The role of production expectations in visual world paradigm linking hypotheses

Judith Degen (jdegen@stanford.edu)
Department of Linguistics, 450 Jane Stanford Way,
Stanford, CA 94305 USA

Stefan Pophristic (stefpop@stanford.edu)
Department of Linguistics, 450 Jane Stanford Way,
Stanford, CA 94305 USA

Abstract

While widely used in psycholinguistics, the linking hypothesis for eye movements in the visual world paradigm is still poorly understood. Recent work on linking hypotheses for referential tasks in particular has found mixed support for the *Referential Belief Link*: that the proportion of looks to a referent in a time window reflects participants' degree of belief that the referent is the intended target in that time window. Here we test the hypothesis that participants' expectations for the utterances observed in an experiment modulate the extent to which the Referential Belief Link holds. This hypothesis is motivated by a simple idea: when utterances are unexpected, listeners engage in additional reasoning to make sense of the observed signal. In a re-analysis of a previous eye movement and incremental decision task dataset, in conjunction with two novel production experiments, we find that the more surprising an observed utterance is, the smaller the correlation between explicit and implicit beliefs is. We discuss the importance of participants' production expectations in research using the visual world paradigm.
Keywords: psycholinguistics; experimental pragmatics; scalar implicature; linking hypothesis; visual world paradigm; eye-tracking

Introduction

The visual world paradigm (VWP) is widely used in psycholinguistics. In the VWP, participants' eye movements are recorded as they listen to unfolding speech while viewing visual scenes like that in Fig. 1. Research using the VWP has had tremnedous success in furthering our understanding of phonetic, lexical, syntactic, prosodic, semantic, and pragmatic processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Allopenna, Magnuson, & Tanenhaus, 1998; Altmann & Kamide, 1999; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Huang & Snedeker, 2009; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014).

The VWP is popular for good reason: eye movements can be interpreted as an indicator of attention that is closely time-locked to the linguistic signal. Language can guide eye movements to a region of interest in a display within 200 ms (Allopenna et al., 1998). By sampling an x/y coordinate every few milliseconds, researchers thus obtain a temporally fine-grained record of participants' language-directed attention over the course of an unfolding utterance. This property has been particularly useful in resolving questions regarding the time-course of online language processing, which typically cannot be addressed using offline measures like forced choice, truth-value judgments, or even more coarse-grained temporal measures like response times from button presses.

Despite its general success, the linking hypothesis for the VWP – that is, how to link observed eye movements to the underlying mental processes that generate them – is still poorly understood (Salverda & Tanenhaus, 2017; Tanenhaus, Magnuson, Dahan, & Chambers, 2000; Allopenna et al., 1998; Magnuson, 2019). The problem of how to interpret eye movement patterns is compounded by the fact that the VWP is used for vastly different tasks (for an overview, see Huettig, Rommers, & Meyer, 2011).

In this work we focus on active referential tasks, in which participants' goal is to identify and select the speaker's intended referent. In such tasks, eye movements are assumed to reflect listeners' active search for or belief in the referent. A way of formalizing this is the *Referential Belief Link* (Degen, Kursat, & Leigh, 2021), first proposed by Allopenna et al. (1998):

$$p_{\text{empirical}}(r|u) \propto p_{\text{belief}}(r = \text{target}|u) \tag{1}$$

This captures the idea that the empirical proportion of looks $p_{\text{empirical}}$ to a referent r in a time window in response to a (possibly partial) utterance u reflects participants' degree of belief p_{belief} that r is the intended target.[1]

This linking hypothesis implicitly underlies much work in the VWP using referential tasks and is thus important to test explicitly. Recent work doing so has found mixed support for it (Qing, Lassiter, & Degen, 2018; Degen et al., 2021). In these studies, previous eye-tracking datasets were re-analyzed with respect to how closely proportions of eye movements to a referent within a time window correlated with explicitly elicited referential beliefs (Qing et al., 2018; Degen et al., 2021). Explicit referential beliefs were collected in an offline *incremental decision task* (similar to gating tasks, Allopenna et al., 1998; Kreiss & Degen, 2020). In a re-analysis of an adjective processing dataset (Leffel, Xiang, & Kennedy, 2016), Qing et al. (2018) found low or no correlations between explicit beliefs and eye movements (all $r \in [0.06, 0.46]$ in the theoretically relevant window of analysis). In contrast, the same methodology applied to a quantifier process-

[1] As Degen et al. (2021) caution, the assumption of proportionality may be too strong. A weaker version is that $p_{\text{empirical}}$ is monotonically increasing in p_{belief}.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ing dataset (Sun & Breheny, 2020) found high correlations (all $r \in [0.79, 0.96]$ in the theoretically relevant window of analysis, Degen et al., 2021).

