expected verb. When participants had just read an unexpected pronoun, there was no difference in RTs to the expected vs. unexpected verb. This pattern continued into the following region.

One possible explanation of these effects is that participants were able to rapidly update their predictions about the verb based on their interpretation of the pronoun. Reading an expected pronoun led them to predict or strengthen their prediction of the expected verb, which was then facilitated, leading to faster RTs relative to unpredicted verbs. On the other hand, when the pronoun was not the one expected, participants may have not predicted the expected verb at all—or at least not strengthened any prior predictions they may have had about the verb—based on the unexpected pronoun. They therefore exhibited similar RTs to both verbs, both being equally unexpected.

One potential alternative to the prediction explanation is that the RTs were not reflective of prediction, but of ease of integration into the semantic context. After the expected pronoun, the expected verb fit the semantic context well and was easy to integrate, leading to faster RTs than for the unexpected verb. After the unexpected pronoun, on the other hand, there may have been difficulty integrating either verb into the semantic context. Although we attempted to make the unexpected verb and sentence continuation plausible following the unexpected pronoun, this sometimes was not the case until the end of the sentence. Participants therefore may still have found it harder to integrate than the expected verb following the expected pronoun.

However, although in the majority of our items the expected and unexpected verbs had equally low cloze probability following the unexpected pronoun, for five of the items, the "unexpected" verb actually had higher cloze probability after the unexpected pronoun than the "expected" verb. An exploratory analysis of these five items (see the supplementary materials) suggests that participants may indeed have been making predictions about the upcoming verb based on even the unexpected pronoun's interpretation, lending support to the prediction interpretation of the results. Given that there were so few items that had this property, though, this support is tentative and simply presents a hypothesis for future research.

Finally, the effects here may be due to a combination of factors. Garrod et al. (1994) used a similar design manipulating pronoun and following verb plausibility in an eyetracking-while-reading study. They argued for a constraint satisfaction model of sentence processing in which readers take pronoun gender, verb bias, and discourse focus immediately into account when interpreting the verb following a pronoun. In the

current experiment, it is possible that predictions about both the pronoun and verb, as well as integration difficulty, all contributed to the effects seen.

Although the source of the results in Experiment 1 is not completely clear, the findings demonstrate that reading a pronoun that is consistent or inconsistent with implicit causality predictions has an immediate effect on the processing of a following expected or unexpected verb. Experiment 2 uses EEG to examine possible sources for that immediate effect.

3. Experiment 2

Experiment 2 addresses two main questions: (i) whether readers can indeed make rapid predictions about the following verb based on pronoun interpretation, and (ii) what effect individual differences in the processing of unexpected pronouns (if they exist) have on the prediction and processing of the following verb. These questions were examined by presenting participants with the same type of sentences as used in Experiment 1 while recording their brain responses using EEG.

Regarding the first question, if the facilitation of the expected verb following the expected pronoun in Experiment 1 was due to participants rapidly making predictions about the upcoming verb based on the interpretation of the pronoun, then we should expect a reduced N400 to the expected versus unexpected verb following an expected pronoun. However, predictions cannot always be updated quickly (Chow et al., 2016, 2018; Liao & Lau, 2020). Because of this, the Experiment 1 results may also have been due to other aspects of processing such as integration. If this is the case in this experiment, then we would either not expect a reduced N400 at the expected verb following the expected pronoun, or we might expect the same facilitation at the expected verb regardless of pronoun. The latter would be predicted if people not only begin making predictions about the pronoun but also the verb prior to reading the pronoun.

As for the second question, previous studies have found three kinds of brain responses to pronouns whose gender violates participants' likely predictions or to other forms of reference with similar violations. First, Van Berkum et al. (2007) found a

P600 response to pronouns violating implicit causality predictions and argued that readers assumed the pronoun contained a morphosyntactic error. A P600 may also be elicited if participants initiate revision or repair strategies either in their representation of the gender of the characters, or in their initial interpretation of the pronoun (Kaan & Swaab, 2003; Kolk & Chwilla, 2007; Nieuwland & Van Berkum, 2008). In this experiment, if participants give higher weight to their top–down implicit causality predictions, they may show a similar response, and treat the unexpected gender on the pronoun as an error, leading to a similar P600 effect at the unexpected pronoun. On the other hand, if participants give more weight to the bottom–up gender information on the pronoun and treat the pronoun as referring to the unexpected character, they may exhibit an N400 due to the unexpected reference (Almor et al., 2017). A final possibility is that participants may simply not know how to interpret the pronoun, resulting in an Nref effect (Canal et al., 2015).

Finally, how people treat the unexpected pronouns may in turn affect their expectations and processing of the following verb. If participants give higher weight to the bottom–up gender information on the pronoun and assume it refers to the unexpected character, resulting in an N400, or if they are unsure who the pronoun refers to and exhibit an Nref, then they would have no reason to predict the "expected" verb, and we would expect no difference in N400 at the verb following unexpected pronouns. If, on the other hand, participants give higher weight to their top–down predictions and assume the pronoun has a morphosyntactic error, they may still have reason to make the same verb predictions regardless of which pronoun they read. This would lead to a similar reduction of the N400 for the expected verb following both expected and unexpected pronouns. However, it may also be the case that dealing with the perceived morphosyntactic error is enough to disrupt or inhibit the prediction process, in which case the reduced N400 to expected verbs may still be stronger following expected pronouns.

3.1. Participants

Eighteen native speakers of English (mean age: 22.4; SD: 2.3; 14 female, 12 male) were recruited at the University of Illinois at Urbana-Champaign (UIUC), and re-

ceived $20–30 or class credit to participate in this study. An additional 18 participants (mean age: 22.4; SD: 5.6; 16 female, 2 male) were recruited at Brigham Young University (BYU) and received $24–30 for their participation. All 36 participants were right handed and had normal or corrected-to-normal vision. All participants gave informed consent, and the experiment protocol was approved by the Institutional Review Boards at both UIUC and BYU.

3.2. Materials

Materials for Experiment 2 included the 36 item sets from Experiment 1 (without the competitor words), as well as 60 additional item sets of similar design. For each item, participants were presented with a preamble setting up the initial context, followed by the target sentence, which included an implicit causality verb leading to the prediction of reference to one of the characters using a specific pronoun. Items were again normed in an online cloze study on Ibex Farm. For the critical items in Experiment 2, the cloze probabilities of the expected and unexpected pronouns were 0.69 ± 0.13 and 0.11 ± 0.09, respectively. After the expected pronoun, the cloze probability of the expected verb was 0.25 ± 0.18, and the unexpected verb 0.002 ± 0.01. Following the unexpected pronoun, the difference between the two verbs was greatly reduced: the expected verb had a cloze probability of 0.03 ± 0.05, and the unexpected verb 0.01 ± 0.04.

The 96 critical item sets were distributed across four lists in a Latin square design, so that each participant saw a total of 24 items in each condition. These were combined with 144 fillers, for a total of 240 items per list. One third of the items were followed by a comprehension question in order to monitor attention. This included 15 questions that specifically asked about the referent of the target pronoun in a given critical item. This was done in order to measure whether participants tended to interpret the pronoun according to the intended gender of the characters in the story.

The full set of critical items can be found in the online supplementary material.

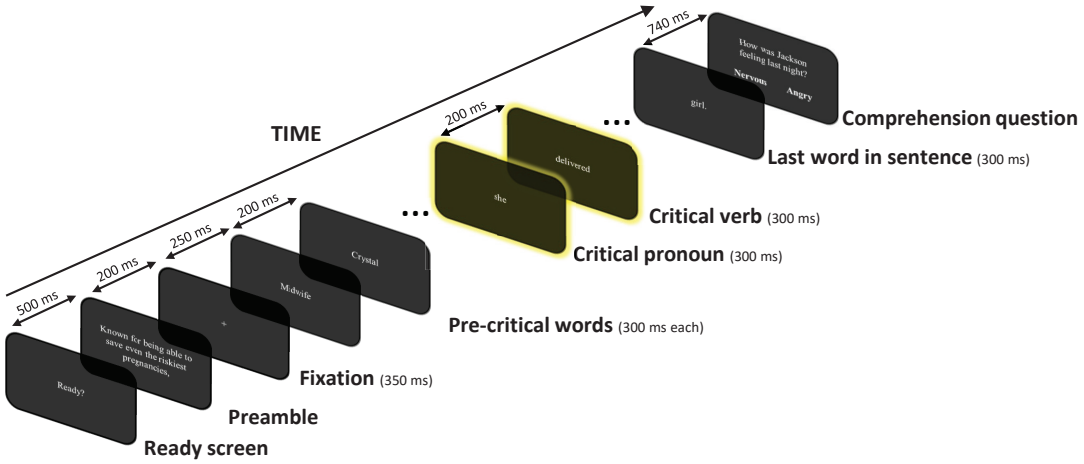

Figure 3. Stimulus presentation in Experiment 2.

