What do eye movements in the visual world reflect? A case study from adjectives

Ciyang Qing, Daniel Lassiter*, Judith Degen*
{qciyang, danlassiter, jdegen}@stanford.edu
Department of Linguistics, Stanford University
460 Serra Mall, Stanford, CA 94305, USA

Abstract

A common dependent measure used in visual-world eye-tracking experiments is the proportion of looks to a visually depicted object in a certain time window after the onset of the critical stimulus. When interpreting such data, a common assumption is that looks to the object reflect the listener's belief that the object is the intended target referent. While this is intuitively plausible (at least for paradigms in which the task requires selecting a referent), relatively little is known about how exactly the proportion of looks to an object is related to a listener's current belief about that object. Here, we test a simple, explicit linking hypothesis: the proportion of looks to an object correlates with the probability that the listener assigns to the object being the target. To test this hypothesis, we supplement the eye-tracking data from Leffel, Xiang, and Kennedy (2016) with an offline incremental decision task to measure participants' beliefs about the intended referent at various points in the unfolding sentence, and assess the extent to which these beliefs predict the eye-tracking data. The results suggest that the degree to which an object is believed to be the referent is only one factor that affects eye movements in referential tasks. Preliminary free production data we have collected for the scenes suggests that utterance expectations also play a role. We discuss methodological implications of these results for experimental linguistics.

Keywords: eye-tracking; visual world; linking functions; gradable adjectives; vagueness; imprecision; semantics; pragmatics

Introduction

Eye-tracking experiments using the *visual world paradigm* (VWP, Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) are widely used in linguistics (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Leffel et al., 2016). In standard VWP tasks, participants view displays of four objects while listening to spoken sentences while their eye movements are monitored (see Fig. 1). A commonly used dependent measure for evaluating whether experimental conditions – that reflect theoretically interesting conditions – differ from each other is the difference in *proportion of looks to a visually depicted object in a certain time window after the onset of a critical stimulus* across condition. When interpreting such data, a common assumption is that looks to the object reflect the listener's belief that the object is the intended target referent. While this is intuitively plausible (at least for paradigms in which the task requires selecting a referent, cf. Salverda & Tanenhaus, 2017), relatively little is known about how exactly the proportion of looks to an object is related to a listener's current belief about that object (but see Allopenna, Magnuson, & Tanenhaus, 1998). Understanding the relation between looks and beliefs is crucial for the theoretical interpretation of eye movement data for the purpose of linguistic theory-building. Here, we test a simple, explicit linking hypothesis: **the proportion of looks to an object correlates with the probability that the listener assigns to the object being the target**. To test this hypothesis, we supplement the eye-tracking data from Leffel et al. (2016) with an offline *incremental decision task* to measure participants' beliefs about the intended referent at various points in the unfolding sentence, and assess the extent to which these beliefs predict the eye-tracking data. The results suggest that the degree to which an object is believed to be the referent is only one factor that affects eye movements in referential tasks. Preliminary free production data we have collected for the scenes suggests that utterance expectations also play a role. We discuss methodological implications of these results for experimental linguistics.

Figure 1: Visual world paradigm used in Sedivy et al., 1999

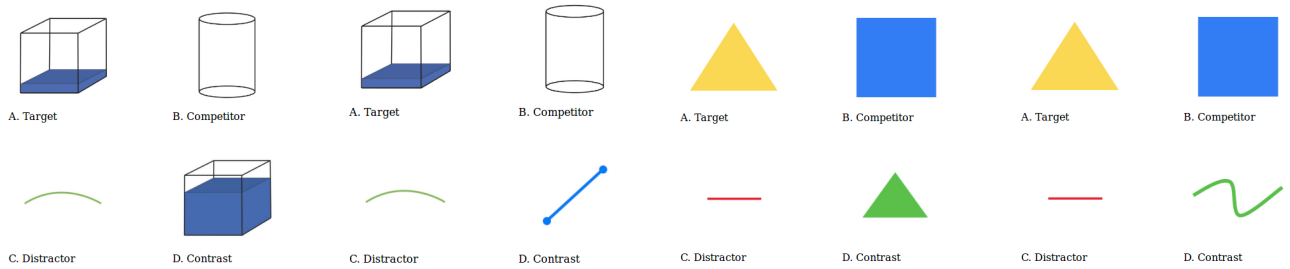
A case study: gradable adjectives

We tested the above linking hypothesis on an eye movement dataset that was collected with the intention of informing the debate over semantic theories of gradable adjectives such as *empty* and *big*. We recap the theoretical motivation for the experiment before focusing on testing the linking hypothesis.

According to degree-based approaches to the meaning of gradable adjectives (e.g., Kennedy, 2007), an object o satisfies a gradable adjective A iff o's degree of A-ness exceeds a standard of comparison θ. There are empirical differences between how *big* and *empty* are interpreted. *Relative* adjectives such as *big* and *tall* are *context-sensitive* and *vague*. In contrast, *maximum* adjectives such as *empty* and *straight* are not (or much less) *vague*: strictly speaking, a glass is empty iff it exhibits a maximum amount of emptiness (i.e., it is completely empty).[1]

The maximum/relative distinction is complicated by the fact that speakers often use these adjectives in an *imprecise* way. For example, it is often acceptable to call a glass empty when in fact there is still a little water in it.