What determines the observed variability in the extent to which the Referential Belief Link holds both across and within studies? Qing et al. (2018) propose an interesting hypothesis, motivated by the idea that there is a tradeoff between *exploration* and *exploitation*: when participants are less familiar with the objects in the scene and the ways of referring to them, eye movements might serve a more exploratory purpose, i.e., to establish the referent options and how speakers might refer to them. In contrast, with more familiarity with possible referents and ways of referring to them, participants might have more resources available for exploiting their signal-driven beliefs.

A prediction of this speculative idea is that if listeners observe a less expected utterance, they will need to explore the scene more, e.g., to evaluate which possible referents are compatible with the utterance, and hence the correlation between proportions of looks and explicit referential beliefs should be weaker. In contrast, if listeners hear a more expected utterance, they can directly exploit the signal, and hence the correlation between proportions of looks and beliefs should be stronger. Qing et al. (2018) found preliminary evidence for this prediction: while the adjectivally modified referring expressions used in the original study were rarely produced naturally in a free production experiment, the single condition in which explicit beliefs were predictive of eye movements was also the condition in which the observed referring expressions had a non-zero (albeit still very low) probability of being produced naturally.

Here we report a direct test of the hypothesis that listeners' expectations for the utterances observed in a visual world eye-tracking experiment modulate the extent to which the Referential Belief Link holds. We do so by testing the extent to which participants' quantifier production expectations, estimated in two free production tasks, predict the correlation between implicit and explicit beliefs as measured in the quantifier processing datasets collected by Sun and Breheny (2020) and Degen et al. (2021).

Research strategy and test domain

We used the displays of Sun and Breheny (2020) (see Fig. 1) to elicit natural referring expressions in two written free production tasks. This allowed us to compute a proxy measure for participants' quantifier production expectations in the original experiment as the surprisal of each quantifier in the two production tasks.[2] If production expectations modulate the strength of the Referential Belief Link, quantifier surprisal should predict the correlations between implicit referential beliefs measured via eye movements in the VWP (Sun & Breheny, 2020) and explicit referential beliefs elicited in an incremental decision task (Degen et al., 2021).

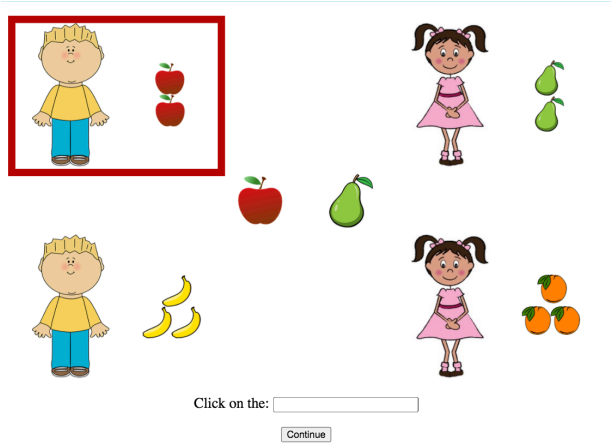

Figure 1: Example display in our written production experiments. The same display types were used in Sun and Breheny (2020)'s eye-tracking Experiment 3 without the red border and text field prompt, paired with instructions like 'Click on the boy with some of the apples.'

We briefly describe the design of the original comprehension studies before introducing our novel production studies and main analysis.