3.3. Procedure

Participants were fitted with an EEG cap and sat in front of a computer screen. Stimuli were presented on a black screen in white 36 pt font using Paradigm (Perception Research Systems, 2007). Each item began with a ready screen, and participants pressed a button on a response pad to continue. After a 500 ms blank screen, the preamble appeared on the screen. After reading the preamble, participants pressed a key to continue. This was followed by a 200 ms blank screen, a fixation cross in the middle of the screen for 350 ms, and another 250 ms blank screen. The rest of the sentence was then presented one word at a time in rapid serial visual presentation (RSVP) format. Each word appeared for 300 ms in the middle of the screen, followed by a 200 ms blank screen, for a total SOA of 500 ms (see Figure 3). The final word of the sentence was followed by a 740 ms blank screen. The comprehension question then appeared for a subset of trials, along with two potential answers. Participants responded by pressing a button on the response pad. The location of the correct response—right or left—was counterbalanced across items. After participants responded, there was a 240 ms blank screen, after which the ready screen for the following item appeared. The experiment began with four practice items. The 240 main items were split into six blocks of 40 items each, with a break between blocks.

3.4. Data acquisition, processing, and analysis

Continuous EEG was recorded from 28 passive tin scalp electrodes mounted on an elastic cap (Electro-Cap International) in an electromagnetically shielded booth at UIUC, and from 30 active Ag/AgCl electrodes mounted in an EasyCap in a non-shielded room at BYU. The configuration of the channels at the two sites was identical with the exception of the two additional channels built into the caps at BYU (FT9 and FT10). To combine the data, those two electrodes were removed from the BYU data, leaving identical cap configurations. Electrodes were placed in accordance with the extended 10-20 system (Jasper, 1958), with reference electrodes on both mastoids. At UIUC, eye movements and blinks were recorded through 3 electrodes: two at the outer canthus of each eye, and one below the left eye. At BYU, an additional electrode was placed above the left eye for bipolar recording of both the HEOG and VEOG. Although bipolar recording of the VEOG component at UIUC was not used, blink artifacts were still large enough in the data to be detected and corrected. The EEG was amplified at UIUC using a BrainAmp DC bioamplifier system, and at BYU using the actiCHamp Plus system (both from Brain Products Gmbh, Gilchin, Germany) and digitised with a sampling rate of 1000 Hz, an online 250 Hz low-pass filter, and a 0.016 Hz high-pass filter (10s time constant), referenced online to the right mastoid. Impedances were held at or below 5 kΩs for the passive electrodes at UIUC (one participant had one electrode at 6 kΩs), and below 20 kΩs for the active electrodes at BYU (one participant had a single electrode at 29 kΩs). Although the single electrodes for these two participants had higher impedance, they were subject to the same filtering and artifact rejection as the other electrodes.

Offline processing was conducted with the Brainvision Analyzer software (Version 2.2.1). A 0.1 to 30 Hz bandpass zero phase-shift Butterworth IIR filter (order 2) was applied to the continuous EEG. After filtering, large artifacts were removed semi-automatically based on a maximum gradient of 100 μV/ms, a maximum peak-to-peak threshold of 200 μV, an overall maximum voltage threshold of ±150 μV, and minimum activity threshold of 0.5 μV. These thresholds were adjusted for each individual based on visual inspection. Individual channels with an excessively large number of artifacts were topographically interpolated by spherical splines (order 4), with a maximum of

two interpolated channels per participant. Independent Component Analysis (ICA) was then used to isolate and remove artifacts related to eye movements and blinks. The data was then re-referenced to the average of both mastoids. The continuous EEG was then epoched from -200 to 1300 ms relative to the onset of the pronoun, with baseline correction based on the 200 ms prior to pronoun onset. This epoch was chosen in order to capture ERP responses to both the pronoun and verb. This was followed by another round of artifact rejection to remove epochs with smaller artifacts (max gradient: 75 μV/ms; peak-to-peak threshold: 150 μV; max amplitude: ±100 μV/ms; low activity threshold: 0.5 μV in a 50 ms moving window). There were very few artifacts in the data for most participants after ICA. The overall rejection rate across participants after all data cleaning was 2.1% (range: 0–15.6%; SD: 3.3%).

To then analyze the effects of pronoun and verb predictability, linear mixed-effects models were fit looking at ERP responses in the 300–500 ms, 500–800 ms, and 800–1000 ms windows after the onset of the pronoun. The first two time windows were selected to capture different potential responses to unexpected pronouns. An N400 would be expected to cause a posterior negativity in the 300–500 ms window, a P600 a posterior positivity in the 500–800 ms window, and an Nref an anterior negativity in both time windows. Since the verbs appeared on the screen 500 ms after the pronouns, the third time window represents the N400 window for the verb (300–500 ms after verb onset). The P600 window for the verb was not examined because the research questions focused only on whether verb predictions, reflected in the N400 response to the verb, are affected by individual differences in pronoun processing.

Fixed effects in the models for the 300–500 ms and 500–800 ms windows included pronoun predictability (sum coded) and anteriority (as an ordered factor with orthogonal polynomial contrasts, based on seven levels of anteriority: Fp, F, FC, C, CP, P, O). Polynomial coding was used in order to be able to test whether anteriority linearly affected the size of the effect of pronoun predictability (i.e. the size of the effect increased or decreased steadily moving from anterior to posterior regions), as well as whether there were quadratic (e.g. the size of the effect increases, then decreases), or more complex effects of anteriority (Brehm & Alday, 2022; Schad et al., 2020). Random intercepts by participants and random slopes for pronoun predictabil-

ity were also included. Hemisphere was not included as a fixed effect in the model because none of the hypothesised ERP responses differ across hemisphere. In addition, random effects for items were not included because the EEG was recorded with separate triggers for each condition, but not for individual items; the models therefore were based on mean amplitude for each participant. The model for these time windows was $\mu V \sim$ `Pronoun*Anteriority + (1 + Pronoun|Participant)`. In addition to the fixed effects in the earlier time windows, the model for the 800–1000 ms window included fixed effects for verb predictability (sum coded) and its interaction with pronoun predictability, with random intercepts by participant and random slopes for pronoun and verb predictability and their interaction. The model this region was $\mu V \sim$ `Pronoun*Verb*Anteriority + (1 + Pronoun*Verb|Participant)`. For both models, only results including the highest-order interaction terms with both cluster and pronoun/verb predictability are reported where significant. Otherwise lower-order terms were examined, as long as pronoun/verb predictability was included. Follow-up comparisons were again computed using the emmeans package.

Mean accuracy on the comprehension questions was also measured for all comprehension questions together, as well as separately for only those questions that had to do with final, offline interpretation of the pronoun. Because the items were created with the intent that the characters would be of a stereotypical gender, these questions were coded as "correct" if the interpretation of the pronoun reflected that stereotypical gender.

The full data and statistical analysis are available in the online supplementary material.

