A Possible Cure for ‘N400 Blindness’ to Role Reversal Anomalies in Sentence Comprehension

Lara Ehrenhofer, Ellen Lau, Colin Phillips
University of Maryland

Abstract

Human comprehenders rapidly and incrementally integrate linguistic information into predictions about upcoming sentence material. Though rare, cases where information does not immediately impact predictions provide important insights into predictive mechanisms. One well-studied case is argument role information, which many studies have shown not to immediately impact the N400, an ERP index of prediction, when argument roles on nouns are reversed from their canonical order. In the current study, our aim was to determine whether verbs are necessary for argument role information to rapidly impact prediction. Instead, we serendipitously discovered a set of role-reversal materials that yield the immediate effects of argument roles on predictions that have so far been largely absent in the literature. In a follow-up experiment, we confirmed our results, again demonstrating an immediate effect of argument role reversal on the N400 in our materials, as well as replicating the absence of N400 effects with the original materials from one of the prior studies. Our results and analyses suggest a new avenue of research into role-reversal anomalies, exploring fine-grained nuance in the contextual properties that determine whether argument position has an immediate impact on prediction.
1. Introduction

Identifying who does what to whom within a sentence is a key aspect of successful comprehension, and it usually requires relating an event described by a verb to participants in that event described by noun arguments. Syntactic information like word position and morphology is often critical for assigning the correct role to each argument; for example, while the bull has different roles in the bull kicked the cowboy and the cowboy kicked the bull, it has the same role in the bull gored the cowboy, the cowboy was gored by the bull, and the bull that the cowboy had gored.

Recent investigations of predictive mechanisms in language comprehension have benefited greatly from work on argument role assignment: although dozens of studies have shown that comprehenders can rapidly and successfully predict upcoming input on the basis of diverse and subtle cues, considerable evidence has suggested that argument role information has a delayed impact on prediction. In offline measures like sentence completion (cloze probability), comprehenders differentiate between bull-as-agent and bull-as-patient events, differentially predicting events such as gore or ride depending on argument role assignment. However, in online measures like the N400, an electrophysiological component that is usually sensitive to the predictability of a word in context (Kutas & Hillyard, 1984), ‘role reversals’ (which cowboy the bull had gored vs. which bull the cowboy had gored) appear to have no immediate impact (Kuperberg et al., 2003; Kolk et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; van Herten et al. 2005; van Herten et al., 2006; Ye & Zhou, 2008; Oishi & Sakamoto, 2010; Stroud &
Phillips, 2012; Chow & Phillips, 2013; Chow et al., 2016b). These kinds of results have motivated recent attempts to model the time course by which different information sources impact predictions (Kukona et al., 2011; Chow et al., 2016a; Kuperberg, 2016; Chow et al., 2018). In the current study, we report a new case in which argument role information is able to rapidly impact predictions, which suggests a new avenue of investigation for better understanding predictive mechanisms in online comprehension.

1.1. Blindness to argument role information in prediction

In the study of language comprehension, the N400 response in event-related brain potentials (ERPs) is one of the most well-known online processing measures to track the predictability of a word relative to the prior context. The N400 response is a negative-going deflection in the ERP response to words and other meaningful stimuli (pictures, environmental sounds) across modalities (Ganis et al. 1996; Orgs et al., 2008). Predictability is often operationalised as the likelihood that participants provide a given word in an online completion task, also known as cloze probability. Much work since the early 1980s has demonstrated that the amplitude of the N400 response to a word in context is systematically reduced with increasing cloze probability, across auditory and visual modalities (e.g. Kutas & Hillyard, 1984; Gunter, Friederici & Schriefers, 2000; Wlotko & Federmeier, 2013). N400 amplitude is similarly sensitive to measures of predictability derived from corpus probabilities (e.g. Strauss, Kotz, & Obleser, 2013; Lau, Namyst, Fogel, & Delgado, 2016). Many researchers have suggested that the amplitude of the N400 is additionally modulated by the processing cost engendered by outright semantic
anomaly (as in She spread her warm toast with socks; e.g., Kutas & Hillyard, 1980; Brown & Hagoort, 1993; Hagoort, Hald, Bastiaansen, & Petersson, 2004), although others have suggested that such effects can be subsumed under differences in predictability (e.g. Kutas & Hillyard, 1984; Lau et al., 2008; Brouwer, Fitz, & Hoeks, 2012).

Because of the extensive prior work establishing a tight link between N400 amplitude and predictability, it is surprising and interesting that this relationship breaks down in role-reversed sentences. Role-reversed sentences are those cases such as the bull that the cowboy gored in which the canonical role assignment on context nouns is revealed to be anomalously reversed at the target verb. A wealth of studies have found that these sentences do not result in higher N400 amplitudes than unreversed controls (the cowboy that the bull gored), but like many anomalous sentences, they do elicit late positivities or ‘P600s’ (Kim & Osterhout, 2005; Brouwer et al., 2012).¹ This is true even when cloze probabilities are explicitly collected to establish that verb predictability is much lower in the role-reversed sentences than in the control sentences (Chow et al., 2016b; Chow et al., 2018).

Early studies often assumed that the P600 was syntax-specific and that the N400 was a reliable index of semantic anomaly detection, and therefore the dominant explanation for the lack of an N400 contrast in role-reversal sentences was that it was due to “semantically attractive” combinations of nouns and verbs, in which the parser constructs the most plausible interpretation regardless of bottom-up syntactic information (Kim & Osterhout 2005; van Herten et al., 2006; Kuperberg, 2007). Much of the early discussion in the literature therefore focused on the debate around the in(ter)dependence of syntactic and semantic processing. Kim & Osterhout (2005)

¹ Note that a few studies in German, Icelandic and Turkish do show an N400 sensitivity to argument role reversals (see Bornkessel-Schlesewsky et al., 2011). We discuss these in greater detail in the General Discussion.
contrasted active and passive sentences such that the subject was either a good agent for the verb or not (The hearty meal was devoured vs. ...devouring...), and measurement on the verb revealed a lack of N400 contrast, although there was a P600 contrast. Kuperberg, Sitnikova, Caplan, & Holcomb (2003) compared responses to sentences whose subjects were related to the verb, but were either a good agent for the verb (“for breakfast, the boys would only eat...”) or a poor agent (“for breakfast, the eggs would only eat...”), and again found no N400 contrast on the verb. Hoeks et al. (2004) extended this finding into Dutch sentences with a word order where both arguments of the verb appear prior to the verb itself. They contrasted active and passive sentences (the javelin-nom was by the athletes-dat thrown vs. the javelin-nom the athletes-acc threw) and once again found no N400 contrast on the verb, despite the strong implausibility of the active sentence.

However, around the same time, it began to be recognised that the P600/late positivity is associated with detection of many kinds of anomalies—not just syntactic, but semantic incongruity of many kinds, as well as orthographic violations (Münte et al., 1998; Patel et al., 1998; Brouwer et al., 2012). This results in a somewhat different perspective on the ERP responses to role-reversals: the fact that role-reversals almost always elicit P600s suggests that comprehenders do detect the semantic anomaly associated with role reversals very rapidly. Also, it became apparent that P600 effects elicited by semantic anomalies were not dependent on the presence of ‘semantic attraction’ between the verb and the arguments (Paczyński & Kuperberg, 2011; Paczynski & Kuperberg, 2012; Stroud & Phillips, 2012; Chow & Phillips, 2013). Therefore, focus has now begun to shift instead to the question of why the same role information does not impact N400 amplitudes through its impact on the predictability of the critical word.
Several recent results suggest that argument role information can impact N400 amplitudes, but that it takes some time. Chow et al. (2018) presented subject-object-verb (SOV) sentences in Mandarin (where the co-verb ba unambiguously indicates the subject and object status of the two NPs) whose final verb cloze probability contrasted depending on word order (canonical or reversed). They found no N400 contrast when the target verb immediately followed the two NPs, but when a temporal adverb (e.g. last week) was inserted between the NPs and the target verb, an N400 contrast emerged for items whose average target probability in the high cloze condition exceeded 40%. Similarly, Momma found no N400 effect of argument role (signalled by case-marking) in Japanese when the noun and verb were presented with an SOA of 800ms, but an N400 effect emerged when the SOA was lengthened to 1200ms (Momma, 2016). Any account of the N400’s systematic blindness to argument role information must therefore explain not only the N400’s insensitivity when the target word is close to the predictive cues, but also the emergence of an N400 contrast when the target word is further away from the predictive cues. We next explore an account that has been proposed to explain this pattern of results.

1.2. Argument roles and prediction mechanisms

Chow et al. (2016) hypothesise that prediction difficulty arises in role-reversal constructions because the process of retrieving a likely verb from memory on the basis of noun+argument-role combinations is not straightforward, for several reasons. First, structural position and case information do not fully determine a noun’s argument role of a noun (subjects could be Agents or Experiencers; in passives, they could be Themes or Patients, to name just a few roles that have
been proposed). Uncertainty around the identity of these roles may result in downweighting them as retrieval cues. Second, some linguistic theories deny the existence of generalised argument roles like ‘Agent’ and ‘Patient’ altogether in favor of event-specific roles like ‘Eater’ (for discussion see Williams, 2015). If such theories are correct, then prior to the verb there is in fact no role information available that could be used to directly map noun+role information to event representations, and a slower inferential/analytical/serial process will be required. For example, a comprehender might initially retrieve events associated with the nouns in a role-independent fashion, and then for each event testing whether it is compatible with the roles of the nouns. This could explain why N400s reflect sensitivity to role information only when the verb is presented after a delay.

