

Probabilistic online processing of sentence anomalies

Laurel Brehm, Carrie N. Jackson & Karen L. Miller

To cite this article: Laurel Brehm, Carrie N. Jackson & Karen L. Miller (2021) Probabilistic online processing of sentence anomalies, *Language, Cognition and Neuroscience*, 36:8, 959-983, DOI: [10.1080/23273798.2021.1900579](https://doi.org/10.1080/23273798.2021.1900579)

To link to this article: <https://doi.org/10.1080/23273798.2021.1900579>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 15 Mar 2021.



[Submit your article to this journal](#)



Article views: 767



[View related articles](#)



[View Crossmark data](#)



Citing articles: 2 [View citing articles](#)

Probabilistic online processing of sentence anomalies

Laurel Brehm^{a,b,c}, Carrie N. Jackson^{a,d} and Karen L. Miller^{a,e}

^aCenter for Language Science, Pennsylvania State University, University Park, PA, USA; ^bMax Planck Institute for Psycholinguistics, Nijmegen, Netherlands; ^cDepartment of Psychology, Pennsylvania State University, University Park, PA, USA; ^dDepartment of Germanic and Slavic Languages and Literatures, Pennsylvania State University, University Park, PA, USA; ^eDepartment of Spanish, Italian, and Portuguese, Pennsylvania State University, University Park, PA, USA

ABSTRACT

Listeners can successfully interpret the intended meaning of an utterance even when it contains errors or other unexpected anomalies. The present work combines an online measure of attention to sentence referents (visual world eye-tracking) with offline judgments of sentence meaning to disclose how the interpretation of anomalous sentences unfolds over time in order to explore mechanisms of non-literal processing. We use a metalinguistic judgment in Experiment 1 and an elicited imitation task in Experiment 2. In both experiments, we focus on one morphosyntactic anomaly (Subject-verb agreement; *The key to the cabinets literally *were ...*) and one semantic anomaly (Without; *Lulu went to the gym without her hat ?off*) and show that non-literal referents to each are considered upon hearing the anomalous region of the sentence. This shows that listeners understand anomalies by overwriting or adding to an initial interpretation and that this occurs incrementally and adaptively as the sentence unfolds.

ARTICLE HISTORY

Received 14 July 2020
Accepted 16 February 2021

KEYWORDS

Sentence processing; visual-world paradigm; subject-verb agreement; implicit negation; noisy channel; good-enough processing

1. Introduction

Language is full of variability with the potential to hinder communication. Speakers produce utterances with varied success, making occasional disfluencies or speech errors, and what is considered acceptable to one speaker might be ungrammatical or infelicitous to another. Yet, despite these potential obstacles, most messages are successfully conveyed from speaker to listener. The present work examines the online processing and ultimate comprehension of anomalous ungrammatical sentences, such as *The key to the cabinets literally *were on the table*, and semantically implausible sentences, such as *Lulu visited the gym without her hat ?off late yesterday night*. By combining an offline measure of sentence interpretation with an online measure of attention to referents, we demonstrate the real-time processing mechanisms that allow the comprehension of these morphosyntactic and semantic anomalies. This provides insight into how attention, meaning and morphosyntax contribute to language comprehension and sheds light on what allows communication to be so robust despite potential pitfalls.

1.1. Non-literal processing

Sometimes understanding anomalous utterances is easy, leading to many circumstances where a literal utterance is efficiently interpreted non-literally. Odd phrases like *The mother gave the candle the daughter* (Gibson et al., 2013) or *Lulu went to the gym without her hat off* (Brehm et al., 2018; Frazier & Clifton, 2015) are often reconstrued to add an element (*The mother gave the candle to the daughter*) or to eliminate an element (*Lulu went to the gym without her hat* _). Similarly, erroneous sentences like *The key to the cabinets *were on the table* are also often interpreted as if they include a grammatical, verb-matching subject noun (*The keys to the cabinets were on the table*; Brehm et al., 2018; Patson & Husband, 2016).

The *good enough processing* (Christianson et al., 2001, 2010; Ferreira, 2003) and *noisy channel* frameworks (Gibson et al., 2013; Levy et al., 2009) suggest mechanisms for how non-literal processing could occur. Broadly, both suggest that our expectations about likely mappings between semantics, pragmatics, and morphosyntactic forms can be used to reconstrue an odd or erroneous utterance. If an utterance has an

unlikely form or unlikely meaning, comprehenders may infer that what was literally said was not what was intended. This is a sensible strategy if we assume that the goal of conversation is the communication of meanings – not literal forms – between speakers and listeners.

The good enough processing framework was originally developed to describe how readers can interpret garden-path sentences non-literally: in these sentences, an initial interpretation that turns out not to be licensed by the sentence's syntax remains available once disambiguating information has appeared (Christianson et al., 2001). Good enough processing suggests that this happens because readers often interpret sentences using fast heuristics, instead of slow algorithmic processing (see Karimi & Ferreira, 2016, for a recent review). Heuristics cause individuals to fail to re-parse a sentence when its form conflicts with a more likely meaning or to incompletely incorporate new information as it is received. The consequence is that an initial interpretation often tends to linger even once new evidence points against it.

The heuristics incorporated under the good enough processing framework are based upon experience at multiple levels of processing. This includes argument structure, such that likely agents are interpreted as agents (Ferreira, 2003) and syntactic structure, such that individuals fail to re-analyse infrequent, garden-path structures and instead stick with an initial parse of a verb's structure (e.g. Christianson et al., 2001). Pragmatics, focus, and word order also cause comprehenders to process shallowly, leading to a variety of cognitive illusions (see e.g. Sanford & Sturt, 2002, for review). Evidence for good enough processing often comes from offline judgments (e.g. Christianson et al., 2001; Ferreira, 2003; Frazier & Clifton, 2011, 2015; Patson & Husband, 2016), and reading studies (self-paced reading; e.g. Christianson & Luke, 2011; Swets et al., 2008; eye-tracking while reading; Slattery et al., 2013), though recent work has expanded into ERPs (Qian et al., 2018). The upshot of this is that we have robust evidence that individuals sometimes ignore how sentences are syntactically, pragmatically, or semantically odd, and this oddness does sometimes lead to difficulty in processing.

The more recent noisy channel theory and related rational approaches to sentence processing were originally developed to account for the phenomenon that sometimes readers fail to notice small errors in the input (see e.g. Gibson et al., 2013). The theory, formalised under a Bayesian equation, says that sentence processing relies probabilistically on an algorithm that combines prior expectations with the observed form, which causes processing difficulty when expectations are not met (connected with the concept of *surprisal*,

Levy, 2008) or causes small errors to be corrected seamlessly (Gibson et al., 2013).

The noisy channel theory suggests that expectations of likely utterances map to what is semantically or pragmatically plausible, such that individuals either fail to notice plausible errors at all (Gibson et al., 2013), or revise them during reading by maintaining uncertainty (Levy et al., 2009). The likelihood of correcting an error or revising an implausible utterance is also affected by the amount of noise in the system (literal noise or signals of speaker unreliability, see e.g. Gibson et al., 2017), and the size of the edit that is needed to convert the literal input to a non-literal interpretation (e.g. Gibson et al., 2013). This is evidenced by answers to comprehension questions (Gibson et al., 2013, 2017), and increases in regressions and overall reading time in studies using eye-tracking while reading (Levy et al., 2009). The upshot of this is that errors can be mentally converted into correct utterances, especially in situations where errors are expected.

In the current paper, we explore predictions of both theories with one set of data, exploring how non-literal processing of anomalous sentences happens dynamically over time. Both frameworks overlap substantially in their predictions: listeners will re-interpret semantically and syntactically anomalous sentences as something that is more likely. Under good enough processing, non-literal processing arises because of learned heuristics about argument structure and word ordering. This might cause an initial interpretation of a sentence to linger in the face of new input, even if it conflicts with the original meaning. Under the noisy channel framework, non-literal processing arises because of prior experience with meaning-to-form mappings. This might lead to a swift repair of a small error, immediately overwriting the original interpretation. By exploring two different anomaly types in a paradigm that simultaneously investigates offline and online processing, we can evaluate whether these mechanisms both contribute to sentence processing. Finally, both frameworks predict that individuals should update their interpretation of sentences based upon incoming information; good-enough processing also predicts that listeners may not always succeed at revising an earlier interpretation, and the noisy channel theory also predicts small errors will be corrected quickly.

1.2. Measuring sentence interpretations online using visual world eye-tracking

Much of the previous work on non-literal processing has focused on the separate consequences of offline judgments and online difficulty during reading (though,

see Qian et al., 2018; Slattery et al., 2013). In the current paper, we integrate online and offline measures, demonstrating using how sentence interpretations change continuously over time using the visual world paradigm. This provides concrete evidence about when different representations are activated during processing, and how this leads to a particular interpretation of a sentence. This evidence then can shed light on how quickly reinterpretation can happen and what cues it uses.

The premise of the visual world eye-tracking paradigm is that an individual's fixations to objects on a computer screen while listening to a sentence serve as a proxy for their attention to referents described in the sentence. Changes to patterns of fixations over time show how an utterance interpretation changes in response to new linguistic information, providing information about the time-course of language processing (e.g. Allopenna et al., 1998; Altmann & Kamide, 1999; Tanenhaus et al., 1995, among many, see Huettig et al., 2011 for review).

Previous work has used visual world eye-tracking to show how sentences with multiple possible meanings are interpreted. For example, one line of work examines responses to syntactic ambiguities such as "Tickle the frog with the feather". Tracking fixations to objects matching each reading of the word "with" (designating an instrument: a frog, a feather vs. a modifier: a frog holding a feather) shows which referents are considered at which points in time (see e.g. Ryskin et al., 2017; Snedeker & Trueswell, 2004; Tanenhaus et al., 1995). Similarly, visual world eye-tracking has been used to demonstrate that comprehension of pronouns (Arnold et al., 2000; Brown-Schmidt et al., 2005; Fukumura et al., 2010) and quantifiers (Grodner et al., 2010; Huang & Snedeker, 2009) entails integration of multiple sources of information over time. In both lines of research, the key principle is that the visual world paradigm demonstrates how various types of information contained within a sentence are quickly and efficiently integrated through time to update an expected sentence interpretation. The current work is inspired by this earlier literature on determining reference. We track changes in participants' fixations, time-locked to critical words, to items reflecting literal and non-literal interpretations of sentences. This extends the visual world paradigm to demonstrate how, and when, the final meaning of an anomalous sentence is obtained.

1.3. Current studies

In two studies, we adapt the visual world paradigm to explore non-literal processing of semantic and

morphosyntactic sentence anomalies. We do so by sub-setting the online data based upon an offline measure of ultimate interpretation; this allows us to show how patterns of visual attention correspond to different interpretations of sentences. We measured interpretation in two ways. Experiment 1 used an offline metalinguistic task where participants selected the image matching the subject of the sentence, such that selections of a non-literal subject index non-literal interpretations. Experiment 2 used an offline elicited imitation task where participants repeated the sentence after a filler task, such that sentences repeated with changes provide an index of non-literal interpretation. The goal was to provide converging evidence for the link between non-literal interpretations and online processing in two different ways. Asking participants to select the subject provides an explicit measure of how the sentence was interpreted and links directly to previous non-literal processing research (e.g. Brehm et al., 2018), but requires use of explicit grammatical knowledge and may draw attention to the critical items, affecting how visual attention is allocated. Elicited imitation taps implicit linguistic knowledge (see Bley-Vroman & Chaudron, 1994; Chrabaszcz & Jiang, 2014; Erlam, 2006), but is quite different from tasks previously used to study non-literal processing. Equivalent results from both tasks therefore would provide a robust picture of the link between online and offline processing.

Across experiments, we examined the processing of two kinds of anomalies; the goal here was to examine broad principles of multiple aspects of non-literal processing. Experiments 1a and 2a examined subject-verb agreement (SVA), while Experiments 1b and 2b examined implicit negation using *without* and the preposition *off* or *on* (Without). These reflect examples of non-literal processing based primarily on morphosyntax (Experiments 1a and 2a) and based primarily on semantics (Experiments 1b and 2b). These two anomalies also differ in their likelihood of occurrence and consistency of interpretation, allowing us to further examine how expectations are affected by prior experience, and how expectations change given experience within an experiment. Agreement errors are common and have a preferred interpretation (the head was correct and the verb incorrect), but blends of implicit negation are uncommon and more varied in interpretation (see e.g. Brehm et al., 2018; Frazier & Clifton, 2015). If these different types of anomalies disclose similar behavioural patterns, this demonstrates general principles of anomaly processing that cut across both non-literal processing frameworks.

Within the experiments, we focused on three research questions. First, we asked whether the anomalies led to

the non-literal interpretation of sentence subjects: this is predicted under both non-literal processing frameworks. Next, we asked if, regardless of interpretation, fixations were driven towards new referents after hearing an anomaly and away from the first referent that was specified in the sentence. This shows the speed with which individuals revise the interpretation of sentence meanings given new input, regardless of the ultimate interpretation. Lingering initial interpretations are predicted under good-enough processing, and fast corrections of small errors are predicted under the noisy channel framework. Finally, we also asked how processing each anomaly changes through the course of the experiment, as individuals develop expectations about the types of anomalies present in it. Changes to processing over time are consistent with the noisy channel framework because expectations result directly from what the comprehender has previously observed; this fundamentally must update over time. Adaptation would also be consistent with experience-driven changes to heuristics under good enough processing because heuristics also develop out of past experience. We walk through how these research questions translate to specific predictions for each anomaly type below.