3.5. Results

3.5.1. ERP responses

Grand average ERP waveforms at the Fz, Cz, and Pz electrodes for each condition as well as topoplots showing the effect of pronoun predictability and verb predictability following expected and unexpected pronouns are given in Figure 4 (ERPs at all electrodes are given in the supplementary materials). Visual inspection of the data suggests

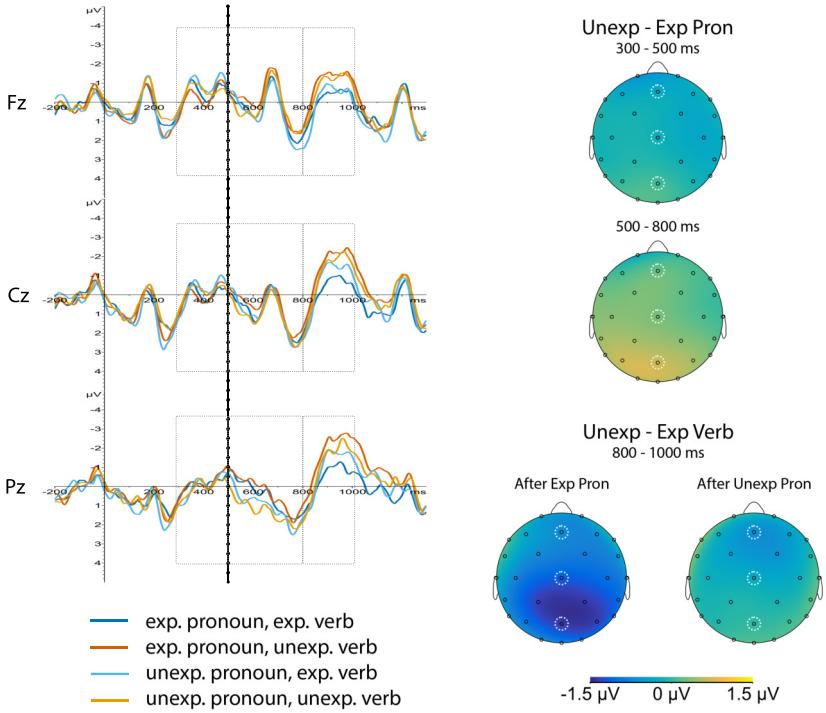

Figure 4. Left: Grand average ERPs for Experiment 2 at Fz, Cz, and Pz. Negative is plotted up. The dotted vertical line indicates the onset of the presentation of the verb. A 15 Hz low-pass filter was applied for plotting ERP waveforms only. Right: Topographic distribution of ERP effects for pronoun predictability in the 300–500 ms and 500–800 ms windows, and for verb predictability following expected and unexpected pronouns in the 800–1000 ms window post pronoun onset.

that relative to expected pronouns, unexpected pronouns elicited a large, broadly distributed positivity, with the strongest effect in posterior regions in the 500–800 ms window after the presentation of the pronoun (i.e. a P600 effect). Furthermore, unexpected verbs appear to have elicited a large posterior negativity relative to expected verbs 300–500 ms after the presentation of the verb (800–1000 ms after the pronoun; i.e. an N400 effect), but only after expected pronouns.

However, inspection of individual participants' responses to the unexpected pronoun revealed that a posterior positivity was only elicited for a subset of the participants. Others showed a strong negativity in response to the unexpected pronoun. In addition, in comments after the experiment was completed, a few participants made observations about the unexpected pronouns that indicated differences in how they were being processed. One, for example, stated that they thought the experiment must be investigating individuals' preferred and non-preferred gendered pronoun use. Another said they noticed some typos in the pronouns. In order to analyze the effect of indi-

vidual differences on pronoun and verb processing, we measured the mean amplitude difference for each participant between the expected and unexpected pronouns over anterior and posterior electrodes for the 300–500 and 500–800 ms windows after the onset of the pronoun, and used k-means clustering to identify groups of participants with similar neural responses. Using the silhouette method in the factoextra package in R (Kassambara & Mundt, 2020), it was determined that the optimal number of clusters was two. Cluster 1 included 15 participants, and Cluster 2 had 21 participants. Which cluster participants belonged to (sum coded) was then included as predictor in the analyses (pronoun regions: `μV ~ Cluster*Pronoun*Anteriority + (1 + Pronoun|Participant)`; verb region: `μV ~ Cluster*Pronoun*Verb*Anteriority + (1 + Pronoun*Verb|Participant)`; results of the models on the grand average data without taking into account individual differences can be found in the supplementary material).

ERP waveforms and topoplots for Cluster 1 are illustrated in Figure 5 and for Cluster 2 in Figure 6. In the 300–500 ms time window after the onset of the pronoun, there was a significant interaction between cluster and pronoun, t = -6.89, $p<0.001$. Follow-up comparisons indicated that for Cluster 1, unexpected pronouns elicited a negativity relative to expected pronouns, z.ratio = -6.11, $p<0.001$; for Cluster 2, unexpected pronouns elicited a positivity, z.ratio = 3.45, $p<0.001$. There was no interaction with anteriority in this time window, so the responses were broadly distributed.

In the 500–800 ms window, there was again a cluster by pronoun interaction, t = -5.75, $p<0.001$, as well as an interaction between cluster, pronoun, and the linear component of the anteriority variable, t = 2.28, p = 0.023. Follow-up comparisons to these interactions revealed that ERPs to the unexpected pronoun were again more negative than to the expected pronoun for Cluster 1, z.ratio = -3.13, p = 0.002, but this negativity was now strongest in anterior channels, z.ratio = 4.03, $p<0.001$. For Cluster 2, the unexpected pronoun produced a positivity, z.ratio = 5.21, $p<0.001$, again with no effect of anteriority, z.ratio = 1.24, $p>0.1$.

In the 800–1000 ms window (i.e. 300–500 ms after the presentation of the verb), there was a significant interaction between cluster, pronoun, verb, and linear anteriority, t = 3.52, $p<0.001$. Unexpected verbs led to a stronger negativity than expected verbs

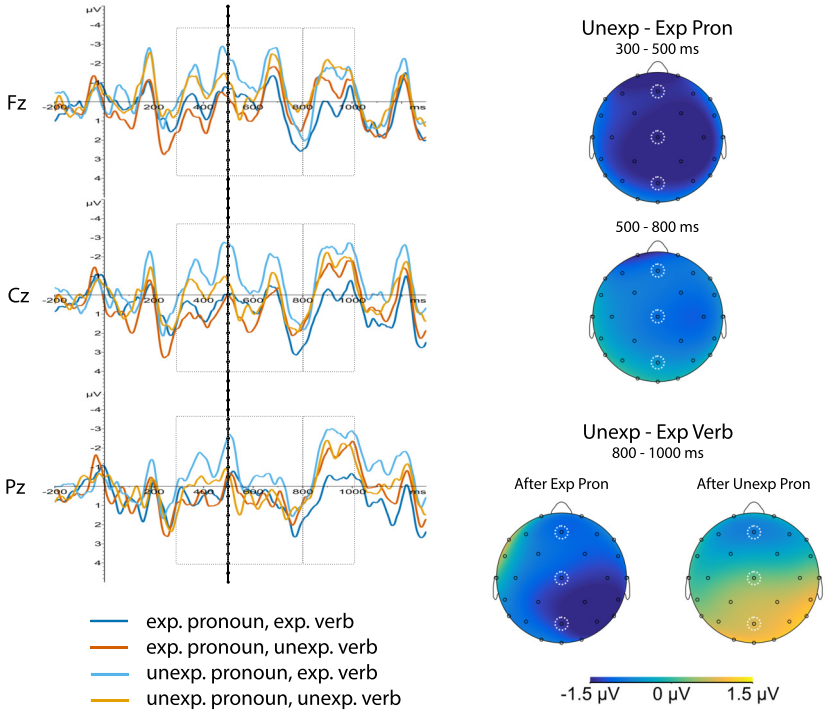

Figure 5. Results for Cluster 1 only. Left: ERPs for Experiment 2 at Fz, Cz, and Pz. Negative is plotted up. The dotted vertical line indicates the onset of the presentation of the verb. A 15 Hz low-pass filter was applied for plotting ERP waveforms only. Right: Topographic distribution of ERP effects for pronoun predictability in the 300–500 ms and 500–800 ms windows, and for verb predictability following expected and unexpected pronouns in the 800–1000 ms window post pronoun onset.