Importantly, this account makes a key distinction between the predictions that comprehenders can make before, as opposed to after, encountering a verb. The Chow et al. (2016a) account assumes that the verb carries unique and critical information about how the nouns relate to the event described by the sentence, and that retrieving likely events is therefore more difficult without this information. This account predicts that verbs may be able to immediately guide role-specific predictions about upcoming nouns. This is because the verb disambiguates key aspects of how the noun relates to the sentence meaning.

Experiment 1 of the current study was designed to evaluate this prediction by measuring ERPs in response to a target verb or noun in contrasting canonical and role-reversed noun-noun-verb (NNV) and noun-verb-noun (NVN) clauses in English. As predicted by the format mismatch account, we found an N400 contrast on target nouns in the NVN conditions. However, unlike previous studies, we also found an N400 contrast on target verbs in NNV
conditions. We addressed this unexpected result in Experiment 2, which replicated the surprising finding of an N400 contrast from Experiment 1, while also confirming the finding of a lack of N400 contrast from the stimuli in a previous study (Chow et al., 2016b, Experiment 1).

2. Experiment 1

2.1. Materials

Test materials (see Table 1 for examples) consisted of sentences containing embedded NNV or NVN clauses. The NNV order was created via an indirect object wh-question, and the NVN order was created via an indirect subject wh-question. In the object-wh clauses, non-canonical OSV order created an opportunity for comprehenders to predict the clause-final verb based on the two preceding nouns. In the subject-wh clauses the canonical SVO order created an opportunity for comprehenders to predict an upcoming noun phrase based on the preceding noun and verb. Targets either matched or mismatched the argument role assignments indicated by the preceding context. This was achieved by reversing the order of the nouns (NNV: which bull the cowboy had ridden vs. which cowboy the bull had ridden; NVN: which jockey had raced the horse vs. which horse had raced the jockey). Following from previous findings that violations of selection restrictions (Kolk, Chwilla, Van Herten, & Oor, 2003; van Herten, Chwilla, & Kolk, 2006) and animacy restrictions (Kuperberg et al., 2003; Hoeks et al., 2004; Stroud & Phillips, 2012; Chow & Phillips, 2013) impact the N400 and P600 ERP components, as much as possible we avoided creating role-reversal sentences that resulted in these types of violations. While
several stimulus sets were taken directly from Chow et al. (2016b), most had to be created anew in order to satisfy the constraints of the current design.

Our experimental design departs from previous studies in fully crossing order and the target’s fit to context. Each context order was therefore paired with both a high-cloze completion which matched the argument role assignment of the preceding context, and the opposite order’s completion, which in this context was low-cloze. In most NVN stimuli sets, the demands of balancing a high cloze value for a completion in one order against that same completion having a low cloze value in the opposite order required us to change nouns between conditions; see Table 1 for details. As a point of comparison, we included two additional sets of high-low cloze contrast items in ‘filler’ sentences of varied structures in which more context was available to predict the critical word. In one set the cloze manipulation occurred on a noun and in another set it occurred on a verb (N-Filler vs. V-Filler conditions). Low-cloze completions were created by exchanging high-cloze targets to produce sentences whose targets were both low-cloze and implausible.

Table 1: EEG Sample Stimuli

<table>
<thead>
<tr>
<th>NNV contexts</th>
<th>The cattle rancher remembered...</th>
<th>mean cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Canonical, high cloze</td>
<td>… which bull the cowboy had ridden out on the range.</td>
<td>36%</td>
</tr>
<tr>
<td>b Reversed, low cloze</td>
<td>… which cowboy the bull had ridden out on the range.</td>
<td>&lt;2%</td>
</tr>
<tr>
<td>c Reversed, high cloze</td>
<td>… which cowboy the bull had gored out on the range.</td>
<td>36%</td>
</tr>
</tbody>
</table>
**d** Canonical, low cloze … which bull the cowboy had **gored** out on the range. <2%

<table>
<thead>
<tr>
<th>NVN contexts</th>
<th>The horse trainer saw...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a</strong></td>
<td>Canonical, high cloze</td>
</tr>
<tr>
<td></td>
<td>… which jockey had raced the horse across the track. 36%</td>
</tr>
<tr>
<td><strong>b</strong></td>
<td>Reversed, low cloze</td>
</tr>
<tr>
<td></td>
<td>… which horse had raced the jockey across the track. &lt;2%</td>
</tr>
<tr>
<td><strong>c</strong></td>
<td>Reversed, high cloze</td>
</tr>
<tr>
<td></td>
<td>… which horse had thrown the jockey across the track. 36%</td>
</tr>
<tr>
<td><strong>d</strong></td>
<td>Canonical, low cloze</td>
</tr>
<tr>
<td></td>
<td>… which gambler* had thrown the horse across the track. &lt;2%</td>
</tr>
</tbody>
</table>

* In NVN stimuli, the need to ensure counterbalanced cloze values for the final noun required initial noun substitutions in conditions c and/or d. These nouns were chosen to be highly associated with the remaining nouns and verbs in the stimulus set.

<table>
<thead>
<tr>
<th>Controls</th>
<th>During the gold rush, prospectors <strong>found/knitted</strong> gold in the Rocky Mountains. 36% vs. 0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-Filler</td>
<td>The environmentally friendly office recycled <strong>paper/nuts</strong> and plastic whenever possible. 36% vs. 0%</td>
</tr>
</tbody>
</table>

In cloze norming for the current study we operationalised cloze using conceptual criteria rather than exact lexical match, as participants often generated a wide range of completions that were often clustered around highly similar concepts. Therefore, roughly synonymous lemmas were counted towards the highest-cloze member of that semantic group (e.g. 20% “eaten”, 5% “devoured”, 5% “consumed” would count as 30% “eaten”). Across filler conditions, high-cloze completions averaged 36% cloze (N-filler SD: 13%; V-filler SD: 15%) and low-cloze completions had 0% cloze (SD: 0%). In the critical manipulation, in NNV contexts, high-cloze conditions averaged 35% cloze (SD: 14%) and low-cloze conditions averaged 1.4% cloze (SD:
2.8%). In NVN contexts, high-cloze conditions averaged 36% cloze (SD: 17%) and low-cloze conditions averaged 2.5% cloze (SD: 4.4%).

To ensure this careful balance of high and low-cloze completions across two sets of four test stimuli each, as well as two sets of fillers, stimuli were developed in an iterative series of 19 web-based cloze norming experiments on Amazon MTurk, comprising data from a total of over 900 unique participants (guaranteed through the use of a Unique Turker ID, https://uniqueturker.myleott.com/). Experiments varied in length from 10 to 50 minutes. Participants were paid to complete sentence fragments which ended prior to the target with “the first thing that comes to mind,” and they were asked to take no more than twenty seconds to complete each sentence. The final test and filler stimuli set were based on cloze values collected from 30 participants each.

We created sixty stimuli of each type (NNV, NVN, N-Filler, V-Filler) and distributed them across four lists in a Latin Square design that ensured that no participant saw more than one condition from each item set. Each list contained 50% high and low-cloze items and an equal proportion of stimulus types, with 240 items in all. Item order was randomised between participants. A full set of experimental items can be found in the Supplemental Materials.

2.2. Procedure

Participants sat comfortably at a distance of ca. 100 cm from a monitor and read sentences that were presented in RSVP using 24-point font. Following Chow et al. (2015), a fixation cross was shown for 500 ms at the beginning of each trial. Words were displayed for 300 ms, with a blank screen for 230 ms after each word (total SOA = 530 ms). Participants answered a yes/no
plausibility question 1000 ms after the last word of each sentence (marked with a full stop). Plausibility was defined as “something that could normally happen.” To ensure greater attention, participants’ accuracy on fillers was displayed at the end of each experimental block with a note encouraging better performance where required. Participants were offered breaks after 40 stimuli or ca. every 10 minutes. Each experimental session took an average of two hours. Participants gave informed consent and were paid $10-$15/hour (payment guidelines changed over the course of the project).

2.3. EEG recording

Continuous EEG measurements were collected from 29 AgCl electrodes placed on the participant’s head using an electrode cap (Electrocap International): midline: Fz, FCz, Cz, CPz, Pz, Oz; lateral: FP1, F3/4, F7/8, FC3/4, FT7/8, C3/4, T7/8, CP3/4, TP7/8, P3/4, P7/8, and O1/2. Scalp electrodes were referenced to the left mastoid online, and in an offline processing step, re-referenced to the average of both mastoids. To track eye movements, the electro-oculogram (EOG) was recorded at four bipolar electrode sites, with two electrodes above and below the left eye recording vertical EOG and a further two electrodes at the outer canthus of each eye recording horizontal EOG. Electrode impedances were below 10 kΩ for all participants (and for all but three participants, they were below 5 kΩ). EEG and EOG recordings were amplified and digitised online at 1kHz with a bandpass filter of 0.1–100 Hz.
2.4. Participants

Data from 24 adult participants from the University of Maryland Linguistics participant pool were included in subsequent analyses (11 female, mean age 22 years, SD: 4.3). All participants were native speakers of American English, right-handed according to the Edinburgh Handedness Test (Oldfield, 1971), with corrected-to-normal vision, no reading disabilities and no history of neurological disease. Data from a further three participants were excluded for poor performance in the plausibility task (lower than 80% accuracy on filler trials), and a further five participants due to excess EEG artefacts.