1.3.1. Subject-verb agreement

In Experiments 1a and 2a, we examine a morphosyntactic anomaly involving subject-verb agreement (SVA). Production errors occur for SVA, such that singular subjects with plural “local” (non-subject) nouns often elicit plural verbs (termed “attraction”, e.g. Bock & Miller, 1991). SVA production variability also occurs based upon dialect (e.g. Tagliamonte & Baayen, 2012; Tortora & den Dikken, 2010) and as a result of the phrase’s meaning (e.g. Brehm & Bock, 2013, 2017). This suggests that multiple inferences about morphosyntax could be licensed about why an SVA anomaly has occurred.

This variability has consequences for SVA processing. Attraction makes anomalous plural verbs relatively felicitous (1a is less bad than 1c; e.g. Pearlmutter et al., 1999; Wagers et al., 2009)¹; this is typically explained using a cue-based retrieval framework: the local noun (*cabinets*) is retrieved instead of the head to control the verb since it has a verb-matching number feature (see e.g. Dillon et al., 2013; Wagers et al., 2009). Anomalous plural verbs (1a/1c) also cause non-literal comprehension of the head, as do plural non-subject nouns (1a/1b), meaning that in both cases, the literal head noun (*key*) is interpreted as plural (*keys*; see Brehm et al., 2018; Patson & Husband, 2016; Schlueter et al., 2019).

- (a) The key to the cabinets literally *were on the table
- (b) The key to the cabinets literally was on the table

- (c) The key to the cabinet literally *were on the table
- (d) The key to the cabinet literally was on the table

Earlier work leaves open whether attraction and non-literal SVA comprehension have the same origins, which make this anomaly an ideal candidate to investigate in a paradigm combining online and offline measures. One interpretation of non-literal SVA comprehension is that a plural number feature on the local noun or verb causes re-construal of the head number (from *key* to *keys*) as a repair (as suggested by Patson & Husband, 2016 and consistent with e.g. Levy, 2008). This means that non-literal SVA comprehension and attraction could arise from the same mechanism: recasting the head’s number from singular to plural to repair the utterance would cause increased plural interpretations and also lead to attraction. Alternatively, attraction could come from mis-retrieving the local noun, not as a repair, following the typical view of cue-based retrieval in agreement processing. Or, attraction could arise as a repair, but non-literal interpretations could arise from a faulty or underspecified memory of the utterance (the gist was “plural”; suggested in Patson & Husband, 2016, and consistent with Tanner et al., 2014). Finally, a mixture of multiple mechanisms could obtain (see Schlueter et al., 2019, for discussion).

In the current work, we examine how attention is allocated to sentence referents varying in number (see Figure 1(a)); we can therefore see when the literal (*key*) and non-literal (*keys*) versions of the head noun, and literal and non-literal versions of the local noun (*cabinet(s)*) receive attention during listening. This sheds light on the link between processing difficulty and non-literal comprehension, asking whether non-literal interpretations of SVA anomalies are due to the quick integration of information as it unfolds or to a late revision process.

1.3.2. Without blends

In Experiments 1b and 2a, we examine a semantically implausible utterance combining two implicitly-negative elements: “without” and the preposition “off”. Use of negation is variable in English such that negative concord (*didn’t have no* = had nothing) and compositional negation (*didn’t have no* = had something) can both be used, depending on the speaker’s dialect (see e.g. Blanchette, 2013; Blanchette & Lukyanenko, 2019). Having two similar forms available to convey one message results in production blends (e.g. Butterworth, 1982; Cutting & Bock, 1997). In the case that speakers have more than one dialect available (e.g. Standardised English and a non-Standardised dialect), the two plans might compete to elicit a blend. This suggests that

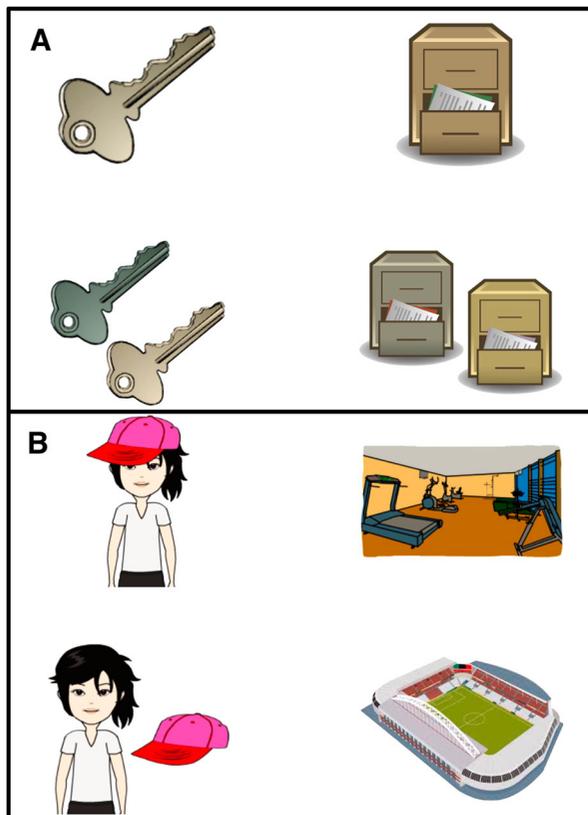


Figure 1. Sample image arrays. A: SVA item (*The key to the cabinet(s) literally was/were on the table*). Top left = literal subject key; bottom left = non-literal subject keys; right pictures = local noun and a non-literal version of it (*cabinet/cabinets*). B: Without item (*Lulu visited the gym without her hat on/off late yesterday night*). Top left = Yes-Accessory agent (the literal interpretation of “without + off”); bottom left = No-Accessory agent (the literal interpretation of “without + on” and non-literal interpretation of “without + off”). Right pictures = location of the verb and a semantic competitor for it.

multiple inferences might be drawn about why Without anomalies are produced (purposefully, reflecting concord, versus in error).

Correspondingly, comprehension of Without anomalies is variable. Potentially-anomalous utterances like 2a have two possible meanings, such that they elicit literal (“Lulu was wearing a hat”) and non-literal interpretations (“Lulu was not wearing a hat”) in contrast to utterances like 2b with no possible concord or blend, which are nearly always understood literally (“Lulu was not wearing a hat”, Brehm et al., 2018; Frazier & Clifton, 2015).

- (a) Lulu visited the gym without her hat ?off late yesterday night
- (b) Lulu visited the gym without her hat on late yesterday night

The suggestion is that the infrequent co-occurrence of “without” and “off” and the plausible site for a blend between two sentence plans (*without her hat on; with her hat off*) can cause listeners to infer the presence of a speech error despite the fact that the utterance is actually grammatical. Frazier and Clifton (2015) propose that this inference derives from the comprehender “reading through” the original utterance, ignoring redundant information when it can be attributed to noise in the system. This would suggest that when individuals take a non-literal interpretation of this anomaly, they should continue attending to the original referent at the anomalous *off*, rather than revising their choice. An alternate possibility might be that the comprehender notices the change in interpretation at the anomalous *off*, and spends time considering the alternate referent before ultimately selecting the original referent as the most plausible candidate for the utterance meaning.

In the current studies, we examine how Without anomalies are comprehended by examining how implicitly-negative elements (*without, off*) affect attention to agents who are wearing an optional accessory like a hat or shoes (Yes-Accessory, see Figure 1(b)) versus minimally-different agents with the accessory beside them (No-Accessory, see Figure 1(b)). The word *without* may lead a listener to the “No-Accessory” interpretation, drawing attention to the “No-Accessory” image. To obtain the literal interpretation of (2a), a listener then needs to revise the initial interpretation, which might be effortful or time-consuming. We ask how this revision occurs, examining whether non-literal interpretations of Without anomalies are due to the online integration of unfolding information or due to offline processes.

2. Experiment 1

In this experiment, we assessed individuals’ interpretations of sentences containing SVA (Experiment 1a) and Without anomalies (Experiment 1b) using a meta-linguistic forced-choice judgment combined with visual world eye-tracking to measure changes in interpretation over time, within trials and across the experiment.

2.1. Method

2.1.1. Participants

Data were collected from 38 members of the Pennsylvania State University community (mainly graduate and undergraduate students); participants were paid \$10 or given course credit in exchange for their participation. Of these, five individuals were excluded because they

were not native speakers of English, and one was excluded due to poor calibration. All 32 remaining participants were native speakers of English, ranged in age from 18 to 28 ($M=20$) and had normal or corrected-to-normal vision and hearing.

2.1.2. Materials and design

A list of critical items appears in Appendix A; items were adapted from previous work (Brehm et al., 2018; Solomon & Pearlmutter, 2004). There were four versions of each SVA item, varied in the number of the local noun (singular/plural) and verb (singular/plural), as in (1a) – (1d) above. There were two versions of each Without item, varied in preposition (*on/off*), as in (2a) and (2b) above. Filler items included datives (e.g. *The devil gave the treasure to the tooth fairy*), numerals (e.g. *Two giraffes were eating grass in the zoo*), and embedded relative clauses (e.g. *The button that fell off the boy's coat was blue*).

Items were recorded by a female speaker of American English from the Washington DC area. All critical item recordings were cross-spliced before the anomalous region of the sentence (the verb in SVA trials and the preposition in Without trials). The first half of each item was taken from a canonical sentence (e.g. a sentence like 1b, 1d, 2b) and the second half was taken from either another recording of a canonical sentence or from an anomalous sentence (1a, 1c, 2a). This was done to match stimuli across sets of items prior to the sentence anomaly.

Images were simple coloured line drawings. These were taken from stimulus databases (Duñabeitia et al., 2017; Rossion & Pourtois, 2004) or public domain sources (Wikimedia Commons, Pixabay, and Flickr), created using Pixton.com, or drawn by the first author. Images for SVA trials (e.g. as in Figure 1(a)) contained one- and two-token version of both the head and local nouns; in the two-token images, the colours of each token were adjusted to make them different. Images for Without trials (e.g. as in Figure 1(b)) contained an agent (half female, half male) with an accessory (“Yes-Accessory”), an agent without an accessory (“No-Accessory”), the theme or argument of the verb, and an item/location related to the theme/argument of the verb. For the “No-Accessory” images, the accessory appeared near, but separate from, the agent.

Filler trials used similar image types. Numeral fillers contained images with one to three tokens of an item, dative fillers contained two people and two semantically-similar objects/animals, and relative clause fillers contained a pair of images with “Yes-Accessory” or

“No-Accessory” version of the agent and a pair of objects with a colour contrast.

The session consisted of 108 trials, divided into four blocks of 27 trials. This included 48 critical SVA items, 24 critical Without items, 20 dative fillers, 8 fillers beginning with a numeral, and 8 fillers containing an embedded relative clause. Items were assigned to 16 lists with a Latin square design such that the four versions of SVA items and two versions of Without items were equally represented and so that the target image was equally likely to appear in each screen location. Within a list, items were presented in a fixed pseudo-randomised order such that an equal number of items of each condition were presented in each block, such that no more than two critical items were adjacent, and such that no two items in the same condition were adjacent. Two participants viewed each list.

2.1.3. Apparatus and procedure

Participants were instructed to listen to sentences and to click on the image from a four-image array that best matched the subject of the sentence. “Subject” was defined for participants at the onset of the experiment with the statement: “The subject of the sentence is the do-er of an action or the thing that is being described”.

Each trial began with a fixation target in the middle of the screen that the participant clicked on to start the trial. Then, four images appeared for 1500 ms of preview (as in Figure 1), a long duration in contrast to previous work which was designed to allow participants sufficient time to “spot the difference” between similar images and to fully encode the entire array. The preview phase was followed by a sentence played over speakers (range of lengths: 2213 ms to 3495 ms). Participants were then given an unlimited amount of time to select the image corresponding to the sentence subject. Once the participant clicked an image, an orange box appeared around the selection and remained there for 2000ms. Participants were offered the chance to take breaks between blocks and the session lasted between 30 and 40 min.

The experiment was run using Experiment Builder on a Dell PC with a 21 × 11.5 inch monitor and an EyeLink 1000 Plus Desktop in remote mode (sample rate 500 Hz) placed 20 inches in front of the computer monitor. Participants were seated about 20 inches in front of the camera. Images appeared in ports that were 6.75 inches wide by 5 inches high, spaced 2.75 inches apart horizontally and 1.5 inches apart vertically; these ports defined the experimental interest areas. All sounds were played over speakers at a volume comfortable to the participant.

2.1.4. Analysis

All analyses were conducted in R (version 3.3.3; R Core Team, 2014) using the package lme4 (version 1.1-13; Bates et al., 2014). Analysis of subject selection used logistic regression with odds of selecting a non-literal target object as the dependent measure. Analysis of eye-tracking data used linear regression with the dependent measure of proportion of fixations per person per trial to competitor objects in 500 ms windows offset from onsets of critical words by 200 ms, reflecting the typical time to launch an eye movement (e.g. Cooper, 1974). These windows were selected because they were of a similar length to the nouns and verbs in the sound stimuli and did not overlap in time; 500 ms also provided a reasonable amount of data per window. Analyses were performed separately for each window without correction for multiple comparisons, as is typically done in the field. However, because the data are not completely independent across time windows due to autocorrelation, we also note in the tables any effects that are not significant when the alpha level was corrected by dividing by the number of time windows within an experiment. All predictors were contrast coded and centered; more detail follows within each experiment. Random intercepts were included for Participants and Items and as many random slopes as justified by the data were included (e.g. Barr et al., 2013), with slopes being removed due to non-convergence (beginning with the highest-order interactions) or due to correlations above 0.9 between them. For eye-tracking analyses, *p*-values were obtained by model comparison.