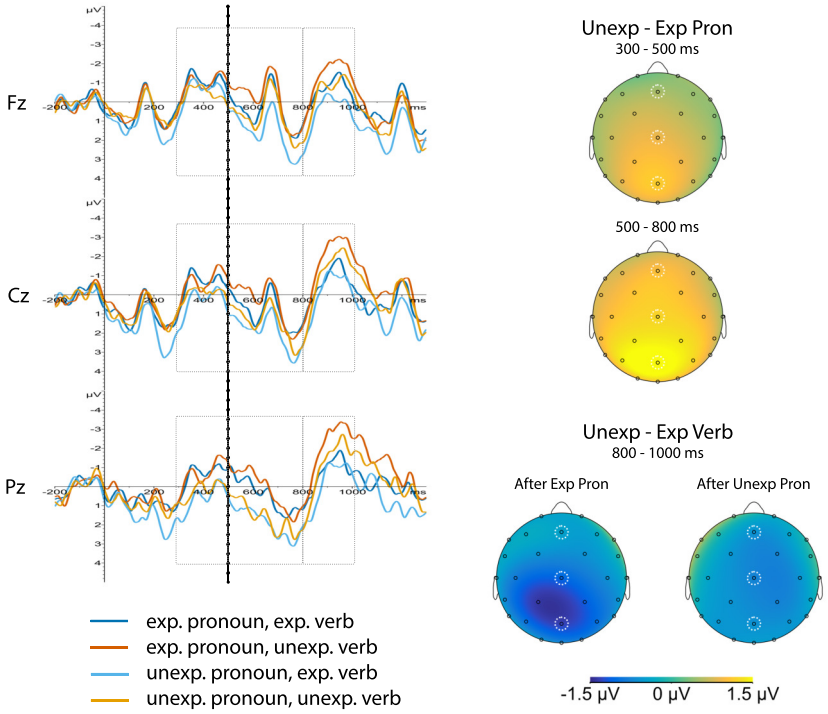

Figure 6. Results for Cluster 2 only. Left: ERPs for Experiment 2 at Fz, Cz, and Pz. Negative is plotted up. The dotted vertical line indicates the onset of the presentation of the verb. A 15 Hz low-pass filter was applied for plotting ERP waveforms only. Right: Topographic distribution of ERP effects for pronoun predictability in the 300–500 ms and 500–800 ms windows, and for verb predictability following expected and unexpected pronouns in the 800–1000 ms window post pronoun onset.

in posterior regions following expected pronouns for both Cluster 1, z.ratio = -2.52, p = 0.012, and Cluster 2, z.ratio = -2.12, p = 0.034. Cluster 2 also showed a significant, albeit weaker, posterior negativity to unexpected verbs following unexpected pronouns, t=-2.20, p = 0.028. By contrast, responses to unexpected verbs following unexpected pronouns in Cluster 1 were *less* negative in posterior regions compared to the expected verb, z.ratio = 3.93, p<0.001.

In sum, participants in Cluster 1 showed a negativity in response to the unexpected pronoun. This negativity was broadly distributed in the 300–500 ms window. In the 500–800 ms window, the effect was smaller, and strongest over anterior electrodes. Participants in Cluster 2, on the other hand, showed a broadly distributed positivity in response to the unexpected pronoun that was strongest in the 500–800 ms window. As for responses to verbs, Cluster 1 showed a posterior negativity to unexpected verbs following expected pronouns, but the reverse following unexpected pronouns, with expected verbs showing a stronger posterior negativity than unexpected verbs. Cluster 2 showed a greater posterior negativity to unexpected verbs than to expected verbs after both expected and unexpected pronouns, but the difference was stronger after expected pronouns.

3.5.2. Question accuracy

Total accuracy for all questions was 0.92±0.04 (mean±SD). For the 15 questions specifically asking about the referent of the pronoun, accuracy was 0.83±0.14. Accuracy between the two clusters was compared using independent-samples t-tests. There was no difference between clusters on overall accuracy (Cluster 1: 0.92±0.04; Cluster 2: 0.92±0.04), t = 0.42, p>0.1, or on questions asking about the referent of the pronoun (Cluster 1: 0.86±0.13; Cluster 2: 0.81±0.14), t = 0.96, p>0.1.

3.6. Discussion

Experiment 2 investigated two main questions: first, whether readers can rapidly update their predictions about upcoming verbs based on pronoun interpretation, and second, whether verb prediction is affected by potential individual differences in pro-

noun processing, as indexed by ERP responses to unexpected pronouns. The results provide evidence that readers can indeed rapidly update their verb predictions after reading an expected pronoun. The effect of reading an unexpected pronoun on verb predictions is modulated by individual differences in how that pronoun is processed.

It is worth noting that the number of participants and items in this experiment is slightly lower than what has been used in some other studies of individual differences in ERPs. For example, Tanner and Van Hell (2014) looked at ERP responses to verb tense and agreement violations and found that some individuals respond with an N400 and others a P600. Their experiment had a total of 42 participants, each of which saw 30 items per condition. The current study, on the other hand, had 36 participants who each saw 24 items per condition. The relatively small number of items is less of an issue when looking at responses to unexpected pronouns, since that analysis collapses across verb types, resulting in 48 items each of expected and unexpected pronouns, a total which exceeds the number of items per condition presented in Tanner and Van Hell (2014) as well as other previous work the current experiment builds on (e.g. Van Berkum et al., 2007). In addition, the total number of participants in the current experiment exceeds that seen in some other studies that have examined individual differences in ERP responses (Canal et al., 2015; Grey, 2022), and the smaller of the two clusters had almost 70% more participants than the smallest group in Tanner and Van Hell's experiment. Given the large number of items per pronoun type, the number of participants in the current study should be sufficient to demonstrate qualitative differences in how individuals responded to unexpected pronouns. As for the verb region, in which there were 24 items per condition, in a recent study, Kim et al. (2023) found that with 25 items, even as little as 10 participants was sufficient to reach over 85% power in examining N400 responses to semantic anomalies. Even though the N400 effect in the current study is smaller in magnitude than that seen in Kim et al.'s study, the number of items in the current study should give enough power to detect N400 effects even in the smaller of the two clusters. In short, even though the number of participants is slightly lower in the current study than in some research, the number of items in each condition in the pronoun regions is higher, and even in the verb region, which had fewer items per condition, the number of items should be enough to detect

the effects of interest with sufficient power.

3.6.1. Updating verb predictions based on pronoun interpretation

For the first research question, it was hypothesised that if verb predictions were based at least in part on reading the expected pronoun, there would be an N400 effect at the unexpected verb relative to the expected verb following the expected pronoun. This prediction was confirmed in both the grand average data and within both clusters of participants. When participants read a pronoun that was consistent with their implicit causality predictions, they subsequently showed an enhanced posterior negativity to unexpected verbs than to expected verbs, the expectation being driven by that pronoun's interpretation.

However, this on its own does not necessarily mean that they were rapidly updating their predictions based on the pronoun. If participants were predictively interpreting the expected pronoun before encountering it (see, e.g., Arnold, 2010; Rohde & Kehler, 2013), then they may have also been making predictions about the verb prior to encountering the pronoun, potentially leading to the N400 effect seen at the verb following expected pronouns. In offline cloze data, when participants were not given any pronoun, they completed the sentence beginning with a gerund form of the expected verb 1% of the time, and with a pronoun or other subject followed by the expected verb another 18% of the time. Although the pronouns used in Experiment 2 items were much more probable than the verb at that point—they had a cloze value of 0.69—the expected verb may still have been preactivated to at least a small degree, if words are preactivated in a probabilistic fashion (Frade et al., 2022). Still, even if participants were predicting the verb prior to the pronoun, the pronoun was likely to boost the activation for the predicted verb; the offline cloze of the expected verb when given the expected pronoun was 0.25.

That the expected pronoun at least significantly boosted the prediction of the expected verb is especially likely when comparing the effect at the verb following expected pronouns to the effects seen following unexpected pronouns, similar to what was seen in the reaction time data in Experiment 1. Failed predictions are subject to lingering activation with gradual decay (Rich & Harris, 2021). Therefore, if the N400

effect seen at the verb following expected pronouns was due only to preactivation of the verb prior to the pronoun, then we might expect there to be a similar N400 effect to the unexpected verb even after an unexpected pronoun, since the expected verb would still have some lingering activation. The fact that the N400 response to unexpected verbs following unexpected pronouns is reduced for Cluster 2 and is even reversed for Cluster 1 indicates that reading the expected pronouns caused at the very least a rapid boost in the preactivation of the expected verb. This is also in line with Sun (2022), who argues that predictions may be boosted based on confirmatory evidence, but not canceled based on disconfirmatory evidence.