2.5. Results

2.5.1. Behavioural results

Behavioural results from the plausibility task for the critical conditions are presented in Table 2. (Note that these represent the outcome of a yes/no behavioural rating task, not the finer-grained assessment of plausibility we pursued in our use of a Likert scale in our post-hoc analysis experiment.) We assessed the contrast in plausibility judgements for test items only. A linear mixed-effects model (calculated using the R package lme4: Bates, Maechler, Bolker, & Walker, 2015) with fixed factors cloze, context and the interaction of cloze and context and random factor participant revealed a significant contrast of plausibility ratings by cloze \( (t = -14.1, p < .001 \text{ ***}) \), but not by context \( (t = -4, p > .5) \), and with no interaction between the two factors \( (t = 1.2, p > .05) \). Plausibility contrasts were closely matched across the different context
conditions, with near identical contrasts in the NNV/NVN conditions, and similarly matched contrasts in the V-filler and N-filler conditions.

Table 2: Percentage of items judged “plausible” by test condition

<table>
<thead>
<tr>
<th>Context</th>
<th>Cloze</th>
<th>% judged plausible (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNV</td>
<td>high</td>
<td>84% (8%)</td>
</tr>
<tr>
<td>NNV</td>
<td>low</td>
<td>32% (20%)</td>
</tr>
<tr>
<td>NVN</td>
<td>high</td>
<td>83% (11%)</td>
</tr>
<tr>
<td>NVN</td>
<td>low</td>
<td>36% (19%)</td>
</tr>
<tr>
<td>V-Filler</td>
<td>high</td>
<td>92% (36%)</td>
</tr>
<tr>
<td>V-Filler</td>
<td>low</td>
<td>15% (27%)</td>
</tr>
<tr>
<td>N-Filler</td>
<td>high</td>
<td>95% (22%)</td>
</tr>
<tr>
<td>N-Filler</td>
<td>low</td>
<td>16% (36%)</td>
</tr>
</tbody>
</table>

2.5.2. ERP results

Trials affected by EOG or other artefacts were removed from further analysis; this affected 15.4% of the final trial sample. Averages were computed separately per participant and per condition, based on a baseline of 100 ms pre-target and 1000 ms post-target. Waveforms for electrode Cz are shown in Figure 1. Plots showing all electrodes can be found in the Supplemental Materials. A repeated measures ANOVA was performed on time windows of 300-500 ms (N400 time window) and 600-800 ms (P600 window), for a centro-posterior region of interest (ROI) of eight electrodes around Cz (Pz; CPz; Cz; P3; P4; C3; C4; CP4). The ANOVA fully crossed cloze (high/low) and context (NNV/NVN). For test items, in the 300-500 ms window, this two-by-two ANOVA revealed a main effect of cloze ($F = 4.5, p < .05^\star$) but not of context ($F = .2, p > .5$), with no interaction of cloze and context ($F=.1, p > .5$). In the
600-800 ms window, this two-by-two ANOVA revealed no main effects of CLOZE (F=1.1, p = .3) or CONTEXT (F = .3, p > .5), with no interaction of CLOZE and CONTEXT (F=.2, p > .5).

In the control comparison, a two-by-two ANOVA fully crossed CLOZE (high/low) and CONTEXT (V-Filler/N-Filler). In the 300-500 ms window, this two-by-two ANOVA revealed a main effect of CLOZE (F = 43.2, p < .001***) and of CONTEXT (F = 7.4, p < .05*), with no interaction of CLOZE and CONTEXT (F=2.6, p > .1). In the 600-800 time window, the main effect of CLOZE was only marginal (F=3.9, p = .06). There was, however, a main effect of CONTEXT (F = 28.3, p < .001***), as well as an interaction between CLOZE and CONTEXT (F=5.1, p = .03). This interaction appeared to be driven by the presence of a late positivity effect in the noun fillers but not the verb fillers.

Figure 1: ERP waveforms at electrode Cz and topographical maps for the difference between low cloze and high cloze conditions in critical items, Experiment 1.
2.6. Discussion

The goal of Experiment 1 was to determine whether the ‘missing’ N400 predictability effect often observed for NNV role reversals would re-emerge when the predictability manipulation is conducted with NVN role reversals and the predicted word is a noun. However, the surprising result was that a significant contrast in N400 amplitude was elicited in the manipulation of cloze across both NNV and NVN role reversals, and there was no indication that the effect was any smaller in the NNV case. In other words, where a large body of previous work has reported no N400 contrast on final verbs in role-reversed sentences, we found precisely such a contrast.

Taken at face value, the N400 results would seem to indicate that our NNV role-reversal materials inadvertently had different properties than the materials used in prior studies. However, given that these results were unexpected and contrast with substantial prior literature, it is important to determine that they replicate. It is also possible that the critical difference lay in
some property of our participant population or their interpretation of the task. To evaluate these possibilities more carefully, we next turn to a follow-up study in which we directly compared ERP responses to our NNV canonical and role-reversed stimuli with responses to an apparently similar set (from Chow et al. 2015) that did not elicit an N400 contrast.

Although the primary focus of the experiment was on the N400 response, we also reported results in the late positivity time-window, which has been of interest in some of the prior literature on role reversals. Although a late posterior positivity is often observed for role-reversals, we did not observe one for our critical items. Given recent accounts in which late posterior positivities are associated with implausibility detection (Van Petten & Luka, 2012), this could relate to the fact that our critical items were perhaps not as strongly implausible as the materials used in prior studies. In the current experiment low-cloze target items were judged implausible about 65% of the time compared to about 85% of the time for the fillers. However, we also observed an unexpected interaction in the late positivity for the filler items, where implausible nouns elicited a positivity and verbs did not, even though both types elicited equivalently high rates of implausible judgments. We return to this late positivity pattern in the General Discussion.

3. Experiment 2

The aim of Experiment 2 was to evaluate potential stimulus differences between the NNV items used in Experiment 1 and those used in Chow et al. (2016b, Exp. 1) that might drive differences in the extent to which argument roles impacted N400 responses at the verb. We used a 2x2
within-subjects design to compare N400 amplitudes to high or low cloze target verbs in sentences with embedded indirect wh-questions (NNV order) drawn from Chow et al. (2016b, Exp. 1) and Exp. 1 of the present study. The experiment design thus crossed cloze (high/low, corresponding to canonical/reversed sentences) and item set (Chow et al., 2016b, or Exp. 1 of the present study).

3.1. Materials

54 items were drawn from the present study’s Experiment 1 and were selected to match the 26% average high cloze and 0% average low cloze found in Chow et al. (2016b, Exp. 1). The remaining 54 items were drawn from Chow et al. (2016b, Exp. 1).² A further 108 filler sets were evenly split between high or low cloze target nouns or verbs (again with a 27% - 0% contrast between high and low cloze conditions). These were also drawn from Chow et al. (2016b) and the present study’s Experiment 1. These fillers were not exactly the same set used in Exp. 1 because of the slightly lower mean cloze value. The full set of experimental items can be found in the Supplemental Materials.

² Because Exp. 1 used Chow et al. 2016b as a model, 4 of the 30 sets used in Exp. 1 were in fact taken directly from Chow et al. 2016b. Although clearly these 4 sets could not have driven any differences between the two studies, for maximal consistency with the original experiments we assigned 2 of the 4 items to the Chow et al. 2016b item set and 2 of the 4 to the set drawn from the current study, Exp 1.
3.2. Procedure

Procedure was as in Exp. 1.

3.3. EEG Recording

EEG recording was as in Exp. 1.

3.4. Participants

Data from 24 participants were analyzed, and these participants were drawn from the same population as in Exp. 1 (17 female; average age 21.9 years, SD: 3.2). Data from a further 7 participants were collected but excluded from the final sample due to technical issues, artefact rejection rates above 40%, multilingualism, or poor accuracy scores on filler items in the plausibility rating task.

3.5. Results

3.5.1. Behavioural results

The behavioural results for the critical NNV items (see Table 3) were analysed as in Exp. 1. As previously, data from participants who scored lower than 80% accuracy on filler items were excluded from further behavioural and EEG analysis; this removed two participants’ data from
the data set. For the remaining participants, a linear mixed-effects model (calculated using the R packages lme4, Bates, Maechler, Bolker, & Walker, 2015, and lmerTest, Kuznetsova, Brockhoff, & Christensen, 2017) with fixed factors CLOZE, ORIGIN and the interaction of CLOZE and ORIGIN and random factor PARTICIPANT revealed a significant contrast of plausibility ratings by CLOZE ($t = -20.6, p < .01 ***$), but not by ITEM SET ($t = 0.25, p > .5$), and with no interaction between the two factors ($t = -0.67, p > .5$). It is also worth noting that unlike in Exp. 1, plausibility ratings were relatively similar across low cloze critical items and filler items at ~20%.