2.2. Experiment 1a: SVA trials

The SVA anomaly reflects a syntactic dependency gone awry. We examine the meaning interpreted from anomalous, ungrammatical utterances such as *The key to the cabinet(s) literally *were on the table* in contrast to canonical utterances containing the verb *was*, providing insight into how morphosyntax is processed non-literally and how attraction arises in comprehension. The prediction was that if non-literal interpretation and attraction arise from the same mechanisms – a recasting of the head or head number – local plural nouns and plural verbs should increase the likelihood of non-literal interpretation and increase fixations to objects other than the literal head noun, but that if the two are fundamentally separate, patterns should differ between interpretation and eye-tracking measures.

2.2.1. Analysis

Analysis of subject selections examined the odds of selecting the plural number competitor to the head

(“Keys”) in the offline interpretation task, with the predictors of Local Noun Type (singular, plural, contrasts of .5, -.5), Verb Type (singular, plural, contrasts of .5, -.5), and Trial Number (scaled and centered). Analysis of eye-tracking data examined the proportion of fixations per participant per trial to the plural version of the head (e.g. “Keys”) in four 500 ms time windows. Time windows began 200 ms after onset of the head noun, the local noun, and the verb, and 700 ms after the onset of the verb, with the aim of capturing processing in response to each critical word, and after hearing the verb and integrating it with the rest of the sentence. Analysis of eye-tracking data was performed only on the trials where the canonical subject was selected, as this was the overwhelmingly dominant response. For the eye tracking analyses, predictors were Local Noun (singular, plural, centered contrasts of .48, -.51), Verb (singular, plural, centered contrasts of .47, -.53) and Trial Number (scaled and centered).²

2.2.2. Results

Subject selection. As shown in Figure 2, plural local nouns and plural verbs increased the odds of selecting the non-literal head “Keys” as the subject of the sentence. Plural local nouns led to selection of the non-literal head 10% of the time, vs. 5% for singular local nouns, and plural verbs led to selection of the non-literal head 14% of the time, vs. 2% for singular verbs. These patterns were confirmed by mixed-effect modelling (see Table 1).

On top of these patterns, we also observed a main effect of Trial Number and an interaction between Trial Number and Verb such that the non-literal head “Keys” was selected more often at the beginning than the end of the experiment (11% in Block 1 vs. 6% in Block 4)³, and the decrease was clearest in the plural verb conditions (dropping from 16% in Block 1 to 12% in Block 4). We did not test whether the interaction between local noun and verb was qualified by trial number: since the non-literal head was selected only on one trial in the singular local noun, singular verb condition, we omitted the three-way interaction between Local Noun, Verb and Trial Number from the model of these data.⁴

Proportion of fixations. Proportions of fixations split by condition appear in Figure 3. Throughout the trial, participants mostly fixated the image of the literal head noun (e.g. “Key”) and its non-literal number competitor (e.g. “Keys”), with few fixations to the images representing the local noun and its competitor (e.g. “Cabinet”/“Cabinets”).

There were minimal differences between sentence versions before the anomalous region of the sentence (see Table 2 for mixed-effect models). In the first time

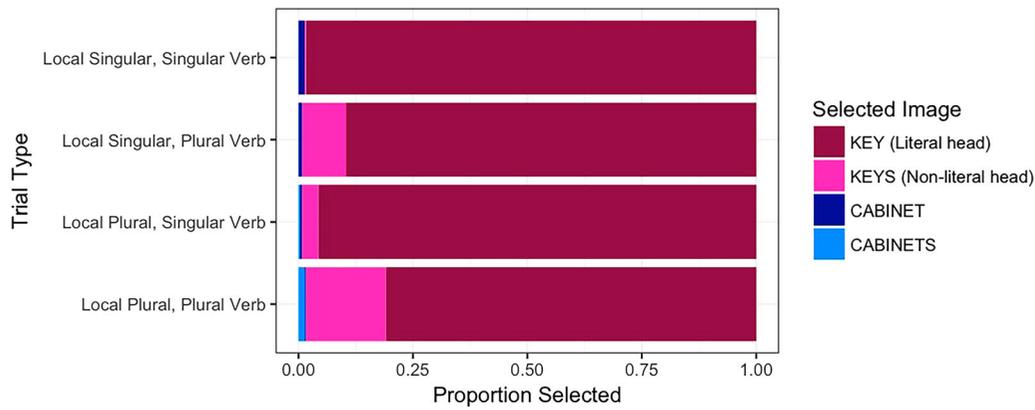


Figure 2. Subject selection for Experiment 1 SVA trials, split by condition.

Table 1. Results from logistic mixed-effect regression of subject selection (odds of selecting head noun number competitor versus all other items) for Experiment 1 SVA trials.

Fixed effects	$\hat{\beta}$	SE	$p(z)$
Intercept	-4.76	0.51	<0.001
Local Noun	-1.85	0.58	0.01
Verb	-4.14	0.77	<0.001
Trial Number	-0.95	0.26	<0.001
Local Noun \times Verb	-1.85	1.14	0.10
Local Noun \times Trial Number	0.17	0.27	0.53
Verb \times Trial Number	-1.46	0.51	<0.01
Random effects	Groups	SD	
	Item	Intercept	0.49
	Participant	Intercept	1.45

window, representing processing relative to the head noun (“key”), there were no significant differences in fixations to the non-literal head (“Keys”). In the second time window, representing processing relative to the local noun (“cabinet(s)”), there was a marginal effect of Trial Number ($p=0.052$) and a significant interaction between Local Noun and Trial Number such that trials at the end of the experiment elicited fewer fixations to the non-literal head (“Keys”) and the reduction was largest for items with singular local nouns (dropping from 32% in Block 1 to 20% in Block 4).

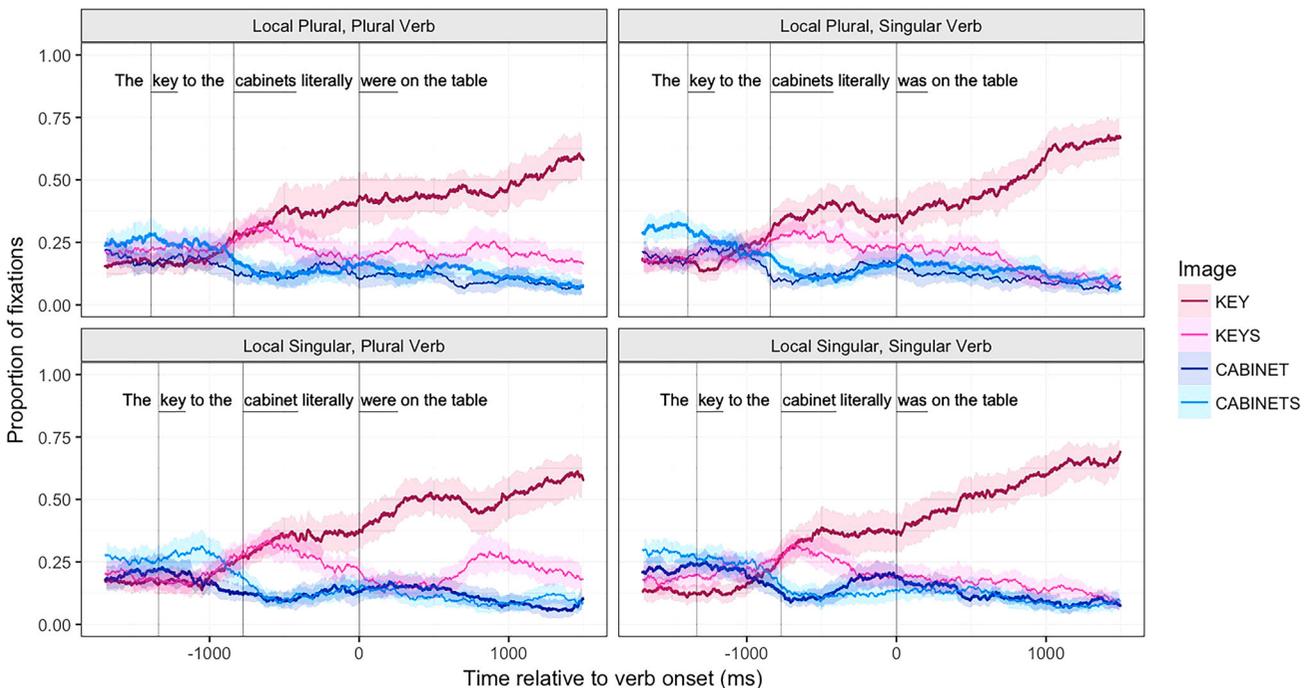


Figure 3. Average proportion of fixations to images in Experiment 1 SVA trials receiving a literal sentence interpretation (KEY selected as the subject), zeroed to reflect verb onset. Vertical lines represent onsets of head nouns, local nouns, and verbs; panels represent sentence conditions. Confidence bands are 95% CIs from a non-parametric bootstrap (1000 iterations) sampled over participants with replacement at 10 msec intervals.

Table 2. Results from linear mixed-effect regression of proportion of fixations to head noun number competitor binned at 500 msec time windows for Experiment 1 SVA trials eliciting a literal-head number interpretation.

Fixed effects	Window 1: Head noun region		Window 2: Local noun region		Window 3: Verb region		Window 4: Post-verb region		
	β	SE	$p(\chi^2)$	β	SE	$p(\chi^2)$	β	SE	$p(\chi^2)$
Intercept	0.25	0.01	<0.001	0.26	0.02	<0.001	0.20	0.02	<0.001
Local Noun	0.01	0.02	0.44	0.00	0.02	0.87	-0.05	0.02	0.11
Verb	-0.01	0.02	0.52	-0.02	0.02	0.32	-0.00	0.02	0.94
Trial Number	-0.02	0.01	0.09	-0.02	0.01	0.05	-0.04	0.01	<0.001
Local Noun × Verb	0.01	0.04	0.78	-0.03	0.04	0.41	0.00	0.03	0.95
Local Noun × Trial	0.02	0.02	0.21	-0.04	0.02	0.02	0.02	0.02	0.15
Verb × Trial	0.00	0.02	0.36	-0.03	0.02	0.15	-0.00	0.02	0.78
Local Noun × Verb × Trial	-0.02	0.04	0.61	-0.00	0.04	0.89	-0.04	0.03	0.25
Random effects			SD			SD			SD
	Item	Intercept	0.04	Item	Intercept	0.04	Item	Intercept	0.04
	Participant	Intercept	0.04	Participant	Intercept	0.06	Participant	Intercept	0.08
		Verb	0.01		Loc Noun	0.04		Loc Noun	0.07
		Trial	0.02		Verb	0.05		Residual	0.30
	Residual		0.33	Residual	Loc Noun × Verb	0.08			0.30
		Groups				0.33			

Effects are italicised when they would be significant under a 0.05 alpha level, but not an alpha level corrected for multiple comparisons.

Larger differences between sentence versions appeared after the anomaly “were” (see Table 2). In the third time window, representing processing relative to the verb (“was/were”) there was a significant main effect of Local Noun such that a plural local noun increased fixations to the non-literal head upon hearing the verb (21% for plural local nouns, 17% for singular local nouns) and a main effect of Trial Number such that fixations to the non-literal head upon hearing the verb decreased during the experiment (from 24% in Block 1 to 14% in Block 4). In the fourth time window, representing processing after hearing and integrating the verb with prior context, there was a significant main effect of Verb such that having heard a plural verb increased fixations to the non-literal head (23% for plural verbs, vs. 14% for singular verbs), and a main effect of Trial Number such that fixations to the non-literal head decreased during the experiment (from 24% in Block 1 to 15% in Block 4).

2.2.3. Discussion

For SVA anomalies (*The key to the cabinet(s) *were/was ...*) containing local plural nouns and plural verbs, the non-literal sentence subject (“Keys”) was selected more often in a metalinguistic judgment task compared to control items. This replicates previous results (e.g. Brehm et al., 2018; Patson & Husband, 2016). Local plural nouns and plural verbs also influenced attention on trials where the literal sentence subject (“Key”) was selected. Within these trials, plural features on local nouns (*cabinets*) and verbs (*were*) led to an increase in fixations to a non-literal version of the head (“Keys”). This shows how conflicting number features lead listeners to consider an alternate referent and provides evidence that attraction and non-literal processing are both supported – at least in part – by repair processes.

These eye-tracking results further suggest that as in the standard cue-based retrieval view of agreement attraction (e.g. Wagers et al., 2009), attraction happens when a plural verb cues retrieval of a plural controller. Existing literature interprets this pattern to mean the local noun controls agreement. However, we observed that a plural version of the head was chosen instead to represent the sentence subject. This difference could be due to the task: as an anonymous reviewer notes, participants tended to fixate both versions of the head noun prior to hearing the verb, which could cause the non-literal head to be encoded, and later, retrieved, following standard cue-based retrieval mechanisms. Alternatively, the comprehender might retrieve the head noun lemma and a plural feature, which are recombined to make a non-literal plural version of the head (see Patson & Husband, 2016 for similar logic).