We therefore see evidence that upon encountering an expected pronoun in an implicit causality context, participants are able to rapidly update their predictions about the following verb.

3.6.2. Individual differences in pronoun processing and their effect on verb processing

Our second research question was how verb prediction is affected by potential individual differences in the processing of unexpected pronouns. The grand average response to unexpected pronouns included a late positivity (P600), consistent with what was found by Van Berkum et al. (2007), who argued that the response indicates that participants at least initially treat the pronoun as containing a gender agreement error. However, examination of individual responses in the current data revealed that a P600 response was only seen in a subset of participants. This indicates that there were indeed individual differences in how participants treated the unexpected pronoun.

In response to the unexpected pronoun, it was hypothesised that if readers rely more on the bottom–up information from the gender cue on the pronoun than to their implicit causality predictions, then one of two responses could be possible. If participants were unsure who the pronoun was meant to refer to, then they might exhibit an Nref effect, similar to what was found for some of the participants in Canal et al. (2015). Or participants may have just been surprised that the unexpected character was being referred to, which may have led to an N400 effect at the unexpected pronoun, caused by reduced reference likelihood (Almor et al., 2017). This appears to have been the case for Cluster 1 participants. In response to the unexpected pronoun,

this cluster showed a broadly distributed negativity 300–500 ms after the presentation of the pronoun. The distribution of this effect suggests an N400 response. However, this was followed by a sustained anterior negativity, suggesting that this cluster also included an Nref response to unexpected pronouns. Cluster 1 participants were therefore either surprised at the mention of the unexpected character or unsure who the pronoun referred to. It may also be the case that some participants responded with an N400 and others with an Nref. However, the k-means clustering algorithm still put the participants together in a single cluster, so possible differences within the cluster were not analyzed. In either case, these participants may have assigned less weight to implicit causality predictions, relying instead on the bottom–up cues given by the pronoun itself.

If, however, participants assigned greater weight to their top–down referential predictions, they might have at least initially interpreted unexpected pronouns as referring to the character predicted by implicit causality. They may then have either assumed that the unexpected pronoun had a gender agreement error or typo, or they may have attempted to repair or revise their initial interpretation or representation, perhaps by revisiting their assumptions about which pronouns characters would use. Either possibility would be expected to induce a P600 effect (Kaan & Swaab, 2003; Kolk & Chwilla, 2007; Nieuwland & Van Berkum, 2008; Van Berkum et al., 2007). This was what was found for Cluster 2 participants, who exhibited a broadly-distributed positivity to unexpected pronouns that was strongest in the 500–800 ms window after the presentation of the pronoun. This suggests that Cluster 2 participants did assign greater weight to their top–down implicit causality predictions rather than relying on the bottom–up gender cues on the pronoun, at least in their initial processing of the pronoun. They may have assumed that the pronouns contained a typo, or they may have revised their initial pronoun interpretation or their initial assumptions about which pronouns characters in the stories would use. This last possibility is supported by one Cluster 2 participant's comment during the experiment. They said, "I thought Talia was a girl's name, but I guess it wasn't." This suggests that they were indeed interpreting the unexpected pronoun in line with their implicit causality predictions, and just assumed that the predicted referent used different pronouns than they initially

thought.

These results reveal individual differences in how readers react to conflict between their top–down implicit causality predictions and the bottom–up gender information on a pronoun. Some readers assign greater weight to the bottom–up gender information and assume that the conflict arises from incorrect referential predictions or ambiguous reference, leading to an N400/Nref; others assign greater weight to their top–down predictions and assume that the unexpected gender on the pronoun is either itself an error, or that it reveals an error in their earlier semantic representation of the characters that requires revision, leading to a P600. Interestingly, these differences in pronoun processing did not appear to persist to sentence-final interpretations, when looking at all questions asking about pronoun interpretation. The two clusters did not differ in their offline comprehension question accuracy, even on the few questions that asked specifically about the pronoun. It appears, then, that regardless of how participants initially reacted to the unexpected pronoun, they ended up interpreting the pronouns the same way in the majority of cases.

However an exploratory analysis of comprehension question accuracy suggests that there was a marginal difference in responses to questions about the referent of unexpected pronouns only ($t = 1.79$, $p = 0.082$), with Cluster 1 participants more likely than Cluster 2 to interpret the pronoun offline as referring to the character that matched it in stereotypical gender (proportion of "correct" question responses with apparent reference to character with stereotypical gender for pronoun: 0.76 ± 0.37 for Cluster 1, 0.42 ± 0.58 for Cluster 2). This was a very small subset of items (each participant saw anywhere from three to twelve items asking about the referent of unexpected pronouns), and the results can therefore be interpreted at most as suggestive that even in offline judgments, Cluster 1 participants may have been more likely to rely on bottom–up information from the pronoun and Cluster 2 participants on top–down referential predictions.

What causes these differences in pronoun interpretation or processing strategy is uncertain. Clusters were determined based on the recorded data itself, rather than on individual characteristics of participants, and the two clusters were fairly uniform in external measures. Based on an exploratory analysis of the clusters, there was no

difference in the reported gender or handedness of the participants in the two clusters or in which university they were attending. However, the two clusters did differ significantly in age, with Cluster 2 having a slightly younger and more uniform age range (Cluster 1: 23.3±6.2 years (mean±SD); Cluster 2: 19.9±1.4 years, $t = 2.43$, $p = 0.021$). It may be the case, then, that the younger participants were more likely to rely on their top–down predictions in interpreting the unexpected pronoun. I will return to this point in the general discussion.

Regardless of the underlying cause of the differences seen in pronoun processing, these differences had a significant effect on participants' verb predictions. Cluster 1 participants' responses to verbs following unexpected pronouns included a less negative response to unexpected verbs than to expected verbs in posterior regions. The morphology of the ERP waveform suggests that this is due to an enhanced N400 to expected verbs. This suggests that the "expected" verbs were more difficult to process than "unexpected" verbs. One possible explanation of the combined response to unexpected pronouns and unexpected verbs in Cluster 1 is that when these participants read the unexpected pronoun, they interpreted it as not referring to the character predicted by implicit causality (even if they were unsure who it was being used to refer to). Based on that interpretation, they rapidly updated their predictions about the verb, and were thereafter surprised to read the "expected" verb, which would be associated with the originally predicted character, leading to an enhanced N400 relative to when that verb appeared after the expected pronoun.

As for Cluster 2, which exhibited a P600 to the unexpected pronoun, if these participants' initial interpretation of the pronoun was indeed consistent with their implicit causality predictions, then they might be expected to make the same predictions about the following verb as when they had read the expected pronoun. This would result in the same N400 effect to the unexpected vs. expected verb, regardless of pronoun. This hypothesis was partially confirmed. Unexpected verbs showed a greater N400 response compared to expected verbs, regardless of pronoun, suggesting that the expected verb was predicted in either case. However, the difference between unexpected and expected verbs was smaller following unexpected pronouns. One possible explanation for this is that having to deal with the unexpected pronoun was enough to limit participants'

ability to make predictions about the verb. This could have led to weaker predictions about the verb, which would account for the relatively weak N400 effect. An alternative explanation is that dealing with the unexpected pronoun caused these participants to not make any new predictions, in which case the N400 effect seen could be due to lingering activation from verb predictions prior to the pronoun, as discussed above. In either case, this response differs from Cluster 1, indicating that individual differences in how individuals responded to the unexpected pronoun did have an immediate effect on their prediction and processing of the following verbs.

4. General discussion

This paper has investigated how quickly readers can update their predictions about a following verb based on pronoun interpretation, as well as whether that is affected by individual differences in how an unexpected pronoun is processed. Stimuli for Experiments 1 and 2 consisted of contexts in which a specific pronoun was highly likely based on implicit causality predictions, and in which a specific verb was also predictable following that expected pronoun. The effects of pronoun and verb predictability were measured using reaction time measures in a maze task (Experiment 1) and with ERPs (Experiment 2).