Sun and Breheny (2020) investigated the processing speed of quantifiers. Specifically they were interested in how: 1) pre-existing low-level associations between quantifiers and set sizes (e.g., the preference for the quantifier "all" with larger set sizes of target objects) affect quantifier processing speed, and 2) how the quantifier used affects processing speed, specifically whether certain quantifiers (e.g., "all" or "some") require an additional process of verifying the relationship between referent target objects and other objects in the scene (what they called the 'residue set' in the center of the display). In order to address these questions, their Experiment 3 manipulated the set size of the target objects (big (3 objects) or small (2 objects)) and the quantifier used to describe the objects ("all", "some" or a number). Visual scenes (such as Fig. 1) were shown with auditorily presented instructions of the form "Click on the | GENDER | that has QUANTIFIER | of NAME'S | NOUN |". GENDER was a gender noun ("boy" or "girl") referring to the target child, QUANTIFIER was one of four quantifiers ("all", "some", "two", or "three"), NAME referred to 1 of 3 characters who were introduced in a background story at the start of the experiment, and NOUN was 1 of 12 target objects (4 kitchenware items, 4 stationary items, or 4 fruit). Thus, a participant who saw the display in Fig. 1 may have heard the instruction 'Click on the boy who has some of Susan's apples.' Participants' eye movements were analyzed in different time windows: Baseline, Gender, Quantifier, Name, and Noun (separated by '|' in the example instruction above).

[2]The use of surprisal rather than probability is motivated by ample evidence showing that processing effort as measured in reading times or in N400 amplitudes in ERP studies is linear in word surprisal, not probability (Levy, 2008; Smith & Levy, 2013; Frank, Otten, Galli, & Vigliocco, 2013).

The theoretical window of interest was the Quantifier window. They found that eye movements to the target increased most rapidly in the Quantifier window in response to number terms. Looks increased more slowly for "all" and "some," modulated to some extent by set size, but participants rapidly looked to the residue set in response to these quantifiers, presumably for verification purposes.

Degen et al. (2021) assessed the Referential Belief Link by replicating Sun and Breheny (2020) in an incremental decision task. They presented the original instructions in written rather than auditory form, but revealed them incrementally, one window at a time.[3] After the participant made a guess about the target in one window, the text in the next window was revealed. They then calculated correlations between participants' explicit beliefs (as measured by referent selections within a window) and proportions of looks to referents in Sun and Breheny (2020). Overall, selection data in the Quantifier window was highly correlated with looking data (all $r > .78$), thus supporting the linking hypothesis. Nevertheless, correlations displayed variability, suggesting the Referential Belief Link was not supported to the same extent across conditions.

The present study In order to test whether participants' expectations for observed quantifiers modulate the extent to which the Referential Belief Link holds, we implemented the design of the original studies as free production tasks. Quantifier surprisal was computed as the negative logarithm of the relative frequency of the quantifier within a specific combination of Sun and Breheny (2020)'s conditions and windows – i.e., a combination of the target object set size (big, small), quantifier condition ("all", "some", number), gender of target child within a trial (boy, girl) and target noun within a trial (1 of the 12 objects):

$$\text{surprisal(det)} = \text{-}\log_2 P(\text{det}) = -\log_2 \frac{\text{freq(det)}}{\text{freq(combination)}}$$

Because log(0) is undefined, if a particular quantifier was never produced, we re-assigned that quantifier a probability of 0.0001, resulting in a surprisal of $\sim$13.29.

We computed quantifier surprisal in two different production tasks. The "purest" form of eliciting production would be to allow participants to produce referring expressions following 'Click on the…' with no constraints whatsoever. However, this would have likely resulted in referring expressions without quantifiers. For example, the target boy in Fig. 1 can be referred to simply as 'the boy with apples.' This is a problem for instance for eye-tracking studies using referring expressions with quantifiers to probe the speed with which scalar inferences are drawn (Huang & Snedeker, 2009; Grodner, Klein, Carbary, & Tanenhaus, 2010; Degen & Tanenhaus, 2016; Sun & Breheny, 2020), since it is not clear that pragmatic reasoning about alternatives naturally unfolds in the face of pragmatically infelicitous utterances. A cooperative speaker who insists on using a quantifier should at least produce one that will allow the listener to rapidly identify the target – here, vague quantifiers like "some" should be dispreferred compared to number terms (especially with the small set sizes considered here, Degen & Tanenhaus, 2015; Sun & Breheny, 2020). Despite the importance of utterance alternatives (and listeners' resulting production expectations) in the computation of pragmatic meaning (Franke, 2014; Goodman & Frank, 2016; Peloquin & Frank, 2016; Degen & Tanenhaus, 2015, 2016; Gotzner, Wartenburger, & Spalek, 2016; Sun & Breheny, 2020), little work has been done to estimate how likely the utterances are that researchers provide participants in eye-tracking experiments on quantifier processing. The current study thus also provides novel, principled estimates of quantifier surprisal in the types of contexts frequently used in experimental pragmatics studies.