Changes in fixations to the non-literal head occurred in separate time windows for nouns and verbs, suggesting their separate influence on interpretation. These patterns also tended to attenuate with exposure, such that at the end of the experiment, non-literal interpretations occurred less often and fewer fixations were made to non-literal referents in time-windows two, three, and four. Combined, these data suggest that, consistent with the noisy channel framework, offline non-literal interpretations can arise due to a fast online repair process, which is impacted by past and recent experience.

2.3. Experiment 1b: without trials

This experiment focuses on processing of Without anomalies, a semantically-odd sentence that could be produced purposefully or as an error. We examine how items such as *Lulu visited the gym without her hat ?off late yesterday night* are interpreted in contrast to plausible control items, asking whether Without anomalies lead to increases in non-literal interpretations and disclosing how interpretations change during processing. The prediction was that if non-literal interpretations arise due to a faulty revision process, fixations to the original referent should be maintained even after hearing the anomaly.

2.3.1. Analysis

The analysis of subject selections examined the odds of selecting the non-literal target image, which was the “Yes-Accessory” image for *without-on* items and the “No-Accessory” image for *without-off* items. Predictors were Preposition type (on, off, contrasts of .5, -.5) and Trial Number (scaled and centered).

The eye-tracking analysis examined the proportion of fixations per participant per trial to a competitor object

(“Yes-Accessory” image, the less likely agent to be selected) in 500 ms time windows beginning 200 ms after the onset of *without* and the preposition *on/off*, and 700 ms after the onset of the preposition, reflecting processing in response to *without*, the preposition, and after integrating the preposition with the rest of the sentence. Preposition was combined with subject selection to yield a three-level variable of Interpretation: “Without-On” items interpreted literally, “Without-Off” items interpreted non-literally, and “Without-Off” items interpreted literally.⁵ These were analysed with centered Helmert contrasts. The first contrast compared trials with different subject selections, contrasting the No-Accessory selections (“Without On”+literal = .08 plus “Without Off”+non-literal = .30) vs. the Yes-Accessory selections (“Without Off”+literal = -0.62); the second contrast compared trials when the No-Accessory picture was selected (“Without On”+literal = -0.22 vs. “Without Off” + non-literal = 0.78).⁶ Scaled and centered Trial Number was also entered as a predictor.

2.3.2. Results

Subject selection. Items containing the non-anomalous preposition *on* led to little variability in responding (see Figure 4), mainly eliciting selection of the literally-interpreted No-Accessory agent as the sentence subject (95% of trials). The rest of the trials were split between the Yes-Accessory agent (3%) representing a non-literal interpretation, and the theme/location (2%), representing confusion about the subject. Items with the potentially-anomalous preposition *off* were interpreted more variably, with the Yes-Accessory agent (literal interpretation) selected on 72% of trials, the No-Accessory agent (non-literal interpretation) selected on 26%, of trials and the theme/location or its foil selected on 1% of trials. This was reflected in a main effect of

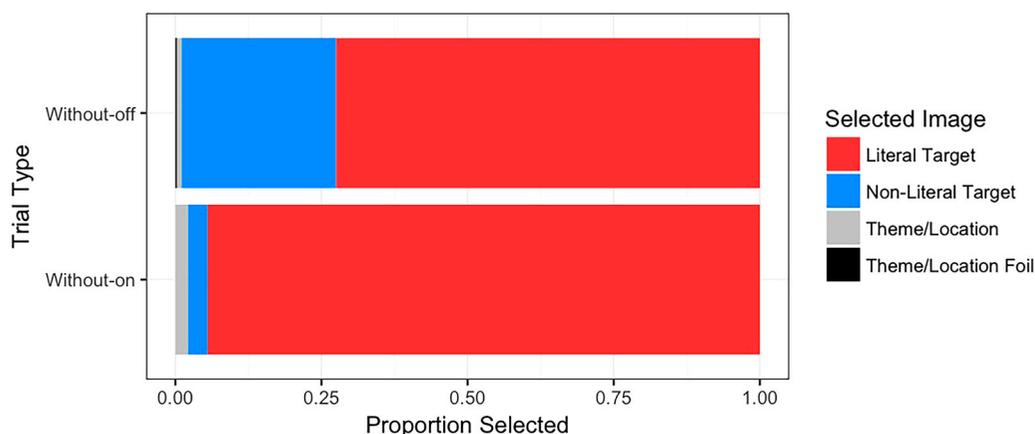


Figure 4. Subject selection for Experiment 1 Without trials, split by condition.

Table 3. Results from logistic mixed-effect regression of subject selection (odds of selecting non-literal subject versus all other items) for Without trials.

Fixed effects	$\hat{\beta}$	SE	$p(z)$
Intercept	-6.92	1.65	<0.001
Preposition	-9.35	3.29	<0.01
Trial Number	-0.96	0.39	0.01
Preposition \times Trial Number	-1.16	0.78	0.14
Random effects	Groups	SD	
	Item	Intercept	5.62
	Participant	Preposition	11.19
		Intercept	0.00

Preposition on non-literal responses in a mixed-effect model; see Table 3.

As Experiment 1a, non-literal response rates dropped through the course of the experiment. In a mixed-effect model, (see Table 3), this was reflected in a main effect of Trial Number, such that 26% of the responses in Block 1 were non-literal but 9% of the responses in Block 4 were non-literal. No other significant main effects or interactions were observed.

Proportion of fixations. As shown in Figure 5, participants tended to fixate both potential agents roughly equally (“Yes-Accessory” and “No-Accessory”) until hearing the word *without*; while there was a slight numeric tendency to prefer the No-Accessory agent, this difference was not statistically reliable. At this point, looks to the No-Accessory agent rose. The

preposition *off* often led to a revision of the original interpretation, such that when the sentence was interpreted literally, looks to the Yes-Accessory agent increased at late time windows of the sentence. Overall, few looks were directed to the location/theme (depicted in grey) or its competitor (not depicted for simplicity).

Mixed-effect model results are shown in Table 4. In the first analysis window, representing processing in response to the word *without*, there was a main effect of Trial Number such that more looks were directed toward the “Yes-Accessory” agent late in the experiment (weighted mean Block 1: 27% vs. Block 4: 31%). Trial Number interacted with the second Interpretation contrast such that for non-literal *Without-Off*-“No-Accessory” interpretations, fixations to the “Yes-Accessory” agent dropped during the experiment (from 32% in Block 1 to 25% in Block 4), while for literal *Without-On*-“No-Accessory” interpretations, fixations to the “Yes-Accessory” agent rose during the experiment (from 22% in Block 1 to 34% in Block 4). This suggests that if participants directed their attention away from the “Yes-Accessory” agent before hearing the sentence anomaly (e.g. for *Without-Off* sentences), they were less likely to select it as the subject.

In the next window, representing processing in response to the preposition, an interaction between

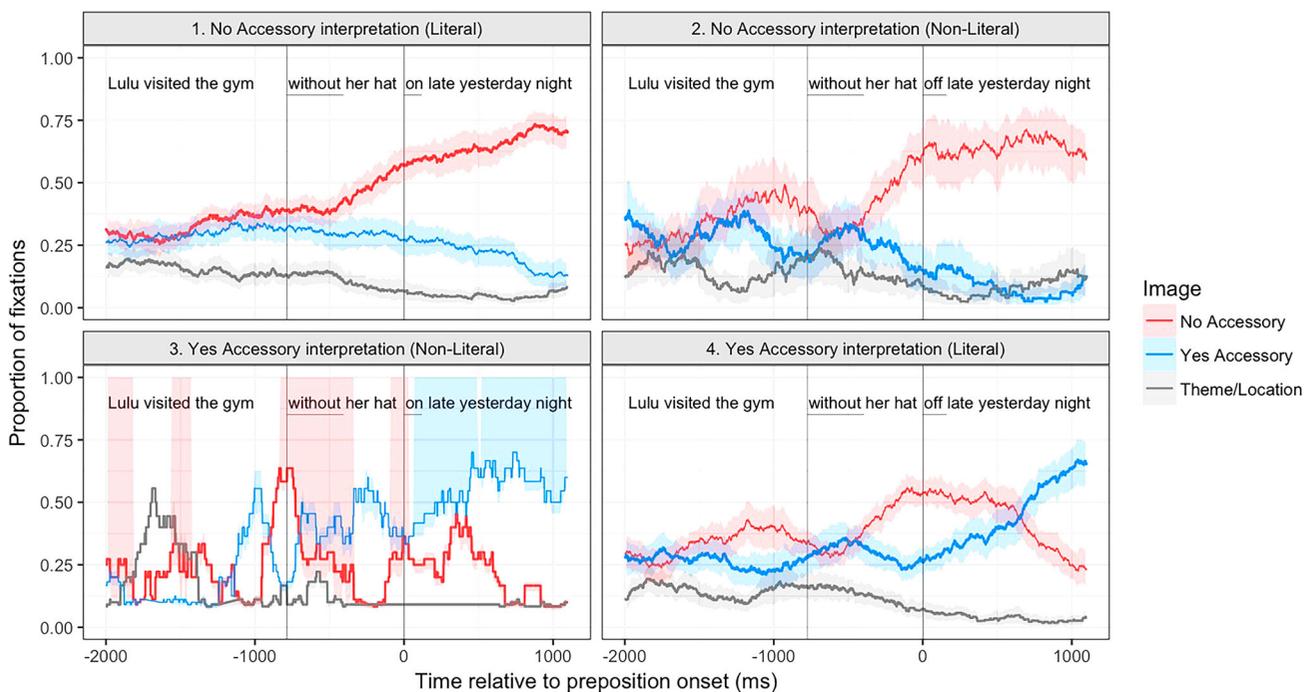


Figure 5. Average proportion of fixations to images in of Experiment 1 Without trials, zeroed to reflect preposition onset. Vertical lines represent onsets of *without* and prepositions; panels represent sentence conditions crossed with sentence interpretation. The non-literal interpretation of “Without-On” (bottom left panel) was not analysed because of an insufficient number of trials but we present it here for completeness. Confidence bands are 95% CIs from a non-parametric bootstrap (1000 iterations) sampled over participants with replacement at 10 msec intervals.

Table 4. Results from linear mixed-effect regression of proportion of fixations to yes-accessory agent binned at 500 msec time windows for Experiment 1 Without trials split by interpretation.

Fixed effects	Window 1: Without region			Window 2 Preposition region			Window 3: Post-Preposition region		
	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$
Intercept	0.28	0.02	<0.001	0.25	0.03	<0.001	0.19	0.02	<0.001
Interpretation 1: (Yes- vs. No-Accessory)	0.02	0.06	0.78	-0.10	0.07	0.13	-0.61	0.06	<0.001
Interpretation 2: (No-Accessory Literal vs. Non-Literal)	-0.06	0.07	0.43	-0.13	0.07	0.08	0.08	0.06	0.19
Trial Number	0.04	0.02	0.02	0.00	0.00	0.77	0.00	0.00	0.34
Interpretation 1 (Yes- vs. No-Acc) \times Trial	0.08	0.06	0.18	0.00	0.00	0.04	-0.01	0.00	<0.001
Interpretation 2 (Yes-Acc Lit vs. Non-Lit) \times Trial	-0.15	0.07	0.03	0.00	0.00	0.40	0.00	0.00	0.28
Random effects	Groups			Groups			Groups		
	Item	Intercept	0.03	Item	Intercept	0.06	Item	Intercept	0.03
		Trial	0.02		Interp. 1	0.11	Participant	Intercept	0.06
	Participant	Intercept	0.08	Participant	Intercept	0.09		Interp. 1	0.20
		Interp. 1	0.14		Interp. 1	0.17	Residual		0.31
		Interp. 2	0.11		Interp. 2	0.10			
		Trial	0.04	Residual		0.33			
	Residual		0.36						

Effects are italicised when they would be significant under a 0.05 alpha level, but not an alpha level corrected for multiple comparisons.

Trial Number and the Interpretation contrast 1 was observed such that more looks were directed toward the “Yes-Accessory” agent at the end of the experiment especially if it was selected as the subject (26% in Block 1, vs. 45% in Block 4). This suggests that participants learned to tune their attention during the experiment in response to the preposition.

All differences were largest in the third time window, representing processing after hearing and integrating the anomaly with prior context. In this window, there was a main effect of the Interpretation contrast 1, such that items for which the “Yes-Accessory” agent was selected elicited more looks to the “Yes-Accessory” agent (61% of fixations, vs. “No-Accessory” weighted mean of 11%). This interacted with Trial Number such that fixations to the “Yes-Accessory” agent increased during the experiment when the “Yes-Accessory” agent was selected as the sentence subject (36% of fixations in Block 1, vs. 70% in Block 4).

2.3.3. Discussion

For Without anomalies (*without her hat ?off*), we observed more selections of a non-literal sentence subject (“No-Accessory”) in a metalinguistic judgment task compared to non-anomalous control items. This replicates previous work (e.g. Brehm et al., 2018; Frazier & Clifton, 2015). We also showed that fixations to possible referents varied depending on whether the sentence contained a potential anomaly and how the sentence was interpreted. When *without + off* was interpreted literally, there were more fixations to the “Yes-Accessory” agent after the word *off* was processed, demonstrating how revisions in interpretation can be prompted by even potentially-anomalous elements. In contrast, when *without + off*

was interpreted non-literally, there were few fixations to the competitor “Yes-Accessory” agent in the same region, such that all trials receiving the same “No-Accessory” interpretation had statistically equivalent patterns of fixations to the competitor object. This underscores the connection between visual attention and inferred meaning, a premise of the visual-world paradigm, even in anomalous sentences. It also underscores how, consistent with the good-enough processing framework, non-literal interpretations arise from the dynamic revision or maintenance of how a sentence is interpreted in response to new information.