4.1. Individual differences in pronoun processing: top–down vs. bottom–up cues

The results of Experiment 1 replicated findings from previous literature (e.g. Koornneef & Van Berkum, 2006; Stewart et al., 2000) that unexpected pronouns cause a slowdown in processing relative to pronouns predicted based on implicit causality. RTs alone, however, do not make it clear what the source of this slowdown is. Participants may have been surprised that the unexpected character was being referred to or simply been unsure who the intended referent of the pronoun was; or they may have seen the unexpected gender on the pronoun as a morphosyntactic error or typo, or that the gender representations they had formed for the characters were incorrect. Experiment 2 examined the possible underlying cognitive processing differences in response

to unexpected pronouns across individuals using ERPs.

Experiment 2 revealed two qualitatively different reactions to unexpected pronouns. Some participants showed an enhanced N400/Nref response. Almor et al. (2017) argue that the N400 is sensitive to reference predictability, and Van Berkum (2009) argues that the Nref appears when there is unclear reference. This suggests that these participants interpreted the pronoun as referring to an unexpected character, in violation of their implicit causality predictions, or that they were unsure who the pronoun referred to. In either case, they placed higher weight on the bottom–up gender information on the pronoun than their top–down predictions about likely next referents.

By contrast, other participants showed an increased P600 in response to the unexpected pronoun. Van Berkum et al. (2007) found a similar response to implicit-causality-violating pronouns and argued that it reflected a perceived gender violation. However, it could also be due to participants revising their initial interpretation or representation of the gender of the characters in the story (Kaan & Swaab, 2003; Kolk & Chwilla, 2007; Nieuwland & Van Berkum, 2008). In either case, participants appear to have given higher weight to their top–down implicit causality predictions and at least initially interpreted the pronoun in line with their referential predictions. Responses to questions specifically asking about the referent of unexpected pronouns suggests that these differences might have continued into offline judgments, with participants exhibiting a positivity to unexpected pronouns being marginally more likely to interpret pronouns based on their top–down predictions than those who exhibited a negativity. However, this possibility is based on a very small number of items, so the question of whether the effects seen here in the ERP responses carries over into later offline interpretations is left for future research.

The results illustrate that there are individual differences in the processing of unexpected pronouns in implicit causality contexts. Some participants are more likely to rely on top–down predictions in reference resolution, and others more on bottom–up information. This study did not, however, directly investigate why individuals would differ in this way. Instead, k-means clustering was used to identify clusters of individuals based directly on the recorded ERP responses to the pronoun. The only significant difference we found between the clusters in terms of individual characteristics measured

was a small difference in age. Cluster 2 participants were on average around three years younger than Cluster 1 participants. Although it seems unlikely that this small difference would cause a change in whether participants are more like to rely on top–down versus bottom–up information in reference resolution, recent societal changes in pronoun usage may have caused the younger participants to have had more exposure to and interactions with individuals who use pronouns other than the ones the participants would initially assume. They may therefore have been more willing to revise their gender representations of the characters in the items in this experiment, which would allow them to interpret the pronoun consistent with their implicit causality predictions.

In addition to a possible effect of age differences between the clusters, previous research has given several additional potential reasons for the difference seen in pronoun processing. Again, because the current study was focused on discovering what differences exist, it is left to future research to investigate which, if any, of these factors may have been at play in the current study. The following discussion thus highlights potential avenues for future research.

First, differences in pronoun processing and the likelihood of using top–down versus bottom–up information may be tied to individual differences in print exposure and language experience. Johnson and Arnold (2021) found that higher print exposure as measured by the Author Recognition Task (ART) is associated with higher use of implicit causality predictions in pronoun selection. Similarly, Langlois and Arnold (2020) found that scores on the ART were associated with the extent to which participants rely on a subject bias in pronoun interpretation, with higher scores leading to higher subject-biases. The ability to make and rely on predictions may also be related to language experience more generally (see Farmer et al., 2005; Hersch & Andrews, 2012; Kaan, 2014; Mani & Huettig, 2012). Therefore, in items such as those used in our study, individuals with higher text exposure may be more likely to rely on top–down predictions in reference resolution.

Second, Nieuwland and Van Berkum (2006) found that the size of the Nref response to ambiguous pronouns depends in part on Reading Span score, which was used to assess language processing ability. Similarly, Qiu et al. (2012) found that in written Chinese, pronouns that mismatch in gender with their antecedent elicit an N400 when

the antecedent and pronoun are separated by a short interval, but a P600 at longer intervals, and argue that this is due to working memory demands. In addition, Huettig and Janse (2015) found using the visual-world paradigm that enhanced working memory capacity and faster processing speed increase participants' ability to make predictions based on article gender. Therefore, differences in pronoun processing in the current study may have been related to individual differences in language processing ability and more general working memory capacity.

Third, Van Berkum et al. (2009) found that ERP responses can be modulated by individual differences in personal views on morality. In the current study, it could have been that individual views on non-binary or transgender issues may have affected whether participants were likely to revise their gender representations of the characters.

Finally, individual differences have also been found in how clear morphosyntactic errors are processed, with some individuals displaying a classic P600 and others an N400 (Grey et al., 2017; Tanner & Van Hell, 2014). In the current experiment, then, at least some of the participants in who exhibited a negativity to unexpected pronouns may have done so as a result of assuming the pronoun contained a morphosyntactic error. If that were the case, it might mean that some participants treated unexpected pronouns qualitatively the same, as containing an error, but had different brain responses based on that. However, this seems unlikely, since the two groups of participants had different responses to the verbs following unexpected pronouns. Participants who exhibited a P600 to unexpected pronouns showed evidence of qualitatively similar predictions regardless of pronoun, while those who showed a negativity to the unexpected pronoun appear to have changed their verb predictions.

4.2. Rapid verb prediction based on pronoun processing

This study also examined whether the processing of a pronoun would have an immediate effect on readers' prediction of the following verb. In Experiment 1, readers had slower RTs for unexpected verbs relative to expected verbs, but only after expected pronouns. This indicates that the processing of the pronoun had an immediate effect on the processing of the following verb. However, although that effect may have been due to rapid prediction based on the pronoun, that is not the only possible explana-

tion of the Experiment 1 results on their own. Participants may have made predictions about the following verb regardless of pronoun interpretation—perhaps even prior to the pronoun—but processing the unexpected pronoun may have simply made processing the following verb more difficult overall, so the effect of prior predictions did not show up.

However, the results from Experiment 2 give further evidence that readers can indeed make rapid predictions (within 500 ms) about the upcoming verb after reading an expected pronoun, and perhaps even after reading an unexpected pronoun. Both groups of participants showed an enhanced N400 to the unexpected verb when it followed an expected pronoun. Because the N400 is more uniquely sensitive to prediction than RTs, if the expected verb was predicted regardless of pronoun interpretation, we might expect similar effects at the verb following either pronoun. Instead, the N400 indicated that expected verbs were facilitated more strongly relative to unexpected verbs following the expected pronoun. There was still some evidence for preactivation of the expected verb following unexpected pronouns in Cluster 2, but the effect was reduced, suggesting that the expected pronoun at the very least caused a boost in the preactivation of the expected verb. For Cluster 1, the evidence is even stronger, since the expected verb showed an even larger N400 response than the unexpected verb following unexpected pronouns, suggesting that not only did the expected pronoun lead to an update in verbal predictions, but the unexpected pronoun may have also.

Interestingly, this is true in spite of evidence that reading a pronoun does not reactivate its antecedent in memory (Lago et al., 2019; Smith & Federmeier, 2018). It was not the case, then, that the expected pronoun caused local priming of the expected verb just based on reactivation of the antecedent in memory and spreading activation to semantically related verbs. Despite the antecedent not necessarily being reactivated in memory, readers still were able to update their predictions about the verb based on the growing conceptual structure. These results are also inconsistent with general conceptual priming of the verb, in contrast to the results in Delogu et al. (2019), since the expected verb was generally a better fit to the context prior to the pronoun. If only general priming was involved, we would expect it to cause the same effects regardless of pronoun.