Table 3: Percentage of items judged “plausible”, Experiment 2

<table>
<thead>
<tr>
<th>Origin</th>
<th>Cloze</th>
<th>Mean % Judged “Plausible” (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ehrenhofer NNV</td>
<td>high</td>
<td>83% (10%)</td>
</tr>
<tr>
<td>Ehrenhofer NNV</td>
<td>low</td>
<td>25% (16%)</td>
</tr>
<tr>
<td>Chow NNV</td>
<td>high</td>
<td>84% (10%)</td>
</tr>
<tr>
<td>Chow NNV</td>
<td>low</td>
<td>22% (13%)</td>
</tr>
<tr>
<td>N-Filler</td>
<td>high</td>
<td>96% (20%)</td>
</tr>
<tr>
<td>N-Filler</td>
<td>low</td>
<td>21% (40%)</td>
</tr>
<tr>
<td>V-Filler</td>
<td>high</td>
<td>93% (26%)</td>
</tr>
<tr>
<td>V-Filler</td>
<td>low</td>
<td>19% (39%)</td>
</tr>
</tbody>
</table>

3.5.2. ERP results

ERPs were analysed as in Experiment 1. 15.3% of the final trial sample were rejected due to artefacts. Waveforms for electrode Cz and scalp maps for the 300-500 ms and 600-800 ms analysis windows are shown in Figures 3 and 4. The statistical analysis was identical to the analysis used in Exp. 1. Plots showing averaged waveforms at all electrodes can be found in the Supplemental Materials.
For test items, in the 300-500 ms window, the two-by-two ANOVA revealed no main effect of \textit{cloze} (F = 1.9, p = .18) or \textit{origin} (F = .03, p > .5), but there was a significant interaction of \textit{cloze} and \textit{item set} (F=9.7, p < .01*). Visual inspection indicates that this interaction was driven by the presence of a standard N400 cloze effect in the Ehrenhofer items, and the absence of an N400 cloze difference in the Chow items. In the 600-800 ms window, this two-by-two ANOVA revealed a main effect of \textit{cloze} (F=19.0, p < .001***) but not of \textit{item set} (F = 1.4, p > .2). In this lahere was a trend towards an interaction of \textit{cloze} and \textit{origin} (F=3.5, p = .08), which was most likely driven by the notably larger late positivity effect in the Chow items.

In the control comparison, as in Exp. 1, a two-by-two ANOVA fully crossed \textit{cloze} (high/low) and \textit{context} (V-Filler/N-Filler). In the 300-500 ms window, this two-by-two ANOVA revealed a main effect of \textit{cloze} (F = 53.8, p < .001***) and of \textit{context} (F = 6.8, p < .01**), with no interaction of \textit{cloze} and \textit{context} (F=.2, p > .5). In the 600-800 time window, there were main effects of \textit{cloze} (F=5.7, p < .05*) and \textit{context} (F = 18.1, p < .001***), but no significant interaction between \textit{cloze} and \textit{context} (F=2.0, p = .17). As in Experiment 1, there was a tendency for the negativity effect to persist longer in the V-fillers than the N-fillers, although in contrast to Experiment 1 no late positivity effect was visible for the N-fillers.
3.6. Discussion

Figure 3: ERP waveforms at electrode Cz and topographical maps for the difference between low cloze and high cloze conditions in critical items, Experiment 2.

Figure 4: ERP waveforms at electrode Cz and topographical maps for the difference between low cloze and high cloze conditions in filler items, Experiment 2.
In Exp. 2, we replicated the results of both Exp. 1 and Chow et al. (2016b, Exp. 1). We found that test stimuli drawn from the Chow et al. (2016b) elicited no N400 contrast, but did show a P600 contrast, between canonical and role-reversed stimuli. On the other hand, despite being cloze-matched with the Chow et al. (2016b) stimuli, test stimuli drawn from Exp. 1 elicited an N400 contrast between canonical and role-reversed conditions, but no P600 contrast.

The two sets of stimuli were matched on both cloze and plausibility judgments, yet yielded different ERP outcomes. This supports the surprising empirical result of Experiment 1 in which role-reversals generated an N400 effect. As this was an unexpected finding, in the next section we describe our exploration of a range of other contextual measures, in an attempt to discover any fine-grained differences in experimental materials that may have affected ERP responses without being reflected in offline cloze completion data.

4. Post-hoc stimulus analyses

NNV stimulus creation in Experiment 1 was constrained by the need to create sets of four sentences that crossed cloze (low/high) with word order. In Chow et al. (2016b), stimuli consisted of low/high cloze pairs, such that one order was paired with its high cloze verb completion, then reversed so that the same completion had a low cloze value. In Experiment 1, by contrast, the additional constraint of needing to cross word orders meant that stimuli could only be included if the high-cloze target for each order did not appear in the cloze completion data for the reversed order. This ensured that offline, each order generated a specific verb prediction that was not at all (or barely) predicted for the reverse order. Additionally, low-cloze
stimuli in the present study needed to have not only a low cloze probability for the target completion, but as far as possible, this low-cloze completion had to be implausible. We undertook four further analyses to evaluate possible differences in stimuli properties, which we tested as a comparison between Experiment 1 NNV stimuli and the materials from Chow et al. (2016b). However, it is worth keeping in mind that Chow et al. (2016b) used slightly different criteria in their computation of cloze responses, as they did not collapse synonymous responses into a single bin as was done in the current study.

Data from these analyses for all experimental items can be found in the Supplemental Materials.

4.1. Target cloze distribution

One possible explanation for the difference in N400 outcomes between the present study and Chow et al. (2016b) is that there may be differences in the target cloze values across stimulus sets. That is, the stimulus sets may differ in terms of how highly predicted the maximum cloze item is for each condition.

However, our estimates do not indicate strong differences in cloze across sets. Within each stimulus set, the average maximum cloze values predicted offline for canonical and reversed NNV contexts differ (Experiment 1 stimuli: $t = 25.69, p < .001***$, mean high cloze: 35%, mean low cloze: 1.4%; Chow et al. stimuli: $t = 24.6, p < .001***$, mean high cloze: 26.6%, mean low cloze: 0%; within Experiment 2: Ehrenhofer-based stimuli: $t = 17.56, p < .001***$, mean high cloze: 28%, mean low cloze: 1.1%; Chow-based stimuli: $t = 13.9, p < .001***$, mean
high cloze: 28%, mean low cloze: 0.4%). Within Experiment 2, a two-way ANOVA listing items as random intercepts confirmed a significant effect of cloze (F = 168.65, p < .001***) but no effect of stimulus set (F = 1.3, p > .05). Although there could be smaller differences in maximum cloze values in each stimulus set that do not meet the criteria for statistical significance, we find no evidence of differences that are strong enough to account for the contrasting presence of an N400 effect.

4.2. Entropy

We next took a broader view of cloze completion data, and asked whether there were differences in the probability distribution of completions beyond the probability of the most frequent completion. If the most frequent offline completion is far more frequent than any of the other completions, this might yield a different online prediction than when the most frequent offline completion is only marginally more frequent than other completions for this sentence fragment. We computed entropy in the terms of Shannon (1948). Within each stimulus set, the distribution of entropy values differed statistically between word orders (Experiment 1 stimuli: $t = 3.9, p < .001***$, mean entropy in high cloze: 3.7, mean entropy in low cloze: 3.25; Chow et al. stimuli: $t = -3.6, p < .001***$, mean entropy in high cloze: 3.39, mean entropy in low cloze: 3.9). A two-way ANOVA showed an effect of cloze (F = 20.9, p < .001***), but no effect of stimulus set (F = 2.67, p > .1), suggesting that any entropy differences between high and low cloze conditions were not drastically different across stimulus sets, and were unlikely to be the cause of the observed N400 differences.
4.3. Frequency

We next compared the log frequencies of target verbs across experiments, based on the well-established finding that N400 amplitudes are impacted by target word frequency (Van Petten & Kutas, 1990) and the possibility that more frequent words might be generated more quickly in prediction. We identified the number of unique target verbs appearing in each condition across the two stimulus sets, and found that both stimulus sets used similar numbers of unique verbs (42 in Experiment 1 stimuli; 50 in verb targets for the stimuli used in Chow et al., 2016b) and obtained verb frequency estimates from the 560-million-word Corpus of Contemporary American English (COCA, https://www.wordfrequency.info/). In Experiment 1, the mean log frequency of the verbs we used was 8.76. This was similar to the mean frequency of verbs used on Chow et al. (2015), which was 8.92, and there was no statistical difference in log frequencies between these two sets of verbs ($t = -0.89, p > .3$). In addition, within Experiment 1, there were no contrasts between log frequency of the verbs used in canonical and reversed orders (log frequency in canonical orders: 9.06, in reversed orders: 8.47, $t = 1.57, p > .1$). Within Experiment 2, there were no contrasts in the means of target log frequency distribution between the two sets of stimuli ($t = -0.8, p > .4$), mean log frequency of target verbs in Ehrenhofer-based stimuli: 8.64, mean log frequency of target verbs in Chow-based stimuli: 8.92). Thus, the observed N400 differences are unlikely to be due to verb frequency differences.
4.4. Subject-verb cosine relationship

As outlined above, examination of the sets of stimuli used in Chow et al. (2016b) and in Experiment 1 did not yield measurable differences between the two stimulus sets in terms of differences in entropy or cloze probability distributions. Although many experiments have observed a relationship between N400 amplitude and cloze probability, the role-reversal literature shows that this relationship is unlikely to be direct or unmediated. Existing models comparing the offline outcome of cloze completion tasks to online prediction mechanisms (Staub et al., 2015, 2010) have yielded useful insights about the possible dynamics between competitors in the process of generating a prediction, but have not formulated clearly what characteristics make a competitor strong (and therefore more likely to be sampled as the winning competitor) or fast (and therefore more likely to reach a certain activation threshold before other competitors).