As in Experiment 1a, all patterns tended to change with exposure: at the end of the experiment, fewer non-literal interpretations occurred and fewer fixations were made to non-literal referents. This demonstrates how exposure influences expectations about what a sentence is likely to mean and confirms the link between anomaly processing and experience with the mapping between likely meanings and likely forms.

3. Experiment 2

In this experiment, we assessed individuals’ interpretations of sentences containing SVA (Experiment 2a) and Without anomalies (Experiment 2b) using an elicited imitation task. We again paired this with visual-world eye-tracking in order to measure interpretation changes during processing. Elicited imitation provides an implicit measure of sentence meaning without requiring a metalinguistic judgment; the question was whether the online and offline patterns in Experiment 1 replicated with an implicit measure of non-literal interpretation.

3.1. Method

3.1.1. Participants

Data were collected from 35 members of the Pennsylvania State University community (mainly undergraduate and graduate students); participants were paid \$10 or given course credit in exchange for their participation. Two individuals were excluded due to missing audio data and one was excluded due to poor eye-tracking calibration. The 32 remaining participants were all native speakers of English, ranged in age from 18 to 36 ($M = 23$) and had normal or corrected-to-normal vision and hearing.

3.1.2. Materials and design

Critical items were identical to Experiment 1. To implement a distractor task for the elicited imitation paradigm, two-thirds of the filler items (24 out of 36, 22% of all items) were altered by swapping nouns between pairs of fillers from Experiment 1 to make them refer to highly unlikely events (e.g. *Three clocks were eating grass in the zoo*).

3.1.3. Apparatus and procedure

The first phase of the trial matched Experiment 1: trials began with a fixation cross, a preview of four images, and a sentence played over the computer speakers. Images remained on screen 4000 ms after the onset of the sentence. The next phase was inspired by Chrabaszcz and Jiang (2014). Participants performed a pragmatic distractor task in which they were given unlimited time to judge “How likely is this event?” by selecting from a three-point scale (Likely, Not Sure, Unlikely) with the mouse. Next, images were re-presented in the same or scrambled locations in order to elicit the production of the sentence presented earlier in the trial. For critical trials, all four images appeared in a new position, counterbalanced such that across participants, each image appeared equally often in each position; the goal of this was to increase the difficulty of the recall task. For filler trials, the new array was identical or had two to four images in a new position. The number of trials, length of experiment, and apparatus matched Experiment 1.

3.1.4. Analysis

Repetitions were coded as to whether they reflected literal or non-literal interpretations, as outlined separately for each sub-experiment below. Logistic regressions on odds of non-literal interpretation and eye-tracking analyses on fixations to the non-literal head were performed as described in Experiment 1.

3.2. Experiment 2a: SVA trials

Experiment 1a examined items with a potential syntactic anomaly (SVA items: *The key to the cabinet(s) *were/was*). This experiment disclosed that a non-literal sentence subject (“Keys”) was selected more often when the sentence contained plural local nouns or plural verbs and that these plural cues also increased attention to a non-literal version of the head noun during processing when the sentence was interpreted literally. The goal of Experiment 2a was to assess whether these patterns replicated with a more implicit offline measure. If so, plural features should again lead to more non-literal inflection-changing repetitions and the pattern of fixations during processing should broadly mirror Experiment 1a.

3.2.1. Analysis

The analysis of sentence repetitions examined the log odds of sentences being produced non-literally, as defined as changes to the number inflections on the two preamble nouns or the verb. We compared these non-literal responses against identical repetitions and other types of repetition error. Identical repetitions were all productions in which all content words were reproduced as in the original prompt, including trials with disfluencies or restarts. The remaining trials were classified as other repetitions. These included all trials where the inflections matched the original prompt but the adverb was misplaced or the preposition was changed. We contrasted the non-literal repetitions with all other repetition types because only the non-literal repetition of inflections could change which picture matched the sentence subject.

The analysis of eye-tracking data examined the proportion of fixations directed at the non-literal head image across the same time windows as described in Experiment 1a. In this analysis, we included all identical and other repetitions: trials in which all noun and verb inflections were reproduced as in the prompt.

As in Experiment 1a, the analysis of repetitions contained the predictors of Local Noun (singular, plural, contrasts of .5, -.5), Verb (singular, plural, contrasts of .5, -.5), and Trial Number (scaled and centered). Analysis of eye-tracking data focused on the comprehension phase of the experiment, with the same windows and analysis procedure as in Experiment 1a. Predictors were Local Noun (singular, plural, centered contrasts of .47, -.53), Verb (singular, plural, centered contrasts of .47, -.53) and Trial Number (scaled and centered).

3.2.2. Results

Pragmatic distractor task. Prompts were judged as “likely” the majority of the time (Local plural, plural

verb: 84%, Local plural, singular verb: 87%, Local singular, plural verb: 78%, Local singular, singular verb: 85%), with the remaining trials more often rated as “unlikely” (Local plural, plural verb: 12%, Local plural, singular verb: 10%, Local singular, plural verb: 15%, Local singular, singular verb: 11%) than “not sure” (Local plural, plural verb: 5%, Local plural, singular verb: 4%, Local singular, plural verb: 7%, Local singular, singular verb: 4%). In a model predicting the odds of “likely” ratings by Local Noun and Verb, with random intercepts by Participants and Items and a random slope for Item by Verb, there were reliably more “likely” ratings when the prompt contained a plural local noun ($\hat{\beta} = -0.36$, $SE = 0.16$, $z = -2.20$, $p > 0.05$) or a singular verb ($\hat{\beta} = 0.78$, $SE = 0.24$, $z = -8.01$, $p > 0.01$).

Sentence repetition. Prompts were repeated identically on slightly more than half of the trials (55% of all trials), with high rates of repetitions that changed part of the sentence but preserved all inflections (other responses; 29% of all trials) and lower rates of inflection-changing repetitions (16% of all trials). Following the tendency for elicited imitations to mirror the statistics of the language (see e.g. Erlam, 2006 for review), most of the inflection-changed repetitions were grammatical (60%). See Table 5 for a fine-grained coding of repetitions.

As shown in Table 5, inflections were changed to form two main types of repairs. For trials containing anomalous plural verbs, the verb was changed from plural to singular on 15% of trials, while the head was changed from singular to plural on only 4% of trials. These two types of repairs are likely to reflect different inferences about the source of the error: verb inflections are likely to be mis-produced, but noun inflections are more likely to be mis-heard (see Brehm et al., 2018 for further discussion). However, given the sparse data, we

collapse both together to analyse the overall odds of inflection changes – our measure of non-literal interpretation in this experiment. This also serves as a more direct comparison to Experiment 1a.

Mirroring Experiment 1a, odds of inflection change were affected by the Local Noun and Verb in the original prompt and by Trial Number. As shown in Figure 6, plural local nouns and plural verbs led to increased numbers of repetitions with inflection changes (light and bright pink bars), with most changes occurring when the local noun and verb were both plural (26% of trials). These patterns manifested as significant main effects of Local Noun and Verb and an interaction between them in mixed-effect models (see Table 6). An interaction between Trial Number and Verb was also observed such that plural verb-containing trials elicited fewer inflection changes late in the experiment (from 31% in Block 1 to 18% in Block 4).

We also qualitatively examined whether the same patterns obtained in non-literal trials that were produced as grammatical versus ungrammatical utterances (see bold versus non-bold items in the “non-literal responses” section in Table 5). The pattern in the grammatical non-literal responses (bright pink in Figure 6) was comparable to the omnibus results, such that these increased when the prompt contained local plural nouns and plural verbs. In contrast, ungrammatical non-literal responses seemed only to be affected by local noun number; this suggests that plural local nouns tended to elicit attraction errors in this elicited imitation paradigm (light pink in Figure 6).

Proportion of fixations. Proportions of fixations on trials in which all inflections were repeated correctly (identical and other repetitions) split by condition appear in Figure 7. Compared to Experiment 1a, participants fixated less often on the head noun and its

Table 5. Response types by category, Experiment 2 SVA trials.

Response type	Loc Plural, Plural Verb	Loc Plural, Singular Verb	Loc Singular, Plural Verb	Loc Singular, Singular Verb
Identical repetition	174	210	216	247
No response	1	1	2	1
<i>Non-literal responses (Inflections changed)</i>				
N1 changed	19	4	2	0
N2 changed	27	34	14	8
V changed	42	21	38	9
N1 and N2 changed	9	2	2	0
N1 and V changed	1	0	0	0
N2 and V changed	1	1	3	1
N1, N2, and V changed	1	1	0	0
<i>Other changes (Inflections preserved)</i>				
Adverb & verb re-ordered	83	86	81	100
Adverb changed	2	8	3	1
Preposition changed	4	3	6	6
All other changes	20	13	17	11

Grammatical utterances in bold.

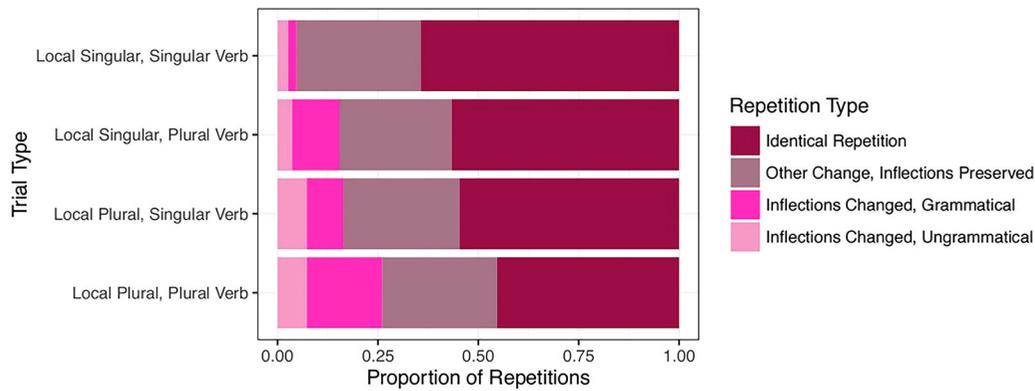


Figure 6. Sentence repetition for Experiment 2 SVA trials, split by condition and whether the inflection change led to a grammatical or ungrammatical utterance.

Table 6. Results from logistic mixed-effect regression of responses with number inflection changes, versus identical and other repetitions, for Experiment 2 SVA trials.

Fixed effects	β	SE	$p(z)$
Intercept	-2.36	0.24	<0.001
Local Noun	-1.12	0.19	<0.001
Verb	-1.27	0.19	<0.001
Trial Number	-0.07	0.09	0.47
Local Noun \times Verb	-0.74	0.37	0.05
Local Noun \times Trial Number	-0.12	0.18	0.52
Verb \times Trial Number	0.77	0.18	<0.001
Local Noun \times Verb \times Trial Number	0.32	0.36	0.37
Random effects			SD
Groups			
Item	Intercept		0.86
Participant	Intercept		0.94

competitor and more often on the prompt's local noun ("Cabinet"/ "Cabinets"), such that all four images received more comparable numbers of looks.

Mixed-effect models (see Table 7) confirmed that there were no reliable differences by condition in proportions of fixations to the non-literal head noun competitor ("Keys") during the head noun and local noun regions (time windows 1 and 2). However, mixed-effect models showed clear differences in proportions of fixations to the non-literal competitor (e.g. "Keys") after the anomalous verb ("were") appeared, as shown in Table 7.

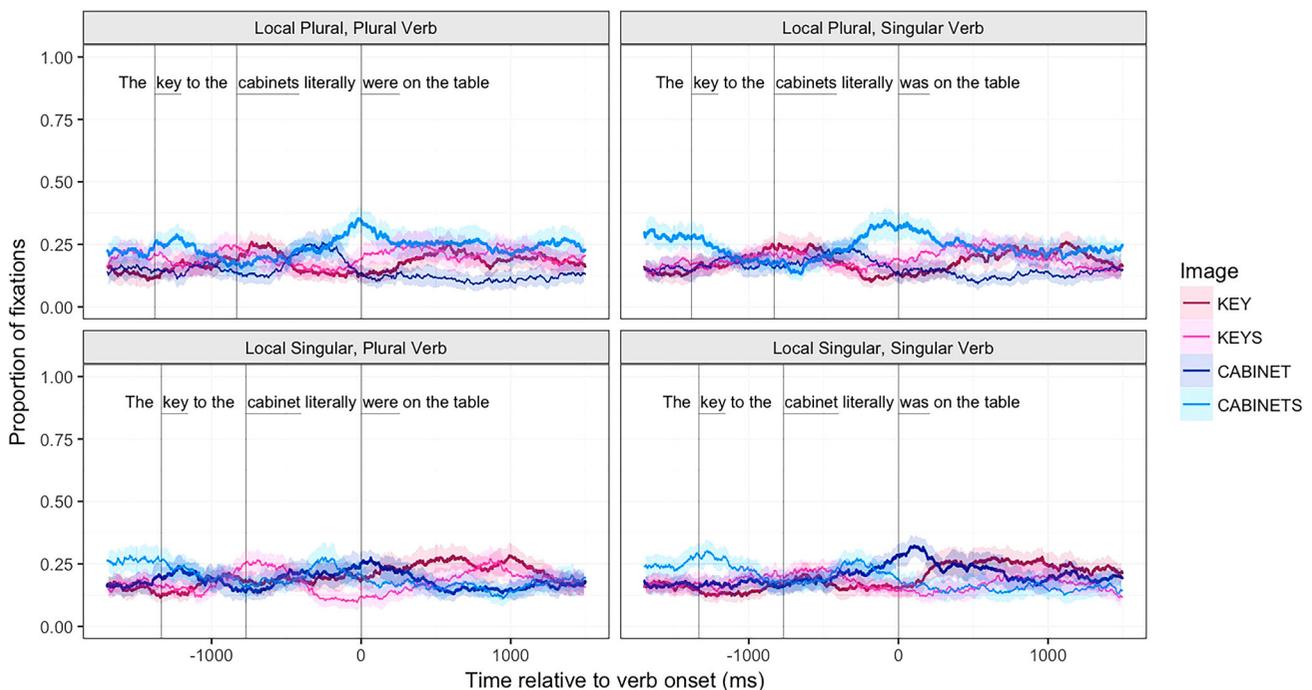


Figure 7. Average proportion of fixations to images in the listening phase of SVA trials where all inflections were repeated veridically (identical and other trials), zeroed to reflect verb onset. Vertical lines represent onsets of head nouns, local nouns, and verbs in audio recording; panels represent sentence conditions. Confidence bands are 95% CIs from a non-parametric bootstrap (1000 iterations) sampled over participants with replacement at 10 msec intervals.