When participants read an unexpected pronoun, the facilitation of the "expected" verb was reduced for Cluster 2 participants and even reversed for Cluster 1. One explanation for this is that participants' ability to make predictions about the following verb was reduced, perhaps due to the additional cognitive demands required in integrating a pronoun that violated implicit causality predictions with the prior context, in struggling to resolve unclear reference, in revising gender representations of the characters, or in dealing with a perceived morphosyntactic error. This would contribute to our understanding about what types of computational processes slow prediction (Chow et al., 2016, 2018; Liao & Lau, 2020). However, there is still evidence that participants were making some verb predictions after the unexpected verbs. Cluster 2 participants may still have initially interpreted both the expected and unexpected pronouns as referring to the same character, warranting similar predictions about the verb, and causing a similar N400 facilitation for expected verbs, even if the prediction was weaker following unexpected pronouns. As for Cluster 1, if these participants assumed that the unexpected pronoun did not refer to the character predicted by implicit causality, they may have specifically not predicted the "expected" verb, leading to a larger N400 when it was seen.

The results from the two experiments together suggest that participants did rapidly update their predictions about the upcoming verb following both expected pronouns and unexpected pronouns, although because the cloze probability of the two verbs following unexpected pronouns was more equal, the evidence for prediction following unexpected pronouns is somewhat weaker. Still, this contrasts with other types of information such as argument roles that can affect verb predictions only at a delay. Chow et al. (2016) present a theory that readers predict verbs by first activating events associated with the event participants without respect to the specific role played by those participants, in a "bag-of-arguments" fashion, and then only later are able to filter those verb predictions based on argument role information (see also Chow et al., 2018; Liao & Lau, 2020; Liao et al., 2022). In theory, one might have expected a similar effect to happen here, since the pronoun is the first bottom–up indication that one of the previous characters is involved in the upcoming verb. However, the fact that the verb can be rapidly predicted based on pronoun interpretation suggests that the implicit

causality context gives a boost to the verb, perhaps by causing readers to predict not only potential referents, but the role those referents will play in the predicted causal event. Further research is necessary to investigate whether a strongly predictive context such as an implicit causality sentence is necessary in order for pronoun interpretation to have such a rapid effect on verb predictions.

4.2.1. Prediction versus integration

An alternative interpretation of the results might be that the effects seen at the verb in both Experiments 1 and 2 were due to integration difficulty, rather than facilitation through prediction. The expected verb was generally a better fit for the context given the expected pronoun than the unexpected verb. Given the unexpected pronoun, however, both the unexpected and "expected" verbs may have been equally (un)fit for the context.

The functional interpretation of the N400 has been debated (see Kutas & Federmeier, 2011). This paper has assumed one common interpretation: that the N400 reflects lexical access mechanisms and is sensitive to the extent to which lexical items are preactivated through prediction (e.g. Kutas & Federmeier, 2000; Lau et al., 2016; Stone et al., 2023). Another view, though, is that the N400 reflects semantic integration processes (e.g. Hagoort et al., 2009; Hagoort et al., 2004), and is only sensitive to prediction insofar as predictable words are often also more easily integrated into the semantic context. Other accounts argue for dual generators, such that an early sub-component of the N400 is sensitive to effects of lexical access, and a later sub-component to semantic integration (Nieuwland et al., 2019).

If an integration view of the N400 were correct, then these results would not necessarily indicate that the interpretation of the pronoun led to rapid updating of predictions about the following verb. Instead, it would just mean that the unexpected verb was more difficult to integrate than the expected verb following the expected pronoun, but not following the unexpected pronoun. This possibility cannot be ruled out by the results in this study, as it did not directly control for predictability versus semantic integration difficulty.

However, several recent studies have directly compared the effects of predictability

and plausibility on the N400. These studies have found a clear effect of word predictability on the N400 even when plausibility, and thus semantic integration difficulty, have been controlled (Delogu et al., 2019; Lau et al., 2016; Mantegna et al., 2019; Sun, 2022). In fact in some cases, the N400 effect has been shown to disappear when words are made predictable even when those words are not plausible in the context (Szewczyk & Schriefers, 2017; Szewczyk & Wodniecka, 2020). These results would not be explainable under an integration account, or even under a dual-generator account in some cases. This study follows these recent findings and assumes that the effects seen do at least in part reflect that the expected pronouns were more predicted than unexpected pronouns, and that the interpretation of these pronouns led to rapid predictions about the following verbs.

4.3. The use of the A-maze procedure and the maze task

This research also has methodological implications. The A-maze procedure for automatic generation of competitor words in the maze task is relatively new (Boyce et al., 2020). Although the maze paradigm is somewhat artificial and may therefore not accurately reflect processing that would occur in a more natural context, these results demonstrate that the procedure, and specifically the generation of competitor words with A-maze, can successfully be used to investigate the processing of unambiguous sentences where the difficulty lies in the predictability of the words given, rather than in ambiguity resolution as was used in earlier A-maze studies, although it may not easily distinguish between possible sources of the processing difficulty. This is especially relevant in situations where in-person data collection is difficult or impossible, but in which remote data collection is possible (not only pandemics, but in researching distant populations when travel is not feasible).

5. Conclusion

This paper investigated two main questions: whether pronoun interpretation in implicit causality contexts can result in rapid predictions about the following verb, and whether that is affected by individual differences in pronoun processing. It was found

that reading an expected pronoun caused immediate facilitation of a following predicted verb. Following unexpected pronouns, verb predictions appear to have differed based on how individuals processed the unexpected pronoun. If participants relied more on their top–down implicit causality predictions to interpret the pronoun, their verb predictions remained relatively unchanged, regardless of pronoun. If they relied more on bottom–up information from the pronoun and interpreted the expected and unexpected pronouns differently, they also made different predictions about the verb. Future research will investigate possible sources for these individual differences in pronoun and verb processing.

Acknowledgments

The author would like to thank the following individuals for assistance in stimulus development and data collection: Sarah Anderson, Sean Cameron, Emily Charny, Stephanie Chen, Siena Christensen, Drew Galbraith, Oxana Kodirova, and Aaron Koroghlanian. He also would like to thank the audience at HSP 2023 as well as the anonymous reviewers for their comments and feedback.

Disclosure statement

The author reports there are no competing interests to declare.

Data availability

The stimuli for this study and the data that support the findings as well as analysis scripts are openly available in OSF at http://doi.org/10.17605/OSF.IO/B2TEM.

Notes

[1] A reviewer correctly notes that referring to people using both their and occupations and names (especially in uppercase letters; e.g. *Midwife Crystal*, *Nervous Parent Jackson*) is not completely naturalistic. Although this may have affected reaction times in this study, whatever effect it had is likely to have been similar across conditions. The major results are therefore unlikely to be due to this aspect of the materials.

References

Almor, A., Nair, V. A., Boiteau, T. W., & Vendemia, J. M. (2017). The N400 in processing repeated name and pronoun anaphors in sentences and discourse. *Brain and Language*, *173*, 52–66. https://doi.org/10.1016/j.bandl.2017.06.003

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560. https://doi.org/10.1080/01690965.2012.658072

Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistics Compass*, *4*(4), 187–203. https://doi.org/10.1111/j.1749-818x.2010.00193.x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082. https://doi.org/10.1016/j.jml.2019.104082

Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, *125*, 104334. https://doi.org/10.1016/j.jml.2022.104334

Canal, P., Garnham, A., & Oakhill, J. (2015). Beyond gender stereotypes in language comprehension: Self sex-role descriptions affect the brain's potentials associated with agreement information-processing. *Frontiers in Psychology*, *6*, 1953. https://doi.org/10.3389/fpsyg.2015.01953

Chow, W.-Y., & Chen, D. (2020). Predicting (in)correctly: Listeners rapidly use unexpected information to revise their predictions. *Language, Cognition and Neuroscience*, *35*(9), 1149–1161. https://doi.org/10.1080/23273798.2020.1733627

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, *33*(7), 803–828. https://doi.org/10.1080/23273798.2018.1427878

Chow, W.-Y., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016). Prediction as memory retrieval: Timing and mechanisms. *Language, Cognition and Neuroscience*, *31*(5), 617–627. https://doi.org/10.1080/23273798.2016.1160135

Cozijn, R., Commandeur, E., Vonk, W., & Noordman, L. G. M. (2011). The time course of the use of implicit causality information in the processing of pronouns: A visual world paradigm study. *Journal of Memory and Language*, *64*(4), 381–403. https://doi.org/10.1016/j.jml.2011.01.001