As mentioned above, the most likely outcome of an entirely unconstrained production task is no quantifier production at all. Given that listeners rapidly update their beliefs about likely utterances in response to exposure (Grodner & Sedivy, 2011; Pogue, Kurumada, & Tanenhaus, 2016; Schuster & Degen, 2020), it is likely that participants in eye-tracking studies, where the same utterance form is observed repeatedly, rapidly form local expectations about likely utterances.

We thus ran two versions of the production task: one in which participants received no exposure to the original comprehension task – allowing for an estimate of non-adaptive quantifier base rates – and one in which participants first completed four comprehension trials from Degen et al. (2021) – allowing for locally adaptive estimates of quantifier rates.

Experiment 1: free production without comprehension trial exposure

Methods

Participants We recruited 51 native English speaking participants in the United States on Prolific. We excluded participants with overall $< 95\%$ accuracy, participants who answered with single word responses on $> 50\%$ of trials, and participants who did not produce target object nouns on $> 50\%$ of trials. These criteria led to no participant exclusions.

Materials and Procedure We used the same design and materials as Sun and Breheny (2020) but changed the task.[4] On each trial, participants told a fictional addressee to click on one of four children in a display by typing into a text box. The target child to communicate was indicated by a red border (see Fig. 1 for an example display).

Each display contained 4 children with an assortment of objects (fruit, kitchenware, or stationary) in four corners of the screen. In the middle, there were extra objects that matched the children's objects. Participants were told that their task was to "*get another player to click on the child*

[3] They collapsed the Name window into the Quantifier window because the name did not provide additional information about the target.

[4] Pre-registrations are available at `https://osf.io/s9fm7`. Experimental materials, data, and analysis scripts are available at `https://tinyurl.com/yp6pyk6p`.

in the red box" by "*complete[ing] the sentence 'Click on the…' by typing what [they] think should come next.*" On each trial they were given the phrase "Click on the" followed by an empty text box into which they could type their answer. Two example sentences were given on the instruction pages: "Click on the boy with two of the apples" and "Click on the boy with some of the apples". These example sentences were included to encourage participants to give responses that could include quantifiers. We only gave two example sentences in order to minimize interference with natural quantifier production. Crucially, participants were not explicitly told that their answers had to match any particular format or be restricted in any way. Therefore they were free to produce any utterance they deemed appropriate.

The experiment followed the design of Sun and Breheny (2020) and included 56 trials: 36 experimental trials, 12 filler trials, and 2 practice trials. Experimental trials were evenly split by whether the target object set size was big (3 objects) or small (2 objects) and by whether participants in Sun and Breheny (2020) originally heard the instruction with the quantifier "all", "some", or a number ("two" or "three"). Targets were counterbalanced for the child's gender (boy or girl) and type of object the child had (fruit, kitchenware, or stationary; total of 12 unique objects).

There were two practice trials which were identical in procedure to the rest of the experiment. For consistency, the practice trial scenarios (sets of children and objects presented in the scene) were taken from Sun and Breheny's (2020) Exp. 3 practice trials. Experimental trial order was randomized.

Results and discussion

Trials were excluded from analysis if participants produced an incorrect gender (n = 3), an incorrect quantifier (n = 6), an incorrect target object noun (n = 17), no target object noun (n = 0), or a single word response (n = 0), resulting in the exclusion of 26 trials (1.5%). We focus on the quantifier window as our main window of analysis.

Quantifiers in participants' productions were classified as belonging to the following categories: "all", "some", number ("two" or "three"), no quantifier, or other (e.g., "most", "least", "both"). Individual surprisal values were calculated for each unique combination of noun (eraser, apple, …) and target set size (big, small).