Table 7. Results from linear mixed-effect regression of proportion of fixations to head noun number competitor binned at 500 msec time windows for SVA trials where all inflections were repeated correctly (identical and other trials).

Fixed effects	Window 1: Head noun region			Window 2: Local noun region			Window 3: Verb region			Window 4: Post-Verb region		
	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$
Intercept	0.20	0.01	<0.001	0.17	0.01	<0.001	0.20	0.01	<0.001	0.21	0.02	<0.001
Local Noun	-0.00	0.02	0.85	0.01	0.01	0.64	-0.07	0.02	<0.001	-0.00	0.02	0.85
Verb	-0.01	0.02	0.68	0.02	0.01	0.13	-0.01	0.02	0.72	-0.03	0.02	0.11
Trial Number	-0.01	0.01	0.11	-0.00	0.01	0.51	-0.03	0.01	<0.01	-0.02	0.01	0.02
Local Noun × Verb	0.01	0.03	0.97	0.02	0.03	0.54	-0.03	0.03	0.36	-0.01	0.03	0.78
Local Noun × Trial	0.03	0.02	0.08	0.02	0.02	0.19	0.02	0.02	0.30	-0.01	0.02	0.59
Verb × Trial	-0.02	0.02	0.28	-0.03	0.02	0.09	-0.02	0.02	0.30	0.05	0.02	<0.01
Loc Noun × Verb × Trial	0.04	0.03	0.22	0.04	0.03	0.22	0.05	0.03	0.12	-0.04	0.03	0.29
Random effects	Groups			Groups			Groups			Groups		
	Item	Intercept	0.03	Item	Intercept	0.04	Item	Intercept	0.04	Item	Intercept	0.06
	Participant	Intercept	0.03	Participant	Intercept	0.04	Participant	Intercept	0.06	Participant	Loc Noun	0.03
	Residual		0.27	Residual		0.27	Residual	Loc Noun	0.01	Residual	Intercept	0.06
									0.29		Loc Noun	0.07
											Verb	0.05
											Trial	0.03
											Loc Noun × Verb	0.03
											Residual	0.29

Effects are italicised when they would be significant under a 0.05 alpha level, but not an alpha level corrected for multiple comparisons.

In the third time window, representing processing relative to the verb, there was a significant main effect of Local Noun such that items containing a plural local noun had increased fixations to the non-literal head (21% for plural vs. 16% for singular local nouns). There was also a main effect of Trial Number such that fixations to the non-literal head decreased during the experiment (from 24% in Block 1 to 17% in Block 4), with no other significant main effects or interactions.

In the fourth time window, representing processing after hearing the verb and integrating it with the prior context, there was a significant interaction between Verb and Trial Number, such that plural verbs increased fixations to the non-literal head, but only in the beginning of the experiment (from 30% for plural vs. 23% for singular in Block 1 to 18% for plural vs. 24% for singular in Block 4). There was also a main effect of Trial Number such that fixations to the non-literal head decreased throughout the experiment (from 26% in Block 1 to 21% in Block 4), with no other significant main effects or interactions.

3.2.3. Discussion

We showed that SVA items (*The key to the cabinet(s) *were/was*) containing local plural nouns and plural verbs elicited more repetitions where changes were made to inflections than items with local singular nouns and singular verbs; this replicates the offline results of Experiment 1a with an implicit task. Non-literal repetitions most often resulted in grammatical utterances, which supports the notion that the main driver of non-literal repetitions is gist memory for the prompt's meaning (as in Erlam, 2006; Potter & Lombardi, 1990). Ungrammatical non-literal utterances were produced more often following plural local nouns; this suggests that in addition to measuring attraction in comprehension, this paradigm also elicits attraction errors due to language production constraints (e.g. Bock & Miller, 1991; see Tanner et al., 2014 for discussion of differences between attraction in comprehension and production).

When we focused on the online processing during trials for which all inflections were correctly reproduced, participants looked less often at the literal and non-literal head nouns ("Key/Keys") and more often at the local noun ("Cabinet"/"Cabinets") than in Experiment 1a, despite listening to the same sentence prompts. As such, the data from Experiment 2a more clearly align with the canonical visual-world pattern of looking to mentioned objects (e.g. Allopenna et al., 1998). However, the most important patterns still replicated what we found in Experiment 1a. Local plural nouns (*cabinets*) and anomalous verbs (*were*) increased

fixations to the non-literal plural head (“Keys”) in the same time windows as in Experiment 1a. Specifically, local plural nouns increased fixations to the non-literal plural head (“Keys”) at the region following the verb anomaly on all trials, whereas plural verbs affected fixations to the non-literal plural head at the final region of the sentence in the beginning of the experiment only. We also observed that fixations to the non-literal head noun tended to attenuate in general throughout the experiment, doing so significantly in time windows 3 and 4. As in Experiment 1a, this represents adaptation to syntactic anomalies within the context of the experiment. Both patterns underscore further that, consistent with the noisy channel framework, non-literal processing occurs due to the quick revision of sentence meanings while listening.

3.3. Experiment 2b

Experiment 1b examined items containing a potential semantic anomaly (Without-off blends: *Lulu visited the gym without her hat ?off late yesterday night*). We showed that these items elicited more selections of a non-literal sentence subject (“No-Accessory”) in a metalinguistic judgment task compared to non-anomalous control items, and demonstrated that fixations to a “Yes-Accessory” agent after hearing the preposition depended on the interpretation of the utterance. The goal of Experiment 2b was to assess whether these patterns replicated with an implicit comprehension task. The prediction was that items containing the potentially-anomalous preposition *off* should lead to more non-literal repetitions where the sentence was reproduced in an “unblended” fashion by either changing the *without* or the preposition, and that patterns of fixations to a non-literal agent after hearing the preposition would depend upon how the utterance was repeated.

3.3.1. Analysis

The analysis of sentence repetitions examined the log odds of sentences being produced in a non-literal fashion by making changes to the preposition or *without* (changing *without* to *with*, deleting the preposition, or changing the preposition), compared to all other repetitions. Identical repetitions were coded as productions where all content words matched the original prompt and included trials with disfluencies or restarts, and the remaining trials were classified as other repetitions. These were most commonly trials where the preposition and *without* matched the original prompt but the name was mis-recalled or the predicate was changed. As in Experiment 1a, we contrasted non-

literal repetitions with all other trials (identical and other trials) because non-literal repetitions involving changes to the *without* and preposition were the trials in which the referent of the sentence could be changed.

The eye-tracking analysis was performed on all trials receiving a clear literal interpretation of the referent (correct and other repetitions) with the same time windows as in Experiment 1b. As in Experiment 1b, we chose to only analyse trials receiving a literal interpretation of the referent (identical and other trials) because there were relatively few non-literal repetitions overall. This analysis included the centered predictor Preposition (*on*: -.48 vs. *off*: .52). Scaled and centered Trial Number was also entered as a predictor.

3.3.2. Results

Pragmatic distractor task. Prompts were judged as “likely” the majority of the time (Without-on: 82%, Without-off: 82%), with the remaining trials more often rated as “unlikely” (Without-on: 15%, Without-off: 12%), than “not sure” (Without-on: 2%, Without-off: 6%). In a model predicting the odds of “likely” ratings by Preposition with random intercepts by Participants and Items and a random slope for each by Preposition, ratings were statistically equivalent for both prepositions ($\hat{\beta} = -0.49$, $SE = 0.55$, $z = -0.90$, $p = 0.37$).

Sentence repetition. As in Experiment 2a, odds of repeating the item identically to the prompt were somewhat low (64% of all trials), with high rates of other repetitions (preserving referent; 30% of all trials). Changes to the *without* or preposition were infrequent (6% of trials), and most of these changed the referent (5% of all trials; see boldface entries in Table 8). The most common type of non-identical repetition was a type of other response where the time phrase at the end of the sentence was mis-repeated (e.g. *last Wednesday night* -> *yesterday evening*, occurring on 23% of trials; note that these were coded as “literal” repetitions for

Table 8. Response types by category, Experiment 2 Without trials.

Response type	Without-On	without-Off
Identical repetition	272	220
No response	1	1
<i>Non-literal responses (Without and/or preposition changed)</i>		
Preposition changed	7	24
Preposition omitted	4	1
Without -> With	1	5
Without-> With; Preposition changed	1	1
<i>Other changes (Without and preposition maintained)</i>		
Time phrase changed	63	87
Name changed	5	8
Verb phrase changed	17	18
Object changed	2	2
Multiple changes	11	17

Referent different than prompt in bold.

Table 9. Results from logistic mixed-effect regression of repetitions containing without/preposition changes, versus identical and other repetitions, for Experiment 2 Without trials.

Fixed effects	$\hat{\beta}$	SE	$p(z)$
Intercept	-3.96	0.47	<0.001
Preposition	-1.20	0.42	<0.001
Trial Number	-0.55	0.20	0.01
Preposition \times Trial Number	-0.33	0.40	0.41
Random effects			SD
	Item	Intercept	0.76
	Participant	Intercept	1.34

the analyses). See Table 8 for responses by category and condition.

As in Experiment 1b, odds of non-literal repetitions of the *without* or preposition (whether preserving or changing referent) were affected by Preposition and Trial Number (see Table 9). As shown in Figure 8 (bright and dark blue bars), anomalous *without + off* items led to significantly more non-literal repetitions than control *without + on* items. Significantly fewer non-literal repetitions occurred at the end of the experiment (from 9% in Block 1 to 3% in Block 4). Qualitatively, most of the changes in non-literal repetitions resulted in a change of referent (86%), especially in the anomalous *without + off* trials (see Figure 8; bright blue bars); this was not analysed statistically due to low power.

Proportion of fixations. Proportions of fixations to trials split by repetition type (non-literal = changes to *without* or preposition; literal = identical and other trials) by condition appear in Figure 9. The overall pattern mirrored Experiment 1b: Participants fixated both agents (“Yes-Accessory” and “No-Accessory”) until hearing the word *without*, when attention was directed to the “No-Accessory” agent. Then, about 800 milliseconds after hearing the preposition *off*, attention was redirected to the “Yes-Accessory” agent if the utterance was interpreted literally.

Mixed-effect models (see Table 10) focused only on literally interpreted trials (upper left and bottom right panels of Figure 9) due to the overall low rates of non-literal sentence repetitions. These models revealed no

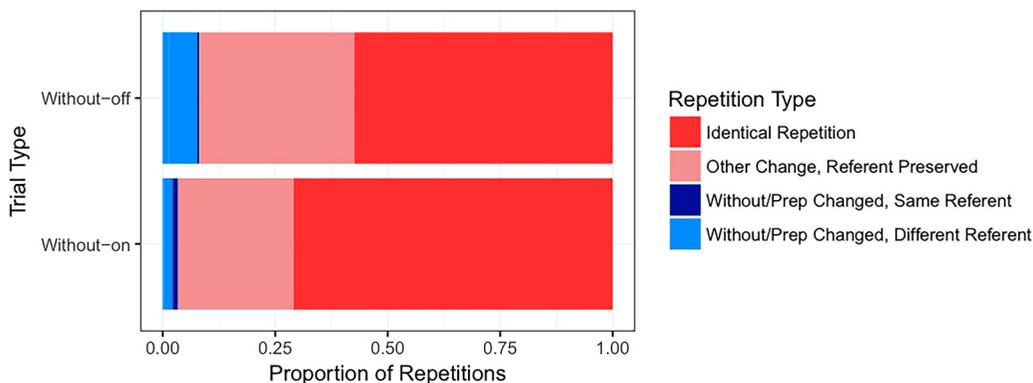
reliable differences in the first two time windows (before the preposition or during processing of the preposition). In the third time window, representing processing after hearing and integrating the anomalous preposition with the rest of the sentence, there was a main effect of Preposition such that *without-off* items received more fixations to the “Yes-Accessory” agent (28% vs. 18%), and a main effect of Trial Number such that more fixations were directed to the “Yes-Accessory” agent throughout the course of the experiment (Block 1: 20%, Block 4: 26%).

Though there were insufficient non-literally interpreted *without-off* items to analyse (comprising only 6% of the data), these showed a similar qualitative pattern as Experiment 1b, with few fixations to the “Yes-Accessory” agent in late time windows and a similar processing profile as the literal *without-on* trials (see Figure 9, upper right panel).