Ettinger (2018) proposes as one possibility the measure of subject-verb cosine relationship (SVCR). The distributional properties of a word in a corpus may be expressed as a vector in a vector space model (VSM). Many different versions of VSMs are widely used in current machine learning and natural language processing tasks. The cosine relationship between two word vectors provides a measure of the extent to which they co-occur in the same linguistic contexts. Ettinger used pre-trained GloVe vectors (Pennington et al., 2014) of 50 dimensions, trained on English Wikipedia and Gigaword corpora, to calculate the cosine between vectors for each embedded subject noun (... which bull the cowboy... or which cowboy the bull...) and the vector for the corresponding target verb (e.g. had ridden). Our and Ettinger’s (2018) calculations showed a similar SVCR between embedded subjects and verbs across high and low cloze.
conditions Chow et al.’s (2016b) stimuli (high cloze mean SVCR: .35, low cloze mean SVCR: .32, \( t = -0.3, p > .75 \)), but found that there was a greater discrepancy in subject-verb cosine relationship between the high and low cloze stimuli in Experiment 1, although this contrast did not reach significance (high cloze mean SVCR: .32, low cloze mean SVCR: .28, \( t = 1.48, p > .1 \)). In Experiment 2, this difference is significant within Ehrenhofer-based stimuli (high cloze mean SVCR: .35, low cloze mean SVCR: .27, \( t = 2.3, p < .05 \)) but not in the items taken from Chow et al. 2016b (results reported above).

4.5. Plausibility: a follow-up experiment

Finally, we attempted to quantify the difference in plausibility of the high cloze verb completion for each context order, with a view to establishing whether there was any difference in the strength of this contrast across the two stimuli sets. Recall from the Materials section of Experiment 1 that we imposed an additional constraint on stimulus creation: for every reversed order, there must be a distinct high cloze target verb that does not appear as a target verb in the canonical order. In other words, both orders make relatively clear predictions, and the predictions for the two orders do not overlap. This was a step further than the already very highly controlled NNV stimuli prepared by Chow et al. (2016b), who required that the high cloze target verb in the canonical order not appear as a target in cloze data for the reversed order, but perhaps crucially, did not impose the converse requirement as we did. In other words, if Chow and colleagues decided to use the canonical item ‘which villager the ghost had scared’ they would confirm that the reversed ‘which ghost the villager had scared’ had low cloze and plausibility values, and in
collecting these ratings, would also learn the highest cloze verb for the reversed item (here, *seen*). However, they did not evaluate responses or ratings when this reversed high cloze verb was substituted in the ‘canonical’ frame (*which villager the ghost had seen*).

We expected robust plausibility rating contrasts between high and low cloze completions within stimulus sets (i.e. within the Experiment 1 stimuli and also the Chow et al. stimuli). However, if there is an underlying difference between the stimulus sets in the degree to which order (canonical/reversed) predicts a role-specific verb, we expected this to be reflected as a between-stimulus set difference in plausibility ratings when high-cloze completions for reversed orders were applied to the canonical order.

We conducted an online plausibility rating task on Amazon Mechanical Turk, comprising a total of 60 participants who rated the full set of NNV stimuli from Experiment 1 and the corresponding materials in Chow et al. (2016b). Cloze completion data from norming studies in Chow et al. (2016b) were used to create an additional set of conditions, such that the resulting Chow stimuli conditions matched those used in Experiment 1, in fully crossing not only cloze (low/high) but also word order (canonical/reversed; see Table 4). Note that in order to increase sensitivity, this follow-up task differed from the within-experiment plausibility judgment which was a binary judgment (plausible/implausible); here we instead used a 7-point Likert scale.

Table 4. Stimulus conditions in ad-hoc plausibility rating experiment, with mean plausibility ratings (Likert scale of 7)

<table>
<thead>
<tr>
<th>Order</th>
<th>Cloze</th>
<th>Example</th>
<th>Set</th>
<th>Mean plausibility rating (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>canonical</td>
<td>high</td>
<td>… which cowboy the bull had <em>gored</em></td>
<td>Exp. 1</td>
</tr>
</tbody>
</table>
b reversed low … which bull the cowboy had gored Exp. 1 2.5 (1.9)
d canonical low … which cowboy the bull had ridden Exp. 1 3.1 (2.0)
d reversed high … which bull the cowboy had ridden Exp. 1 6.0 (1.3)
a canonical high … which villager the ghost had scared Chow 6.3 (1.1)
b reversed low … which ghost the villager had scared Chow 2.5 (1.8)
c *canonical low … which villager the ghost had seen Chow 3.7 (2.2)
d *reversed high … which ghost the villager had seen Chow 5.8 (1.5)

The conditions in italics were not included in Chow et al. (2016b). We created them for this plausibility experiment from Chow and colleagues’ original cloze data, using the maximum cloze target.

The resulting 120 stimulus items (60 items of 4 conditions per set) were distributed across four experimental lists in a Latin square design which ensured that each participant saw only one condition per item. Participants were asked to individually rate sentences on a Likert scale ranging from 1 (very implausible) to 7 (very plausible). One participant was excluded due to illicitly repeating the task. For all participants, two item sets were excluded for technical reasons. For an additional 14 participants, a further three item sets were excluded for technical reasons. For each item in each condition, there were 13-14 ratings, spread among a total of 55 participants.

We first tested the contrast in plausibility ratings between reversed order high cloze completions and their low cloze counterparts between the two stimulus sets. In simpler terms: we analysed the plausibility rating contrast between which bull the cowboy had ridden/#which cowboy the bull had ridden (stimuli from Exp. 1), which ghost the villager had seen/#which villager the ghost had seen (stimuli created to round out the experimental paradigm in Chow et
al., 2015), and evaluated whether the contrast in plausibility ratings was significantly larger within the Exp. 1 stimuli than the Chow-based stimuli.

We ran a mixed-effects model to investigate the between-stimuli set contrast in plausibility ratings between reversed contexts paired with high-cloze verbs (condition d) and the corresponding canonical context paired with that same verb to yield a low-cloze completion (condition c). The fixed effects were the interaction of condition (c/d) and stimulus set (Experiment 1/Chow et al., 2016b). We also included random intercepts for items and participants. Statistical results showed an effect of condition (high-cloze reversed order (d) was rated as significantly more plausible than the low-cloze reversed order condition (c), \( t = 36.6, p < 0.001*** \), an effect of stimulus set (the difference in plausibility rating between high and low cloze conditions was smaller in Chow et al. 2016b items than Experiment 1 items, \( t = 3.9, p < 0.001*** \)) and an interaction of condition and stimulus set (\( t = -7.37, p < 0.001*** \)).

In addition, we tested the size of the contrast in plausibility ratings between the Chow and Experiment 1 materials for canonical high cloze stimuli (condition a) and their reversed counterparts (condition b). We repeated the statistical approach outlined in the last paragraph to compare between-item set plausibility ratings for canonical orders with high-cloze verb (condition a) and the reversed order paired with that same verb to yield a low-cloze completion (condition b). The fixed effects and intercepts replicated the design of the previous model. Here, there was again an effect of condition, with low-cloze completions rated as significantly less plausible than high-cloze completions (\( t = -57.7, p < 0.001*** \)). However, there was no effect of stimulus set (\( t = -0.268, p > .5 \)), nor was there any interaction between stimulus set and condition (\( t = 0.232, p > .5 \)).
These statistical analyses support our hypothesis: the two sets of stimuli from Experiment 1 and Chow et al. (2016b) were well-matched in terms of the plausibility contrast between high and low cloze conditions for the canonical order (a vs. b), but not in terms of the high and low cloze conditions for the reversed order (c vs. d). This suggests that the implausibility constraint in Experiment 1 stimuli led to the selection of stimuli that gave rise to more distinctive, non-overlapping offline prediction profiles, which may have contributed to the novel finding of a contrast in online prediction profiles.

5. General Discussion: Argument roles in prediction and retrieval

Experiments 1 and 2 were designed to investigate the impact of argument role information on prediction. We based the design of Experiment 1 on a sizable body of EEG studies that have reported no N400 contrast between canonical and role-reversed sentences (... which cowboy the bull had gored, where bull is the Agent of goring; #which bull the cowboy had gored, where cowboy is the Agent of goring). The starting-point for our experimental design was a suggestion from Chow et al. (2016) that the reason for the lack of N400 contrasts in the studies which used NNV word orders is that verbs contain information that is crucial to comprehenders being able to rapidly use argument roles in prediction. In Experiment 1 we therefore examined whether role reversals elicited N400 contrasts on the final noun in NVN contexts, and included the standard NNV contrast as a control comparison. To our surprise, we found that N400 amplitudes were higher in role-reversed than canonical contexts in both conditions, in other words failing to replicate the previous NNV results. In Experiment 2, we showed that this surprising result is
likely due to stimulus properties, as we replicated the N400 contrast in our own NNV stimuli and, in the same participants, replicated the absence of N400 contrast in the NNV stimuli from a prior study (Chow et al. 2015).