Fig. 2 shows the mean surprisal for each quantifier category. Of the quantifiers included in the original study, number terms were produced most frequently (1182 of the 1810 trials), yielding a very low surprisal of 0.6 (for both the big and small set size). Much less frequently produced was "some" (48 of the 1810 trials), yielding a relatively higher surprisal of 5.1 (for the big set size) and 6.2 (for the small set size). Perhaps most surprisingly, "all" was never naturally produced, leading to a very high surprisal value (at ceiling). Of note is that the second most preferred utterance type did not include a quantifier at all (e.g., "Click on the boy with apples"). This confirms the intuition that using no quantifier at all to describe the target objects is among the most expected alternatives. All of the reported differences were supported by a mixed effects linear model predicting surprisal from dummy-coded fixed effects of quantifier (reference level: 'some'), set size (reference level: 'big'), and their interaction, with random by-noun intercepts. There were no significant set size effects (all $|t| < 1.5$).

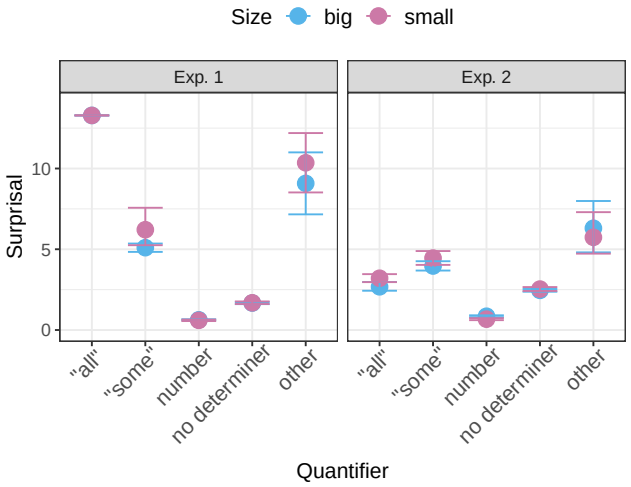

Figure 2: Mean surprisal of quantifiers in Exps. 1 and 2. Error bars indicate 95% bootstrapped confidence intervals. Values were calculated separately for the two different target object set sizes (big and small).

Overall, then, the instructions widely used in studies on pragmatic inferences based on quantifier use (Huang & Snedeker, 2009; Grodner et al., 2010; Degen & Tanenhaus, 2016; Sun & Breheny, 2020) are highly unexpected, raising a question about the generalizability of the results of these studies to contexts where non-number quantifiers are expected. However, as acknowledged above, listeners rapidly adapt to the updated statistics of their linguistic environments. Thus, a fairer estimate of quantifier production expectations might be obtained after brief exposure to the original comprehension task. This was the goal of Exp. 2.

Experiment 2: free production with comprehension trial exposure

Methods

Participants We recruited 51 native English speaking participants in the United States from Prolific. We excluded participants using the same criteria as in Exp. 1, which led to the exclusion of (n = 2) participants.

Materials and Procedure Exp. 2 was identical to Exp. 1 with the exception of 4 additional practice trials. The scenarios (children and objects shown on the screen) were identical to trials in Degen et al. (2021)'s incremental decision study. However, the target utterances were slightly modified so that all possible quantifiers from Sun and Breheny (2020) would be observed within the four trials. Participants were given an

instruction of the form "Click on the | GENDER | that has QUANTIFIER of NAME'S | NOUN |" incrementally and selected what they believed to be the target of the utterance in each window (marked by '|'). The four practice trial sentences were:

1. Click on the boy that has two of Susan's apples.
2. Click on the girl that has some of Susan's apples.
3. Click on the boy that has three of Michael's scissors.
4. Click on the girl that has all of Michael's rulers.

These practice trials were then followed by the instructions and practice trials from Exp. 1. Participants were never explicitly told that their answers had to match any particular format or be restricted in any way.

Results and discussion

Trials were excluded from analysis if participants produced an incorrect gender (n = 2), an incorrect quantifier (n = 6), an incorrect target object noun (n = 16), no target object noun (n = 0), or a single word response (n = 13), resulting in the exclusion of 37 trials (2.1%).

Surprisal means by quantifier and target set size are shown in Fig. 2. Notably, while number was produced at similar rates as in Exp. 1 (1019 out of 1771 trials; surprisal = 0.86 (big set) and 0.67 (small set)), both "some" (106 out of 1771 trials; surprisal = 3.97 (big set) and 4.46 (small set)) and especially "all" (240 out of 1771 trials; surprisal = 2.68 (big set) and 3.21 (small set)) were produced more frequently. In fact, "all" was preferred over "some", though both were marginally more preferred with big than with small sets. As in Exp. 1, these differences were confirmed by mixed effects models.