3.3.3. Discussion

For Without anomalies (*without her hat ?off*) compared to control items, sentence prompts were more likely to be repeated with changes to the *without* or the preposition, leading to changes to the referent of the sentence. This replicates the findings of Experiment 1b in an elicited imitation task, providing further evidence for the link between non-literal interpretations and remembered meanings (as in e.g. Erlam, 2006; Potter & Lombardi, 1990).

As in Experiment 1b, we observed that fixations to candidate referents were modulated by the preposition presented in the sentence, such that when *without + off* was interpreted literally, there were more fixations to the “Yes-Accessory” agent after the preposition *off* was presented. This confirms the link between online processing and offline measures of interpretation, consistent with both non-literal processing frameworks. Finally, as in Experiment 1 and Experiment 2a, patterns of

**Figure 8.** Sentence repetition for Experiment 2 Without trials, split by condition.

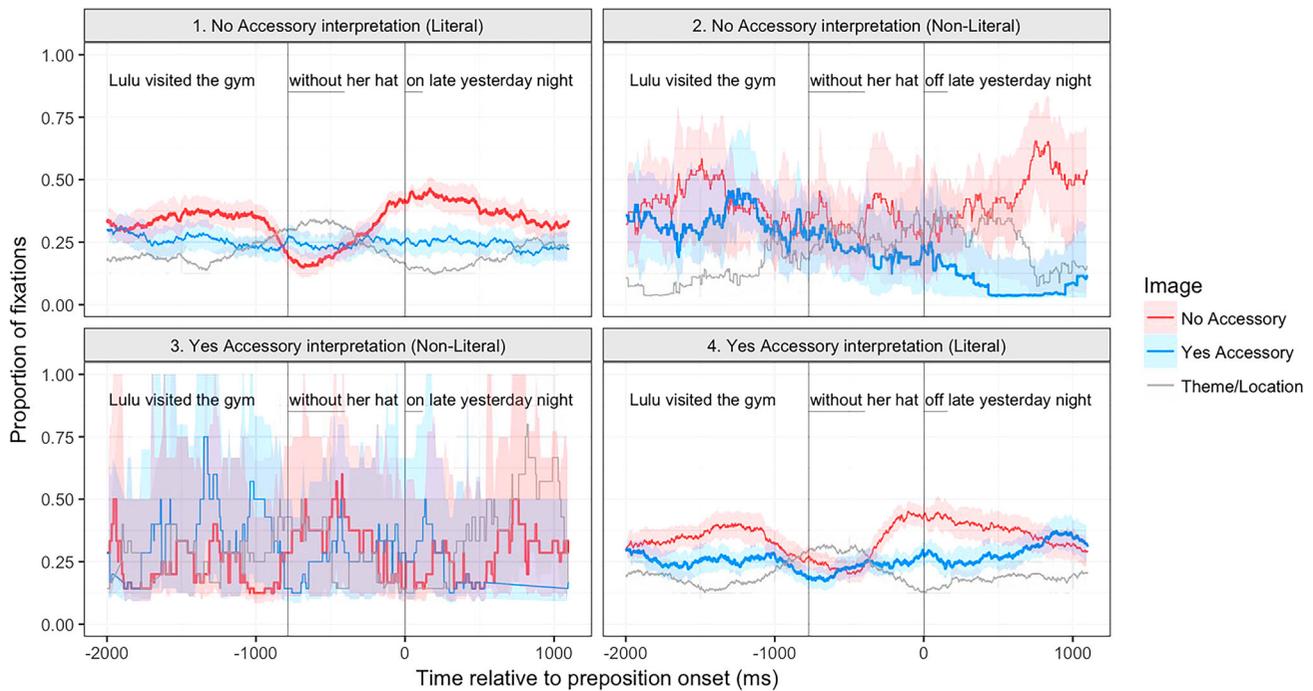


Figure 9. Average proportion of fixations to images in the listening phase of Without trials in Experiment 2, zeroed to reflect preposition onset. Vertical lines represent onsets of without and prepositions; panels represent sentence conditions crossed with sentence interpretation. Confidence bands are 95% CIs from a non-parametric bootstrap (1000 iterations) sampled over participants with replacement at 10 msec intervals. Only literal interpretations (identical and other repetitions) were submitted to analyses due to the sparse number of non-literal repetitions (involving changes to without and/or preposition).

Table 10. Results from linear mixed-effect regression of proportion of fixations to yes-accessory agent binned at 500 msec time windows for literally-interpreted (identical and other repetitions) Experiment 2 Without trials.

Fixed effects	Window 1: Without region			Window 2: Preposition region			Window 3: Post-Preposition region		
	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$	$\hat{\beta}$	SE	$p(\chi^2)$
Intercept	0.19	0.02	<0.001	0.29	0.02	<0.001	0.23	0.02	<0.001
Preposition	-0.01	0.02	0.76	0.02	0.02	0.45	0.10	0.03	<0.001
Trial Number	0.01	0.01	0.56	0.00	0.00	0.23	0.00	0.00	0.03
Preposition \times Trial Number	-0.02	0.02	0.36	0.00	0.00	0.93	-0.00	0.00	0.71
Random effects	Groups			Groups			Groups		
	Item	Intercept	0.04	Item	Intercept	0.04	Item	Intercept	0.00
		Preposition	0.03		Trial Num	0.05	Participant	Intercept	0.06
	Participant	Intercept	0.06	Participant	Intercept	0.05		Preposition	0.10
	Residual		0.27		Preposition	0.05	Residual		0.32
					Trial Num	0.00			
				Residual		0.31			

Effects are italicised when they would be significant under a 0.05 alpha level, but not an alpha level corrected for multiple comparisons.

interpretation and processing changed through the experiment, suggesting adaptation to anomalies with exposure.

4. General discussion

We investigated the time-course and outcomes of processing sentence anomalies non-literally by combining visual world eye-tracking with two different offline measures of interpretation. Experiment 1

used a metalinguistic subject selection task, while Experiment 2 used an elicited imitation task. In both experiments, we examined the processing of subject-verb agreement anomalies (SVA, *The key to the cabinet(s) *were*) and blends of implicit negation (*Without, without her hat ?off*). These two attested anomalies primarily draw upon morphosyntactic and semantic processing, respectively, providing different ways of looking at how online processing leads to offline interpretations.

For both morphosyntactic (SVA, Experiment 1a and 2a) and semantic (Without, Experiment 1b and 2b) anomalies, shifts in attention happened shortly after hearing anomalous or potentially anomalous elements. This shows how quickly changes in interpretation can occur: consistent with current versions of good enough processing (Karimi & Ferreira, 2016) and noisy-channel theories (Gibson et al., 2013), revisions to interpretation happen online. For SVA items, number cues on nouns and verbs led to the consideration of a head noun differing in number at and after hearing the sentence's verb. For Without items, the semantic information in *without* and *on/off* led to the consideration of a different agent upon hearing each word. In both cases, the revision happened as new information appeared.

This provides evidence that participants interpret anomalous sentences incrementally, using the material in the sentence piece-by-piece as it is processed to shape their understanding of sentence meaning. This shows how processing sentence anomalies occurs using the same mechanisms and cognitive infrastructure as non-anomalous sentences (see e.g. Altmann & Kamide, 1999; Snedeker & Trueswell, 2004; Tanenhaus et al., 1995, among many). These results therefore demonstrate the utility of pairing visual-world eye-tracking with offline methods. Doing so shows precisely how and when interpretations of sentences change during processing, and discloses that graded changes in interpretation can be present even when a sentence is ultimately interpreted literally.

4.1. *Contrasting online and offline measures*

A key contribution of the current work relates to the integration of online and offline measures. By comparing measures at different points in processing, we reveal that not all evidence of online difficulty leads to offline interpretation changes. This is essential to consider when examining online or offline measures alone, as many studies in the field of non-literal processing have done. Specifically, we showed that in the anomalous SVA trials (Experiments 1a and 2a), attention was drawn to a plausible non-literal referent even when the literal referent was eventually selected. This shows how anomalies capture attention in situations even when the ultimate interpretation is canonical and highlights the need to consider online processing even when offline performance seems unperturbed. Similarly, in the anomalous Without trials (Experiments 1b and 2b), attention was sometimes drawn to an alternate, repairing referent, and sometimes left on the original referent. This shows how attention to candidate referents can also

be a predictor of the eventual interpretation and highlights how online performance leads to an offline interpretation. Non-literal processing involves both quick revision of interpretations and sustained maintenance of interpretations, and these data disclose that not all online difficulty is obvious when the analysis is limited to offline interpretations.

Multiple offline measures can be used to index sentence interpretations. In order to fully describe the link between online and offline measures of processing, we ran two experiments using different measures. In Experiment 1, we used an explicit measure of interpretation, while in Experiment 2, we used an implicit measure of interpretation; both experiments used the same online task. In both experiments, there was clear evidence for increased rates of non-literal interpretations and increased attention to repairing referents in sentences containing potential anomalies. There were, however, two important differences. First, individuals were more likely to take a non-literal sentence interpretation when asked directly to evaluate sentence meaning; this was particularly clear when contrasting Experiment 2b versus Experiment 1b. This shows a possible role of experimental task in how individuals encode or interpret the meaning of sentences, and it underscores what changes in processing when individuals are asked to repeat speech, versus simply listen.

Second, we observed differences in how attention was recruited based upon the task (e.g. as in Griffin & Bock, 2000); this was particularly clear when contrasting Experiment 2a and Experiment 1a. Despite identical stimuli across the two experiments, the implicit SVA trials in Experiment 2a elicited attention to all four images, while in Experiment 1a, individuals fixated mainly on the item that reflected the sentence subject. This pattern is consistent with individuals focusing more on the auditory than the visual input because we required them to repeat the sentence fragment. The explicit direction of attention changed the overall rate of fixating any object, though it had little impact on relative rates of fixations to images by condition and over time. The implication is that when listening to sentence prompts, individuals explicitly direct their attention based upon the experimental task, which can change overall patterns of online processing but may not affect condition-wise differences. One therefore needs to consider which offline task to use in measuring sentence interpretation; while the overall pattern is comparable, nuances in what is elicited for each task might be important for the researcher's question.

A final consequence of our approach is that it allows us to link non-literal processing with existing sentence processing theories, building a fuller description of

language processing at multiple levels of analysis. As discussed in the Introduction, cue-based retrieval is the currently dominant mechanistic description of agreement error comprehension (e.g. Dillon et al., 2013; Wagers et al., 2009). It states that an erroneous element promotes retrieval of an unlicensed but feature-matching controller from memory. As discussed in Experiment 1a, cue-based retrieval also fits the observed eye-tracking data: individuals attend to both versions of the head noun (“Key”, and “Keys”) before the site of the anomaly, meaning that both are active in memory to be later retrieved. We suggest this means that noisy channel edits, especially those that correct dependencies, could be supported by cue-based retrieval of elements from memory. This points to the importance of considering cognitive mechanisms such as memory and attention in non-literal processing frameworks. We believe this is a promising direction for future work.

4.2. Experience-driven changes in interpretation

While both types of anomalies led to clear evidence of non-literal processing, there were also some differences in how this manifested across anomalies. This sheds important light on the link between experience and the processing of anomalous utterances, an underlying driver of both non-literal processing theories. In Experiment 2a we observed that even anomalous SVA trials tended to be interpreted and repeated with the literal head number, showing that the verb, not the head noun, was typically perceived as the erroneous element in the dependency. This matches statistics learned from the world: agreement errors are typically operationalised as a mis-inflected verb rather than a mis-inflected subject noun (e.g. as in Bock & Miller, 1991), and fits with our earlier work that inferences of mis-inflected verbs increase when speaker-centered sources of noise increase (Brehm et al., 2018). In contrast, we observed that the Without trials elicited more variable interpretations (Experiment 1b) and imitations (Experiment 2b). To our best estimation, these anomalies, while attested (e.g. as in Frazier & Clifton, 2011, 2015), are extremely infrequent. This low frequency could lead to more variability in processing: we suggest that given the low likelihood of utterances of this type, individuals had weaker expectations and were therefore more variable in identifying the locus of the anomaly.

Across items there could also be differences in the likelihood of expecting an anomaly and the type of repair chosen for it because of properties like real-world plausibility, the frequency of nouns in plural versus singular forms, or the co-occurrence of nouns

with *without on/off*. To look at this question, we examined item-level differences in non-literal responding for the subject-verb items, where we had the most item-level power. Items did vary in rates of non-literal responding but did so both within and between experiments. *The vase for the flower(s)* elicited the fewest non-literal responses in Experiment 1a (0%) and *The tag for the gift(s)* elicited the fewest non-literal responses in Experiment 2a (3%). *The banana for the monkey(s)* elicited the most non-literal responses in Experiment 1a (20%) and *The author of the novel(s)* elicited the most non-literal responses in Experiment 2a (63%). The relative inconsistency of non-literal response rates by item across the two experiments might result from the fact that the item set, used in earlier work (e.g. Brehm et al., 2018; a subset of the items also repeat from Patson & Husband, 2016) might not vary sufficiently along important dimensions of real-world experience; a more nuanced exploration of these factors would be worthy of future work.