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569. https://doi.org/10.1016/j.bandc.2019.05.007

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

Farmer, T. A., Christiansen, M. H., & Kemtes, K. A. (2005). Sentence processing in context: The impact of experience on individual differences. *Procedings of the Annual Meeting of the Cognitive Science Society*, *27*.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

Frade, S., Pinheiro, A. P., Santi, A., & Raposo, A. (2022). Is second best good enough? An EEG study on the effects of word expectancy in sentence comprehension. *Language, Cognition and Neuroscience*, *37*(2), 209–223. https://doi.org/10.1080/23273798.2021.1955140

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, *33*(1), 39–68. https://doi.org/10.1006/jmla.1994.1003

Gibson, E. (2006). The interaction of top–down and bottom–up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, *54*(3), 363–388. https://doi.org/10.1016/j.jml.2005.12.005

Green, J. J., McCourt, M., Lau, E., & Williams, A. (2020). Processing adjunct control: Evidence on the use of structural information and prediction in reference resolution. *Glossa: A Journal of General Linguistics*, *5*(1). https://doi.org/10.5334/gjgl.1133

Grey, S. (2022). Variability in native and nonnative language: An ERP study of semantic and grammar processing. *Studies in Second Language Acquisition*, *45*(1), 137–166. https://doi.org/10.1017/s0272263122000055

Grey, S., Tanner, D., & van Hell, J. G. (2017). How right is left? Handedness modulates neural responses during morphosyntactic processing. *Brain Research*, *1669*, 27–43. https://doi.org/10.1016/j.brainres.2017.05.024

Grüter, T., Takeda, A., Rohde, H., & Schafer, A. J. (2018). Intersentential coreference expectations reflect mental models of events. *Cognition*, *177*, 172–176. https://doi.org/10.1016/j.cognition.2018.04.015

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proc NAACL*.

Haeuser, K. I., & Kray, J. (2022). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, *43*(5), 1193–1220. https://doi.org/10.1017/s0142716422000364

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 819–836, Vol. 4). MIT Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hersch, J., & Andrews, S. (2012). Lexical quality and reading skill: Bottom-up and top-down contributions to sentence processing. *Scientific Studies of Reading*, *16*(3), 240–262. https://doi.org/10.1080/10888438.2011.564244

Huettig, F., & Janse, E. (2015). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. https://doi.org/10.1080/23273798.2015.1047459

Husband, E. M. (2022). Prediction in the maze: Evidence for probabilistic pre-activation from the English *a/an* contrast. *Glossa Psycholinguistics*, *1*(1), 7, pp. 1–19. https://doi.org/10.5070/g601153

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, *10*(2), 370–375. https://doi.org/10.1016/0013-4694(58)90053-1

Johnson, E., & Arnold, J. E. (2021). Individual differences in print exposure predict use of implicit causality in pronoun comprehension and referential prediction. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.672109

Kaan, E. (2014). Predictive sentence processing in L2 and L1. *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and visualize the results of multivariate data analyses* [R package]. Version 1.0.7. https://CRAN.R-project.org/package=factoextra

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language*, *56*(3), 384–409. https://doi.org/10.1016/j.jml.2006.09.003

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44. https://doi.org/10.1093/jos/ffm018

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1-2), 1–37. https://doi.org/10.1515/tl-2013-0001

Kim, A. E., McKnight, S., & Miyake, A. (2023). *How variable are the classic ERP effects during sentence processing? A systematic resampling analysis of the N400 and P600 effects* [Manuscript submitted for publication]. https://doi.org/10.31234/osf.io/3k4sp

Kolk, H., & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language*, *100*(3), 257–261. https://doi.org/10.1016/j.bandl.2006.07.006

Koornneef, A. W., & Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, *54*(4), 445–465. https://doi.org/10.1016/j.jml.2005.12.003

Kush, D., & Dillon, B. (2021). Principle B constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language*, *120*, 104254. https://doi.org/10.1016/j.jml.2021.104254

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/s1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, Vol 62*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lago, S., Namyst, A., Jäger, L. A., & Lau, E. (2019). Antecedent access mechanisms in pronoun processing: Evidence from the N400. *Language, Cognition and Neuroscience*, *34*(5), 641–661. https://doi.org/10.1080/23273798.2019.1566561

Langlois, V. J., & Arnold, J. E. (2020). Print exposure explains individual differences in using syntactic but not semantic cues for pronoun comprehension. *Cognition*, *197*, 104155. https://doi.org/10.1016/j.cognition.2019.104155

Lau, E. F., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra*, *2*(1). https://doi.org/10.1525/collabra.40

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews. Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lenth, R. V. (2021). *emmeans: Estimated marginal means, aka least-squares means* [R package]. Version 1.6.2-1. https://CRAN.R-project.org/package=emmeans

Liao, C.-H., & Lau, E. (2020). Enough time to get results? An ERP investigation of prediction with complex events. *Language, Cognition and Neuroscience*, *35*(9), 1162–1182. https://doi.org/10.1080/23273798.2020.1733626

Liao, C.-H., Lau, E., & Chow, W.-Y. (2022). Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language*, *126*, 104350. https://doi.org/10.1016/j.jml.2022.104350

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., ... Wolfsthurn, S. V. G. Z. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*, 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, *1118*(1), 155–167. https://doi.org/10.1016/j.brainres.2006.08.022

Nieuwland, M. S., & Van Berkum, J. J. A. (2008). The interplay between semantic and referential aspects of anaphoric noun phrase resolution: Evidence from ERPs. *Brain and Language*, *106*(2), 119–131. https://doi.org/10.1016/j.bandl.2008.05.001

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596x(92)90039-z

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773. https://doi.org/10.1006/jmla.1995.1033

Perception Research Systems. (2007). *Paradigm stimulus presentation* (Version 2.5.0.68) [computer software].

Qiu, L., Swaab, T. Y., Chen, H.-C., & Wang, S. (2012). The role of gender information in pronoun resolution: Evidence from Chinese. *PLoS ONE*, *7*(5), e36156. https://doi.org/10.1371/journal.pone.0036156

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.2.1). Vienna, Austria. https://www.R-project.org/

Reuter, T., Dalawella, K., & Lew-Williams, C. (2020). Adults and children predict in complex and variable referential contexts. *Language, Cognition and Neuroscience*, *36*(4), 474–490. https://doi.org/10.1080/23273798.2020.1839665

Rich, S., & Harris, J. A. (2021). Unexpected guests: When disconfirmed predictions linger. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 2246–2252. https://escholarship.org/uc/item/7q0000c8

Rohde, H., & Kehler, A. (2013). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927. https://doi.org/10.1080/01690965.2013.854918

RStudio Team. (2021). *RStudio: Integrated development environment for R* (Version 2022.12.0). Boston, MA.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Smith, C. M., & Federmeier, K. D. (2018). What does "it" mean, anyway? examining the time course of semantic activation in reference resolution. *Language, Cognition and Neuroscience*, *34*(1), 115–136. https://doi.org/10.1080/23273798.2018.1513540

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, *42*(3), 423–443. https://doi.org/10.1006/jmla.1999.2691

Stone, K., Nicenboim, B., Vasishth, S., & Rösler, F. (2023). Understanding the effects of constraint and predictability in ERP. *Neurobiology of Language*, *4*(2), 221–256. https://doi.org/10.1162/nol_a_00094

Sturt, P. (2013). Syntactic constraints on referential processing. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 136–159). Psychology Press.

Sun, Y. (2022). *Phonological prediction in spoken language processing* [Doctoral dissertation, University of Illinois at Urbana-Champaign].

Szewczyk, J. M., & Schriefers, H. (2017). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. https://doi.org/10.1080/23273798.2017.1401101

Szewczyk, J. M., & Wodniecka, Z. (2020). The mechanisms of prediction updating that impact the processing of upcoming word: An event-related potential study on sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1714–1734. https://doi.org/10.1037/xlm0000835

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301. https://doi.org/10.1016/j.neuropsychologia.2014.02.002

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Van Berkum, J. J. A. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science*, *20*(9), 1092–1099. https://doi.org/10.1111/j.1467-9280.2009.02411.x

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158–171. https://doi.org/10.1016/j.brainres.2006.06.091