These results suggest that of the three quantifiers typically included in studies on scalar inferences – "all", "some", and small number terms like "two" and "three" – only the number terms are naturally preferred to the unquantified alternative in production. Given the evidence that listeners have probabilistic production expectations that track the actual statistics of words in the world (Levy, 2008; Frank et al., 2013), the results suggest that only number terms are naturally expected in this paradigm. This may explain why number terms in such studies typically lead to much faster target identification than either "all" or "some" (Huang & Snedeker, 2009; Degen & Tanenhaus, 2016; Sun & Breheny, 2020).

Correlation Analysis

We can now assess the hypothesis that the extent to which the Referential Belief Link holds is modulated by participants' production expectations. To do so, we tested whether quantifier surprisal is a predictor of the correlation between implicit and explicit beliefs as reported by Degen et al. (2021) for the quantifier window of Sun and Breheny (2020).[5]

To this end we ran two linear models that predicted the quantifier window correlations (computed at the level of unique combinations of quantifier, target set size, and target child gender) from the surprisal values obtained in Exps. 1 and 2, respectively. Fig. 3 shows Quantifier window correlations against surprisal values.

Surprisal as estimated in Exp. 1 was a marginally significant predictor of the correlations (β=-0.006, SE=0.003, t=-2.003, p <.08). Surprisal as estimated in Exp. 2 was a significant predictor of the correlations (β= -0.039, SE=0.011, t=-3.438, p <.007). Exp. 2 surprisal was a better predictor of correlations than Exp. 1 surprisal, as evidenced in twice the variance explained (Exp. 1 adjusted $R^2 = 0.22$, Exp. 2 adjusted $R^2 = 0.50$). This improvement is largely driven by the difference in estimates for "all" surprisal and provides indirect evidence that listeners indeed rapidly formed experiment-specific quantifier expectations.

These results thus suggest that the more expected the quantifier was in the original experiment, the more participants' explicit beliefs predicted implicit beliefs, i.e., the more strongly the Referential Belief Link held.

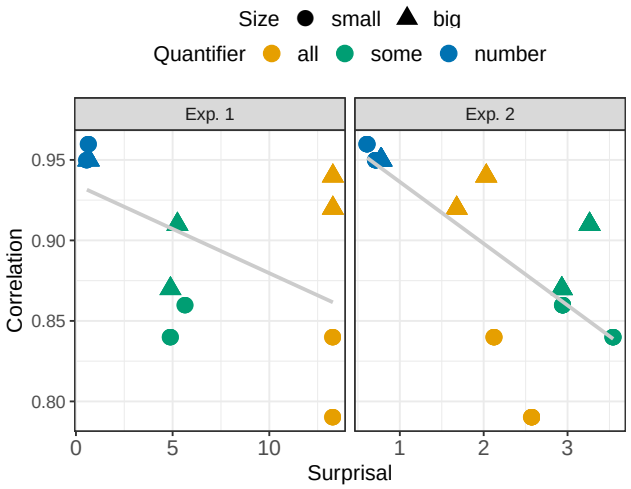

Figure 3: Correlation between implicit and explicit beliefs in the quantifier window of Sun and Breheny (2020) and Degen et al. (2021) against quantifier surprisal. Gray line indicates regression line. Surprisal was calculated over combinations of quantifier, target object set size, and target child gender.

General Discussion

Despite numerous attempts at shedding light on the issue of how to link observed eye movements to the underlying mental processes that generate them (Salverda & Tanenhaus, 2017; Tanenhaus et al., 2000; Allopenna et al., 1998; Magnuson, 2019), linking hypotheses for the visual world paradigm are

[5]We could have run this same analysis in other time windows if there was any variability in word surprisal in those windows, but there wasn't. Surprisal values were close to zero in the gender window (because participants almost always produced the expected gendered noun "girl" or "boy"), and zero in the noun window (because participants always produced the expected noun).

still underdeveloped. In this work we tested a particular hypothesis about why a frequently implicitly assumed linking hypothesis – the Referential Belief Link – appears to variably hold across and even within studies. We found that the extent to which the Referential Belief Link holds for the quantifier processing dataset of Sun and Breheny (2020) was modulated by participants' likely quantifier production expectations. This result dovetails with the result previously reported by Qing et al. (2018) and Degen et al. (2021) that correlations between implicit and explicit beliefs increased across subsequent time windows, as participants presumably became more certain about the intended target.