Changes in processing and interpretation over the course of the experiments also inform the link between experience and processing anomalous utterances. We clearly showed that individuals adapted across the course of the experiment. For both types of anomalies and both types of offline tasks, there were fewer selections of non-literal subjects at the end of the experiment than the beginning, fewer non-literal repetitions, and there was less attention paid to the non-literal referent, especially at late time windows in the critical sentences. This suggests that on all measures for both types of sentences, participants adapted to relative frequency of the anomalous items presented in the experiment, and directed their attention correspondingly, such that competitors became less distracting with experience. This replicates prior work on adaptation to novel or infrequent structures from self-paced reading (e.g. Fine et al., 2013; Fraundorf & Jaeger, 2016) and visual world eye-tracking (Ryskin et al., 2017), and builds upon work showing that even anomalous utterances lead to structural priming (Ivanova et al., 2017). The implication is that we interpret utterances using biases and heuristics about the mapping between intended meaning and observed form, and that these mappings change with experience. While neither the noisy channel theory nor the good enough processing framework make explicit predictions about adaptation over time, both accounts are fully consistent with experience-based changes in processing. Under the noisy channel theory, experience would change the comprehender’s expectations of various sentence types, while under the good enough processing framework, experience could change the likelihood of deploying heuristics

in processing or might cause a comprehender to develop and deploy a new heuristic to easily process sentence meanings.

4.3. Key principles of non-literal processing

The finding that individuals can quickly change their interpretation of a sentence upon encountering an anomaly suggests that individuals maintain uncertainty while processing and adapt accordingly. Earlier work has suggested two ways that uncertainty affects processing. Initial interpretations can be overwritten by a repair that posits an error in the initial input (e.g. the noisy channel theory; Gibson et al., 2013) or can be maintained longer than is ideal (e.g. good-enough processing, Christianson et al., 2001); both lead to non-literal processing. We observed both of these patterns in the current experiments. Our data suggest that non-literal SVA interpretations happen via repair, such that an anomaly forces a revision of the original sentence subject's number (from singular to plural), similar to how repair disfluencies encourage a re-parse (e.g. Arnold et al., 2003; Lowder & Ferreira, 2016). Importantly, while the repair is more frequent for sentences with ungrammatical plural verbs, it also occurs for sentences with grammatical singular verbs (as first discussed in Patson & Husband, 2016), and regardless of eventual interpretation, plural nouns and plural verbs both triggered looks to a corrected version of the head. This demonstrates that while ungrammatical utterances are more likely to trigger a repair to the literal utterance, perfectly grammatical but unexpected ones do too. In contrast, non-literal Without interpretations happen by maintaining an original interpretation (e.g. "No-Accessory") in the face of potentially-incongruous input, suggesting the lingering influence of an initial syntactic parse (e.g. Christianson et al., 2001; Slattery et al., 2013), while the literal ones happen because of a successful reparse of the initially-positing input that is no longer "good enough" but fully veridical. This is particularly clear in Experiment 1b, where differences in early time windows disclosed that if participants directed their attention away from the "Yes-Accessory" agent before hearing the sentence anomaly, they were less likely to form a non-literal sentence interpretation.

These differences underscore that while there are many cases that are covered well by both the noisy channel framework and good-enough processing, such as garden-path sentences, there are important differences in the primary predictions of each framework. We show that core properties of the noisy channel theory ("overwrite and repair") and the good-enough processing framework ("persist in the

face of conflict") provide important and non-identical ways to describe how listeners understand anomalous sentences. We have shown the primary importance of speedy repair for the morphosyntactic SVA anomalies, and the primary importance of lingering representations for the semantic Without anomalies; this implies that mechanisms central to both theories are important to language processing, but does not clearly show whether each is uniquely important in each case. To refine our knowledge of how each mechanism impacts comprehension, and in order to distinguish the relative contribution of each mechanism in everyday language use, future work is needed. Doing so could build a fully-integrated non-literal processing theory that cuts across algorithmic and cognitive levels of processing, uniting two frameworks previously developed to account for different types of anomaly processing. We have taken a step towards unifying these theories in the present work, showing that both describe important principles comprehenders can use to process non-literally. We hope that the paradigm outlined in the present work can be used as a tool to carefully build further theory. For example, one could investigate how anomalies of different grain sizes (words, affixes, phonemes/letters) at different processing levels (semantic, syntactic) with varied real-world frequencies affect what mechanisms comprehenders use to understand realistically anomalous speech; doing so would further properties of interest to both good-enough and noisy-channel processing.

5. Conclusion

Language processing is a hard task that feels remarkably easy. By linking the domains of non-literal processing and visual world eye-tracking, we show that individuals process language anomalies quickly and efficiently, adapting to new information as it appears. By integrating meaning, form, and expectations, we show that individuals can repair or overlook anomalies present in their linguistic input. This leads to the successful transmittal of a message even when its form has gone awry.

Notes

1. Note the adverb in these sentences: in agreement comprehension studies, adding an adverb before the verb serves as a spillover region to better separate processing of local noun and verb number (e.g. Wagers et al., 2009).
2. Centering contrasts allows the intercept and other main effects to reflect the average of the two conditions.

3. While trial was entered as a continuous predictor, we report means by block for ease of exposition.
4. If the three-way interaction is included in the model, only Verb approaches significance ($p = .051$) and all fixed effects are correlated at .9 or above.
5. There were too few Without-On non-literal interpretations to analyze fixations.
6. These are weighted unevenly due to the differing number of observations in each cell; cells with fewer observations get more weight to center the factor so that main effects reflect the average of conditions.

Acknowledgements

Supported by a grant from the National Science Foundation (NSF OISE 0968369; PI: J. F. Kroll; co-PIs: P. E. Dussias and J. G. van Hell). Thanks to Neil Shook for assistance with stimuli and participant testing, to Quiqi Cong for assistance creating stimuli, to Alison Kelly and Katherine Muschler for assistance with participant testing, and to Jennifer Rhee and Adrianna Shevlin for assistance with transcription; thanks to Rachel Ryskin for feedback on study design and analysis and to Antje Meyer for manuscript comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Science Foundation [grant number 0968369].

References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language*, 38(4), 419–439. <https://doi.org/10.1006/jmla.1997.2558>
- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264. [https://doi.org/10.1016/S0010-0277\(99\)00059-1](https://doi.org/10.1016/S0010-0277(99)00059-1)
- Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76(1), B13–B26. [https://doi.org/10.1016/S0010-0277\(00\)00073-1](https://doi.org/10.1016/S0010-0277(00)00073-1)
- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal there, um, new information. *Journal of Psycholinguistic Research*, 32(1), 25–36. <https://doi.org/10.1023/A:1021980931292>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory & Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Dai, B. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-13). <https://cran.r-project.org/web/packages/lme4/index.html>
- Blanchette, F. (2013). Negative concord in English. *Linguistic Variation*, 13(1), 1–47. <https://doi.org/10.1075/lv.13.1.01bla>
- Blanchette, F., & Lukyanenko, C. (2019). Unacceptable grammars? An eye-tracking study of English negative concord. *Language and Cognition*, 11(1), 1–40. <https://doi.org/10.1017/langcog.2019.4>
- Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Lawrence Erlbaum.
- Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1), 45–93. [https://doi.org/10.1016/0010-0285\(91\)90003-7](https://doi.org/10.1016/0010-0285(91)90003-7)
- Brehm, L., & Bock, K. (2013). What counts in grammatical number agreement? *Cognition*, 128(2), 149–169. <https://doi.org/10.1016/j.cognition.2013.03.009>
- Brehm, L., & Bock, K. (2017). Referential and lexical forces in number agreement. *Language, Cognition & Neuroscience*, 32(2), 129–146. <https://doi.org/10.1080/23273798.2016.1234060>
- Brehm, L., Jackson, C. N., & Miller, K. L. (2018). Speaker-specific processing of anomalous sentences. *The Quarterly Journal of Experimental Psychology*, Advance Online Publication.
- Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory & Language*, 53(2), 292–313. <https://doi.org/10.1016/j.jml.2005.03.003>
- Butterworth, B. (1982). Speech errors: Old data in search of new theories. In A. Cutler (Ed.), *Slips of the tongue and language production* (pp. 73–109). Walter de Gruyter.
- Chrabaszcz, A., & Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: A cross-linguistic comparison. *Second Language Research*, 30(3), 351–379. <https://doi.org/10.1177/0267658313493432>
- Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4), 368–407. <https://doi.org/10.1006/cogp.2001.0752>
- Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 36(2), 538. <https://doi.org/10.1037/a0018027>
- Christianson, K., & Luke, S. (2011). Context strengthens initial misinterpretations of text. *Scientific Studies of Reading*, 15(2), 136–166. <https://doi.org/10.1080/10888431003636787>
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1), 84–107. [https://doi.org/10.1016/0010-0285\(74\)90005-X](https://doi.org/10.1016/0010-0285(74)90005-X)
- Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition*, 25(1), 57–71. <https://doi.org/10.3758/BF03197285>
- Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory & Language*, 69(2), 85–103. <https://doi.org/10.1016/j.jml.2013.04.003>

- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2017). Multipic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology*, 1–24. <https://doi.org/10.1080/17470218.2017.1310261>
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3), 464–491. <https://doi.org/10.1093/applin/aml001>
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203. [https://doi.org/10.1016/S0010-0285\(03\)00005-7](https://doi.org/10.1016/S0010-0285(03)00005-7)
- Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10), e77661. <https://doi.org/10.1371/journal.pone.0077661>
- Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of Memory & Language*, 91, 28–58. <https://doi.org/10.1016/j.jml.2016.05.006>
- Frazier, L., & Clifton, C. (2011). Quantifiers undone: Reversing predictable speech errors in comprehension. *Language*, 87(1), 158. <https://doi.org/10.1353/lan.2011.0024>
- Frazier, L., & Clifton, C. (2015). Without his shirt off he saved the child from almost drowning: Interpreting an uncertain input. *Language, Cognition & Neuroscience*, 30(6), 635–647. <https://doi.org/10.1080/23273798.2014.995109>
- Fukumura, K., van Gompel, R. P., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *The Quarterly Journal of Experimental Psychology*, 63(9), 1700–1715. <https://doi.org/10.1080/17470210903490969>
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056. <https://doi.org/10.1073/pnas.1216438110>
- Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, 28(6), 703–712. <https://doi.org/10.1177/0956797617690277>
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279. <https://doi.org/10.1111/1467-9280.00255>
- Grodner, D. J., Klein, N. M., Carbarry, K. M., & Tanenhaus, M. K. (2010). Some and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42–55. <https://doi.org/10.1016/j.cognition.2010.03.014>
- Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3), 376–415. <https://doi.org/10.1016/j.cogpsych.2008.09.001>
- Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171. <https://doi.org/10.1016/j.actpsy.2010.11.003>
- Ivanova, I., Branigan, H. P., McLean, J. F., Costa, A., & Pickering, M. J. (2017). Do you what I say? People reconstruct the syntax of anomalous utterances. *Language, Cognition and Neuroscience*, 32(2), 175–189. <https://doi.org/10.1080/23273798.2016.1236976>
- Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, 69(5), 1013–1040. <https://doi.org/10.1080/17470218.2015.1053951>
- Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086–21090. <https://doi.org/10.1073/pnas.0907664106>
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Empirical Methods in Natural Language Processing (EMNLP)*, 234–243.
- Lowder, M. W., & Ferreira, F. (2016). Prediction in the processing of repair disfluencies: Evidence from the visual-world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1400. <https://doi.org/10.1037/xlm0000256>
- Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *The Quarterly Journal of Experimental Psychology*, 69(5), 950–971. <https://doi.org/10.1080/17470218.2014.992445>
- Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory & Language*, 41(3), 427–456. <https://doi.org/10.1006/jmla.1999.2653>
- Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory & Language*, 29(6), 633. [https://doi.org/10.1016/0749-596X\(90\)90042-X](https://doi.org/10.1016/0749-596X(90)90042-X)
- Qian, Z., Garnsey, S., & Christianson, K. (2018). A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition & Neuroscience*, 33(2), 227–254. <https://doi.org/10.1080/23273798.2017.1379606>
- R Core Team. (2014). *R: A language and environment for statistical computing version 3.1.2*. R Foundation for Statistical Computing. <http://www.R-project.org/>
- Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2), 217–236. <https://doi.org/10.1068/p5117>
- Ryskin, R., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2017). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 43(5), 781–794. <https://doi.org/10.1037/xlm0000341>
- Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, 6(9), 382–386. [https://doi.org/10.1016/S1364-6613\(02\)01958-7](https://doi.org/10.1016/S1364-6613(02)01958-7)
- Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, 10, 1002. <https://doi.org/10.3389/fpsyg.2019.01002>
- Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory & Language*, 69(2), 104–120. <https://doi.org/10.1016/j.jml.2013.04.001>
- Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49(3), 238–299. <https://doi.org/10.1016/j.cogpsych.2004.03.001>

- Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, 49(1), 1–46. <https://doi.org/10.1016/j.cogpsych.2003.10.001>
- Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1), 201–216. <https://doi.org/10.3758/MC.36.1.201>
- Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation & Change*, 24(02), 135–178. <https://doi.org/10.1017/S0954394512000129>
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268 (5217), 1632–1634. <https://doi.org/10.1126/science.7777863>
- Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, 76, 195–215. <https://doi.org/10.1016/j.jml.2014.07.003>
- Tortora, C., & den Dikken, M. (2010). Subject agreement variation: Support for the configurational approach. *Lingua. International Review of General Linguistics. Revue internationale De Linguistique Generale*, 120(5), 1089–1108. <https://doi.org/10.1016/j.lingua.2009.04.004>
- Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory & Language*, 61(2), 206–237. <https://doi.org/10.1016/j.jml.2009.04.002>