To our knowledge, this is the first time that an N400 contrast has been found in NNV role-reversed sentences in English. Among many previous NNV role-reversal studies in other languages, only a handful have reported N400 contrasts when the critical verb immediately follows the nouns. Because our study was not designed with this goal in mind, we cannot give a definitive account here of these surprising results. However, this finding clearly offers an interesting new avenue for investigating the online processing of argument role information. In what follows, we use our findings as a starting point for a more fine-grained discussion of how predictions are generated on the fly, and why argument roles sometimes may and sometimes may not impact predictions at short processing latencies.

5.1. Previous cases of N400 role reversal effects

To our knowledge, no previous studies have reported NNV role-reversal N400 effects in English. However, Bornkessel-Schlesewsky, Schlesewsky and colleagues have observed such N400 effects in German (Schlesewsky & Bornkessel-Schlesewsky, 2009) and in a subsequent cross-linguistic comparison, also reported such effects in Turkish and a subset of tested cases in Mandarin and Icelandic (Bornkessel-Schlesewsky et al., 2011). They argue that these cross-linguistic differences in the presence or absence of role-reversal N400s are associated with word order flexibility: speakers of languages with more word order flexibility make use of ‘sequence-independent’ combination processes that impact the N400, but speakers of languages with stricter word order make use of ‘sequence-dependent’ processes that do not.
While the cross-linguistic dissociation reported by Bornkessel-Schlesewsky et al. (2011) is interesting, several design factors make it somewhat challenging to directly integrate these findings with the current ones. First, as these authors assumed a view of the N400 focused on plausibility rather than predictability, they did not collect cloze data on the contrasts reported, so that it is unclear whether the contrasts were well-matched for predictability across the languages that show different results. Second, some of these cases include additional complexities, such as temporary ambiguity in the role assignment (e.g. due to morphological syncretism) which is resolved by an agreement cue at the main verb or auxiliary (in German), or role-assignment that is disambiguated by the lack of explicit case-marking (in Turkish). As Bornkessel-Schlesewsky et al. (2011) discuss, these complexities may drive different processing strategies that explain some of the cross-linguistic ERP differences observed. However, these differences in processing strategies relate to language-specific grammatical properties. In providing an explanation for why two sets of English stimuli with the same grammatical properties yield different N400 patterns, we must therefore look beyond an account based on such properties.

5.2. How could the format mismatch account explain the present results?

In the Introduction, we described one existing type of account of the N400’s argument role insensitivity from Chow et al. (2016), which we will call the ‘format mismatch account’. At the most general level, the format mismatch account hypothesises that predictions based on role information in NNV sequences are difficult because the ‘format’ of the role information encoded in the NN sequence is not immediately compatible with the format in which information about
verbs is stored in long-term memory, making this type of information insufficient to rapidly retrieve verb predictions. As discussed, some intuitions about why this might be are the fact that the structural position of ‘subject’ (which is what is available) may not bear a one-to-one mapping to a single thematic role, and the possibility that conceptual event representations are not encoded in terms of generalised thematic roles. In order to explain the re-emergence of N400 effects at a delay — and the fact that participants do converge on similar ‘predicted’ offline completions — this kind of account must posit that even though rapid item-to-item retrieval is not available to feed predictions, a slower process can still accomplish the same objective.

One particular instantiation of such a format mismatch account suggested by Chow et al. (2016a) is that verb prediction involves two sequential mechanisms. First, a parallel search mechanism retrieves all events (or all verb lemmas) associated with the preceding nouns, regardless of argument role assignment, which yields an ordered list of events ranked by activation. Second, a slower serial mechanism sequentially checks items in this list, ascertaining for each event whether a candidate matches the argument role assignment of the preceding nouns, until a match is encountered. This account thus attributes the lack of N400 contrast when the verb appears shortly after the arguments to the initial retrieval of non-role-specific verb candidates, and the emergence of the N400 contrast at longer processing latencies (Chow et al., 2018) to the conclusion of the role-specific generation mechanism.

This version of the format mismatch account is unable to explain the present results, because it posits that role information is incorporated relatively slowly in all cases. If role information impacts the N400 in some cases but not others, when timing, structure, and predictability are all held constant, this version of the format mismatch account has no
a straightforward way to account for the variability we find between Experiment 1 and prior findings, as well as the variability we found within Experiment 2. Put another way, the assumption inherent in Chow et al. (2016a) is that the lexical-associative mechanism generates all candidates that are consistent with either argument role assignment. That is, regardless of whether the incoming sentence fragment is *which villager the ghost had*… or *which ghost the villager had*…, the lexical-associative mechanism always generates both *haunt* and *see*, thus yielding no contrast in N400 amplitude on *haunt* when it follows a role-reversed context. This is consistent with Chow et al.’s (2016b) maximum offline cloze values of ca. 26% in canonical and 22% in reversed sentence fragments, which show that when participants generate completions offline, they are able to produce distinct verb predictions for each order.

Adapting the format mismatch account so as to explain the results of Experiment 1 would require the first ‘rapid retrieval’ process to be adjusted such that it can sometimes bias towards a role-appropriate candidate, without directly using argument role information in the process. This would preserve the format mismatch account’s hallmark assumption that argument role information is impossible to use at short processing latencies in the absence of a verb. For instance, activation of verb candidates might differ depending on the order of the nouns, e.g. as a result of decaying activation over time. That is, at the time of testing predictions for verbs, the verbs activated by the first noun might collectively be less activated than the verbs activated by the second. If *bull* generates verbs including *gore*, but the activation of *gore* decays while the comprehender encounters *cowboy* and generates verbs including *ride*, this may be a source of higher activation for *ride* (leading to a low N400 amplitude for *ride* in *which bull the cowboy had…*) and lower activation for *gore* (leading to a high N400 amplitude for *gore* in *which bull
the cowboy had…). This would yield an N400 contrast like the one observed in Experiment 1 and the related stimuli used in Experiment 2. However, if the rapid retrieval mechanism can be biased towards certain role-specific event predictions, it is unclear why the same mechanism would not yield an N400 contrast in all role reversals, which is inconsistent with the previous results in the literature.

Alternatively, it would be possible to adjust the search mechanisms’ operating speed, e.g. by making the serial search process slow in some cases and fast in others. This could be achieved in one of two ways. Either by assuming a way for the role-specific mechanism to engage earlier or later in the prediction process, so as to yield the timing contrasts between Chow et al. (2018) and the present study. Or by assuming that the role-specific mechanism itself operates at varying speeds. If it is the case that the lexical-association mechanism generates both haunt and see, and that these are equally strong predictions, then both of these event predictions should be at the top of the list for the serial search mechanism to check against the preceding context’s argument role assignments. It is necessary to assume that any operation of the role-specific mechanism is slow in order to yield the timing contrasts found in Chow et al. (2018), which requires that checking and rejecting a verb candidate prior to moving down the list and identifying a correct item takes long enough that this mechanism’s output cannot impact prediction at short latencies. If we assume that the role-specific mechanism can begin its processing earlier under specific circumstances, there would have to be some triggering factor during the lexical-association process that would allow the role-specific mechanism to get a head start in checking candidates in the materials for the present study, but not for Chow et al.’s (2016b) stimuli. Or, if we assume that the role-specific mechanism is inherently variable in its processing speeds, it would be
necessary to posit a reason why it would have been systematically slow in processing verb candidates in Chow et al. (2018), but not in the present study. Again, given that exploration of the possible distinctions in the cloze completion data for each stimulus set yielded no systematic difference, it seems that any contrast in the role-specific mechanism’s operating speed must be caused by some underlying difference that is not reflected in offline cloze data.

In summary, updating the existing version of the format mismatch account repeatedly runs into the same difficulty: how to adjust this account’s mechanisms such that they would generate verb predictions that yield an N400 contrast for the stimuli used in the present study, but no such contrast for the stimuli used in prior work. However, this account’s strength lies in its ability to explain the timing contrasts found in prior work by Chow et al. (2018) and Momma (2016).

5.3. How could a probability-based account explain the present results?

Kuperberg (2015) suggests an alternative account for the absence of N400 role-reversal effects, which we will call the ‘probability-based account’. The probability-based account supposes that the contrast between role-blind and role-specific prediction generation lies in the “reliability” of the bottom-up information that is used to infer the underlying event. In Kuperberg’s (2015) account, the choice between using the full set of available information (i.e. nouns including their word order, acting as proxy for argument roles) or a reduced subset of available information (nouns only, excluding word order information) is determined by which of these sets of information yields greater certainty in prediction. This assumes a contrast in the events that can be predicted from two nouns: for instance, if the knowledge that ghost and villager are the
arguments yields a very high-probability prediction for serving events, but the role-specific knowledge \textit{villager-AGENT} and \textit{ghost-THEME} yields only a diffuse prediction for many possible verbs, then the role-independent prediction would be more strongly weighted. It may seem surprising that more specific contextual information is a less reliable as a predictor. But this is possible if, for example, events involving a \textit{ghost} and a \textit{villager} are dominated by events where \textit{ghost} is the subject and \textit{villager} the object, and that configuration is strongly constraining, e.g., predicting \textit{haunt}, while the reversed configuration of arguments is rarer and less constraining.