While we consider this work a promising first step towards understanding the applicability of linking hypotheses in the VWP, we hasten to list the caveats of this work. First, it is clear that a lot of variance in correlations remains unexplained by quantifier surprisal. This is related to a second issue: while we have shown that quantifier surprisal predicted the correlation between implicit and explicit beliefs, we have not yet provided a cognitive model of eye movements in the visual world. In this respect, linking hypotheses for eye-tracking during reading are somewhat better developed (e.g., Bicknell & Levy, 2010).[6] A step in this direction for the visual world might be to use surprisal as a parameter that toggles between explicit referential beliefs and random looking; or between referential beliefs and structured looking. One reason for structured looking is given by Sun and Breheny (2020)'s study: participants surprised by the use of "some" or "all" did not look randomly at objects in the display, but instead looked to the residue set for verification that the observed quantifier correctly applies to the target set under consideration.

Further, the current work is limited to referential tasks, i.e., tasks in which the listener's goal is to identify and select the speaker's intended referent. Other tasks, e.g., certain passive listening tasks which have been shown to elicit predictive eye movements (Altmann & Kamide, 1999), will require a different linking hypothesis.

This work raises interesting questions regarding the role of production expectations in experimental pragmatics comprehension studies. The fact that referential beliefs elicited in an offline selection task were more weakly correlated with the eye movement data precisely in those conditions where the observed quantifier was more surprising, provides support for accounts of pragmatic inference that ascribe delays in target identification not to inference computation cost per se (Bott & Noveck, 2004; Huang & Snedeker, 2009, 2018), but to additional sense-making processes that must be engaged when observing surprising language (Degen & Tanenhaus, 2015; Sun & Breheny, 2020). Future work should further distentangle the relative costs contributed by inference computation vs. additional sense-making processes.

Finally, we have shown here that there is good reason to believe that production expectations play an important role in visual world paradigm linking hypotheses. Future work should assess the generalizability of this finding by extending the investigation across a wider set of contexts and linguistic forms.

[6]Bicknell and Levy (2010) spell out a computational model of eye movement control during reading in which eye movement decisions are made to obtain (possibly noisy) visual information, which the reader uses in Bayesian inference about the form and structure of the sentence. It is an interesting question to what extent insights from the reading literature might be applied to the visual world.

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. *JML*, *38*(4), 419–439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1168–1178).

Bott, L., & Noveck, I. (2004, oct). Some utterances are underinformative: The onset and time course of scalar inferences. *JML*, *51*(3), 437–457.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Degen, J., Kursat, L., & Leigh, D. (2021). Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. In *Proceedings of the 43nd annual meeting of the cognitive science society* (Vol. 43).

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172–201.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society.*

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi: 10.1016/j.tics.2016.08.005

Gotzner, N., Wartenburger, I., & Spalek, K. (2016). The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, *8*(1), 59–95. doi: 10.1017/langcog.2015.25

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010, jul). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Grodner, D. J., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (Eds.), *The processing and acquisition of reference* (pp. 239–272).

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*, 376–415.

Huang, Y. T., & Snedeker, J. (2018, May). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, *102*, 105–126. doi: 10.1016/j.cogpsych.2018.01.004

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.

Kreiss, E., & Degen, J. (2020). Production expectations modulate contrastive inference. In *Proceedings of CogSci 42*.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In *Semantics and Linguistic Theory* (Vol. 26, pp. 836–854).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139.

Peloquin, B., & Frank, M. C. (2016). Determining the alternatives for scalar implicature. In *Proceedings of the annual meeting of the cognitive science society.*

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.02035

Qing, C., Lassiter, D., & Degen, J. (2018). What do eye movements in the visual world reflect? A case study from adjectives. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 89–110). Wiley.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285. doi: 10.1016/j.cognition.2020.104285

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sun, C., & Breheny, R. (2020). Another look at the online processing of scalar inferences: An investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience*, *35*(8), 949–979.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580.

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*, 1632 – 1634.