Chow et al. (2016) argued that this contrast was not borne out in the role-reversal stimuli that showed the lack of an N400, based on offline cloze completion data. In the next section, we discuss in greater detail why cloze data may be too blunt an instrument for determining the predictability of a given completion and how that may affect N400 amplitudes. For now, however, if we assume that the NNV contexts used in the present study (but not those used in Chow et al., 2015) generated competing verb candidates at the lexical association stage of processing, the probability-based account can explain the contrast in N400 outcomes for the present study and past work. According to this explanation, in the present study, assuming that the assuming that the role-specific contexts are more constraining, in both canonical and reversed orders, than the role-neutral contexts, then the role-specific information should be used to generate verb predictions, resulting in the N400 contrasts elicited by the materials from the current study.

The probability-based account is not designed to explain the timing phenomena observed in Chow et al. (2016b) and Momma (2016), and this is a significant shortcoming. In principle, however, one could appeal to the possibility that the verb predictions generated by a
‘bag-of-arguments’ lexical-associative search mechanism might not be generated all at once but might instead have different timing profiles. For example, if verb candidates that are each strongly associated with one or the other word order are generated quickly, creating immediate competition, then this could trigger the choice of the role-specific candidate at short processing latencies, yielding N400 contrasts, as in the present study, whereas other cases might generate a mix of faster and slower verb predictions, with the consequence that a role-blind candidate might be unopposed at short processing latencies, yielding a lack of N400 contrast, while role-specific candidates might be activated slightly later, resulting in an N400 contrast at longer latencies.

5.4. Prediction, competition and candidate generation

Investigating how argument role information impacts which predictions are generated on the fly begins with investigating how candidates for upcoming sentence material are generated even in contexts that do not involve argument role information. For decades, researchers have relied on the cloze task (Taylor, 1953) as an offline means to probe the expectations readers generate as they complete a sentence task, and N400 amplitude has fairly consistently been shown to track offline cloze probabilities (Kutas & Hillyard, 1980; Kutas & Hillyard, 1984; Kutas & Federmeier, 2000; Lau et al., 2008). Mediating between the format mismatch and probability-based accounts of verb prediction relies heavily on argumentation around the relationship between the online and offline prediction of verbs in NNV contexts. Kuperberg’s account rests on a supposition that Chow et al.’s (2016b) stimuli asymmetrically predicted events, with one order more likely to generate a strong verb candidate than the other. Chow et
al.’s (2016a) rebuttal hinges on the fact that their offline cloze measures reflect each noun order’s equally strong ability to generate a verb candidate. The authors implicitly assume (following a long tradition) that offline prediction parallels online prediction. However, experimental results concerning role reversals challenge this assumption, either by showing that offline cloze contrasts do not entail online N400 contrasts (Chow et al., 2016b), or by highlighting variability in the extent to which cloze matches online prediction (as in the present study). This section therefore explores what generation and selection processes cloze probability actually reflects, and how this might in turn impact experimental findings on prediction to yield the divergent patterns (N400 contrasts or lack thereof) that we observe here and in prior literature.

Cloze probability collapses at least three distinct sources of contrast which may have different impacts on N400 amplitudes. It is uncertain whether N400 amplitude reflects the degree of constraint in the sentence context (the extent to which the sentence context restricts the number of possible continuations) or the degree to which a given word is predicted in context (Van Petten & Luka, 2012). A sentence context might be highly constraining, such that a target with a 40% completion rate is a relatively poor completion. To adapt a well-known example, *He mailed the letter without a stamp/an address* might have cloze values of 60% for *stamp* but 40% for *address*. Yet 40% might be the highest cloze completion in a less constraining sentence like *She left the house without her coat/shoes/umbrella*. In addition, the distribution of the competing completions could matter. A target might have a cloze probability of 40% in context, but it might be the highest-probability completion among two others each comprising 30% of the remaining completions, or it could be the winning candidate amongst ten other completions that are each generated with a 6% probability. These distinctions are difficult to control for in experimental
investigations. However, a promising account by Staub (2015) addresses the issue of the distribution of completion candidates.

There is evidence that the N400 elicited by an unexpected word is unaffected by how strongly constraining the sentence context is (Kutas & Hillyard, 1984; Federmeier et al., 2007). However, these studies provide less evidence on the impact of sentence constraint on moderately predictable words.

Staub et al. (2015) argue that the distribution of cloze responses reflects a process in which different lexical candidates ‘race’ to be the first to reach a threshold level of activation. Competitors whose activation rises faster are more likely to win the race, and hence more likely to be selected as the winner in a close task. The authors’ computational simulations demonstrate how a candidate’s speed relative to other candidates affects the dynamics of candidate selection, i.e., which completion ultimately has the highest cloze probability. The authors presume that various factors affect the speed with which a lexical competitor is activated, including the word’s conditional probability in the context, its lexical frequency, and its degree of lexical association to words in the context. It is left unspecified how lexical activations rise towards threshold, and whether activations rise at the same rate, or whether some factors drive different activation gradients than others. If activations are differentially affected by different factors at different points in time, then it might be possible to capture the timing effects found by Chow, Momma, and others.

We mention this account here because it explores the possibility that different types of verb candidates might be more or less predicted in online and offline measures. In principle a verb candidate whose activation rises steadily throughout the prediction process might might end
up with the highest cloze probability offline, but not surface as a favoured prediction early in the process. By contrast, a verb candidate whose activation initially rises quickly might surface early as a favoured prediction in online measures, resulting in facilitated N400 amplitudes, but might not garner the highest cloze probability in offline measures.

We are not the first to highlight these issues. Hoeks et al. (2004) conducted an offline cloze norming task revealing that participants found it more difficult to come up with completions for reversed sentences, e.g. a sentence in which *javelins* is the agent of an event that has *athletes* as its theme. We have no data on the ease of generating sentence completions from the 900+ participants in our cloze norming studies, and even if these were available, it would still be necessary to identify the underlying factors why one word order would more easily yield verb predictions than another. If it were possible to identify such a factor, however, and demonstrate that at short latencies, the stimuli in Chow et al. (2016b) indeed yielded event predictions that only matched one word order, while the stimuli in the present study yielded event predictions compatible with both word orders, the probability-based account could explain why the role-specific verb candidate was predicted in the present study, resulting in a contrast in N400 amplitude. If the search results of the lexical-association mechanisms yield two probable events (e.g. generating both *serve* and *tip* following the context … *which waitress the customer* …), this reduces the overall certainty with which either of those events is predicted. However, the generation process based on the full set of available cues (i.e. nouns including argument role information) yields only one event prediction, which is therefore predicted with higher certainty. It remains to be explained why, in the case of the N400’s short-term insensitivity to argument role information, the generation process initially generates verb candidates that are incompatible
with the preceding context’s argument roles, yielding a facilitated N400 for role-reversed verb targets, and how these inappropriate candidates are nonetheless suppressed enough that they do not appear at all in offline cloze completion data.

Another reason to suspect that cloze is too blunt an instrument for determining online predictions is that among all of the post-hoc analyses we conducted in attempting to pinpoint a contrast in the Chow et al. (2016b) materials and the present materials, all measures based on cloze data and completion distributions failed to find differences between the two stimulus sets. However, the contrast in subject-verb cosine relationship (as calculated by Ettinger, 2018) between the stimulus sets inspired us to explore how this corpus-based computational measure might align with sentence processing.
5.5. Encoding structural information in noun entries: The promise of SVCR

As outlined in Section 4, one measure that found a reliable contrast between the stimuli developed for the current study and by Chow et al. (2016b) is Ettinger’s (2018) calculation of the subject-verb cosine relationship for the two sets of stimuli. Her results suggest that the embedded subjects in Chow et al. (2016b) were found in similar contexts as target verbs across both the canonical and reversed conditions, whereas there was a greater divergence for the same measure in the present study’s NNV stimuli. For Chow and colleagues’ stimuli, this means that the embedded subject of a canonical sentence (e.g. ghost) was just as likely to co-occur with haunt as the embedded subject of a reversed sentence (e.g. villager). By contrast, in the stimuli used in the present work, bull was more likely than cowboy to co-occur with gore.

Vector space similarity is a measure of whether words share distributional similarities. For a pair of words with a high cosine distance, i.e., high similarity, this means that those words tend to occur in similar contexts. It does not mean that the words tend to co-occur with each other. In the case of the word vectors used by Ettinger (2018) the vectors reflect limited information about the structures where the words appear. However, we speculate that subject-verb pairs that show a high cosine similarity with each other may be good agent-verb pairings. Hence it is possible that if the SVCR values are similar between the high and low cloze conditions, as in the Chow materials, then we may tentatively assume that the subjects are good agents for the verb in the high and low cloze conditions alike. Meanwhile, if the SVCR values differ between the high and low cloze conditions, as in the materials from the current study, then we may assume that one noun is a better agent for the verb than the other.
One possibility is that nouns are stored with a distribution of probabilities over their likelihood of being the agent of a specific verb, and that what drives N400 amplitude variation in these studies is primarily the relationship between the noun-as-agent and the verb. Influential work in the syntactic ambiguity resolution literature has argued that probabilistic information about syntactic frames and argument structure is stored with individual verbs (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Kello, 1993; Trueswell, Tanenhaus & Garnsey, 1994; McRae, Ferretti, & Amyote, 1997). In parallel with the verb complement case, then, nouns could be stored with probabilistic information about what verbs they occur with as agents. For example, bull and cowboy type nouns from the present experiment’s stimuli might be stored with probabilities like those in Table 5.

Table 5: Mock distribution of nouns over argument roles with specific verbs (Experiment 1 stimuli)

<table>
<thead>
<tr>
<th>-as-Agent</th>
<th>Bull</th>
<th>Cowboy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gorer</td>
<td>Rider</td>
</tr>
<tr>
<td>Bull</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trampler</td>
<td>Lassoer</td>
</tr>
<tr>
<td></td>
<td>Kicker</td>
<td>Whipper</td>
</tr>
<tr>
<td>Rider</td>
<td>30%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Rider</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Rider</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Rider</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Rider</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Rider</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>
Kicker 10%
Gorer 0%
...
...

On the other hand, *villager* and *ghost* type nouns might be stored with probabilities similar to the ones in Table 6.

Table 6: Mock distribution of nouns over argument roles with specific verbs (Chow et al., 2015 stimuli)

<table>
<thead>
<tr>
<th>-as-Agent</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haunter</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Scarer</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Seer</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Exorciser</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Villager</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seer</td>
<td>35%</td>
<td></td>
</tr>
<tr>
<td>Exorciser</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Haunter</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Scarer</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

The primary contrast between the two probability distributions is that for the Experiment 1 stimuli, there is a clear contrast between the identity of the verbs that are good fits for either noun in the agent role, whereas for the Chow stimuli, there is greater overlap. In generating verb predictions, this could mean that the initial lexical-associative mechanism generates all verb candidates through this pattern of distributions. In the Chow et al. (2015) stimuli, this results in
the generation of a similar group of verbs, resulting in facilitated N400s for these targets even when they do not match the role assignment as well as other candidates. These can only be differentiated at later processing stages through the activity of the role-specific serial search mechanism, which checks whether the argument role assignments of both nouns fits a given verb candidate. In the Experiment 1 stimuli, on the other hand, the distribution over probabilities of argument roles with specific verbs generates different pools of verb candidates in the lexical-associative generation mechanism, as there is less overlap in terms of the agent-of-verb candidates stored with the two nouns. This means that role-specific verb predictions (e.g. in the case of bull, this would be gore, from the high probability of Gorer) will have higher levels of activation than non-role-specific verb predictions (e.g. ride, from the low probability of Rider). This would result in a contrast in facilitation for ride vs. gore in the which cowboy the bull had… prediction context. The later stage of checking via the role-specific search mechanism then serves to confirm, rather than rectify, the verb prediction.

This adjustment to the Chow et al. (2016) account preserves some of its key features, while addressing those aspects of the account that were unable to explain the results of Experiment 1. The important insight from that account is that argument role information may be difficult to use in prediction if it is divorced from verb information.

Note that in this new account, it is not the case that nouns are stored with a distribution over abstract argument roles, without verb information (e.g. bull [80% Agent, 20% Patient], etc.). If that were the case, we should expect these assignments to always dominate verb generation, such that no matter the actual argument role assignment, the verbs generated at the lexical-associative stage would be mostly compatible with bull-Agent (even in the context of
which bull the cowboy had…). If this were the case, we should expect no N400 contrast between the canonical and reversed conditions, as the same verb would be generated in either order. However, by positing that nouns are stored with combined verb-argument roles (Gorer, Rider, Hauntee, etc.) we provide the possibility of a more nuanced distribution over likely upcoming verbs, which is able to take the current argument role assignment into account from the bottom-up input, but is also sensitive to probabilistic variability in the verbs that these nouns co-occur with.

One important caveat to consider in further developing this account is the results of the “argument substitution” conditions in Chow et al. (2016b). Chow and colleagues’ low-cloze conditions came from two different sources. In the role reversal manipulation, target verbs had a cloze probability of 0% due to the reversal of the preceding nouns (… # which waitress the customer had served). In the substitution manipulation, target verbs had a low cloze probability because of the replacement of the first NP with one that could not be a participant in the event denoted by the verb (… which tenant / # realtor the landlord had evicted). In the substitution manipulation, the identity of the subject NP was held constant across both conditions, yet there was a strong N400 contrast between these two conditions. If subject-verb cosine relationship were the primary driver of N400 amplitude, this would predict a lack of contrast when the subject is held constant between conditions. However, the results from Chow et al. (2016b) may complement, rather than undermine, the subject-verb cosine relationship account of the N400 contrast in Experiment 1. Under the account of verb prediction advanced in Chow et al. (2016a, 2016b), an initial stage lexically pre-activates all verbs that are compatible with both NPs in the clause. In the case of the substitution conditions, the two conditions use different lexical items
(tenant + landlord vs. realtor + landlord). While the subject itself (landlord) is constant across these two conditions, the identity of the participants available for prediction leads to the generation of disparate verb predictions, and this part of the verb generation process may be the source of the N400 contrast. On the other hand, in the reversal conditions in Chow et al. (2016b), as well as the NNV conditions of Experiment 1, the same noun phrases are used, but in a different word order. In these cases, the difference in subject-verb cosine relationships among conditions is a further contributor to the amplitude of the N400, such that only a contrast in subject-verb cosine relationships between conditions (as in Experiment 1) leads to a difference in N400 amplitudes. While there were N400 contrasts on target verbs in both the Chow et al. (2016b) substitution conditions and the NNV reversal conditions in Experiment 1, they could plausibly be derived from different sources: the disparate identities of the pair of NPs being used to generate verb predictions in the Chow et al. (2016b) substitution conditions, and the disparate subject-verb cosine relationship in the reversal conditions in Experiment 1.

5.6. Late positivities and component overlap

Although our study was designed to focus on N400 sensitivity to role-reversals, our results also indicate variability in the amplitude of the late positivity that is more consistently reported for role-reversals. Most notably, in Experiment 2 we observed a much smaller late positivity for role-reversals in our stimuli than in the Chow stimuli, which resulted in a marginally significant interaction in the 600-800ms time-window. We similarly observed no significant late positivity for role-reversals in our stimuli in Experiment 1. We also observed an unexpected late difference
between the noun and verb fillers in Experiment 1, in which the semantically anomalous noun fillers elicited a late positivity but the semantically anomalous verb fillers did not.

The differences in the amplitude of the late positivity between the Chow and Ehrenhofer stimuli raise the question of whether component overlap from this late positivity might contribute to the different patterns observed during the N400 time-window. The N400 and the late positivity associated with semantic anomaly have a similar central-posterior distribution in EEG, and because they often co-occur, the absolute timing of the late positivity is difficult to determine. In principle, one could imagine a scenario in which the Chow stimuli do in fact elicit the same underlying N400 differences as the Ehrenhofer stimuli, but where those differences are obscured by an early-onset positivity in the opposite direction. We cannot evaluate this possibility further based on the current data, but we think it warrants future investigation, perhaps in a method like MEG, which is better able to separate the contributions of spatially separated neural generators.

6. Conclusion

This investigation tested the hypotheses of two different prediction-based accounts of the lack of N400 contrast on target verbs in role-reversed NNV and NVN sentence contexts. As predicted by the format mismatch account of verb prediction, we found an N400 contrast between target nouns in canonical and reversed NVN contexts. Unexpectedly, we also found an N400 contrast on target verbs in canonical and reversed NNV contexts. Explorations of the present study’s stimuli, and comparison with the stimuli used in Chow et al. (2015), yielded no measurable
contrasts between the stimuli in terms of entropy or maximum cloze distribution. However, we did discover a contrast in offline ratings of the plausibility of canonical and role-reversed sentences, and in a subject-verb cosine relationship measure of contextual similarity. We have outlined one possible account that emphasises the probability of the noun-as-agent co-occurring with different verbs, and we hope that these unexpected results will motivate a broader range of future work that can provide further insight into the time course of argument structure computation and linguistic prediction.

Acknowledgements

The authors would like to thank Allyson Ettinger and Phoebe Gaston for their contributions to the sections on entropy and SVCR. Furthermore, we are indebted to a dedicated group of research assistants who helped prepare this experiment and collect data: Jon Burnsky, Sade Evans, Maggie Kandel, Hanna Muller, and Neomi Rao. This work was supported in part by a Fulbright Fellowship to LE, by NSF award BCS-0848554 to CP and by NSF NRT Training Grant DGE-1449815 to the University of Maryland.

Note on Supplementary Materials

We have included all test items from Experiments 1 and 2 in the Supplemental Materials. We conducted post hoc analyses only for NNV items in Experiment 1, therefore extra information on frequency and SVCR are not available for NVN items.
References