[1]There is a third class of gradable adjectives such as *bent* and *dirty*, which only requires a minimum degree as the standard θ. Following Leffel et al. (2016), we do not consider such adjectives.



(a) Max (*empty*), contrast (b) Max (*empty*), no contrast (c) Rel (*big*), contrast (d) Rel (*big*), no contrast

Figure 2: Stimuli used in Leffel et al., 2016. Critical sentences are of the form "click on the [adj] [noun]"

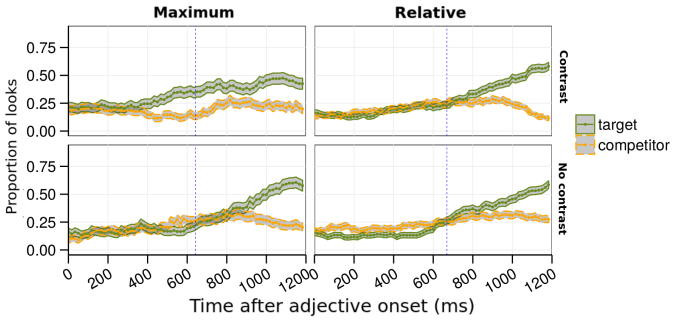


Figure 3: Proportions of looks on targets and competitors for different adjective types (columns) across contrast conditions (rows) from Leffel et al.'s (2016) visual world study. Blue lines indicate 200ms after average noun onset.

There is a consensus in the literature that the interpretation of relative adjectives involves resolution of the standard θ based on contextual information. However, theories differ in terms of how they analyze maximum adjectives and in particular their imprecise uses. According to recent probabilistic approaches, imprecise uses of maximum adjectives can be captured by a unified model of the contextual resolution of θ, and the differences between maximum and relative adjectives follow from different world knowledge about the various properties denoted by gradable adjectives (Lassiter & Goodman, 2013, 2015; Qing & Franke, 2014a, 2014b). Following Leffel et al. (2016), we call such approaches the *semantic hypothesis about imprecision* (**HS**). In contrast, Leffel et al. (2016) proposes a *pragmatic hypothesis* (**HP**), according to which maximum adjectives always use maximum degrees as θ and imprecise uses are due to an additional pragmatic mechanism that relaxes their strict literal meanings.

Leffel et al. (2016) attempted to use VWP to adjudicate between these two hypotheses. They conducted a variant of Sedivy et al.'s (1999) experiment (Fig. 1), in which participants saw displays of four objects and their task was to take actions according to auditory stimuli such as "touch the tall glass." Among the four objects, there was one that uniquely satisfied the full DP *the tall glass* (the *target* object). In addition, there was a *competitor* object that satisfied the adjective but not the noun (the tall pitcher). In half of the displays (the *contrast* condition) there was an object (the *contrast*) that satisfied the noun but not the adjective (e.g., the short glass). The rest of the objects were *distractors* that satisfied neither the adjective nor the noun (e.g., the comb and the key). The main finding, now a classic effect, is what has since been termed the *Referential Contrast Effect* (Sedivy, 2003): there was a difference in proportions of looks between the contrast and no-contrast conditions when only the adjective information was available, such that listeners looked more to the target in the presence of a contrast member, presumably as a result of pragmatic reasoning about the adjective only being necessary to distinguish two members of a contrast pair.

Building on Sedivy et al. (1999), Leffel et al. (2016)'s study leveraged the Referential Contrast Effect to test the processing of both maximum and relative gradable adjectives (Fig. 2). Crucially, the competitor object always exhibited a higher degree of the property denoted by the adjective than the target, and in the case of maximum adjectives the competitor satisfied the adjective perfectly (e.g., the perfectly empty cylinder in Fig. 2). Thus, in the case of maximum adjectives the target object was described by the adjective only in an imprecise way.

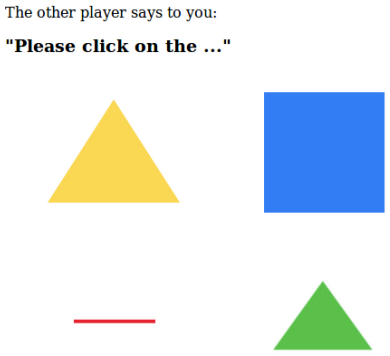


Figure 4: Visual stimulus in the prior window in the incremental decision task. In the adjective window, the adjective (e.g., *big*) was additionally displayed; in the noun window, the noun (e.g., *triangle*) was also displayed.

They observed a Referential Contrast Effect for maximum adjectives but not for relative adjectives (Fig. 3). They concluded that this favors **HP**, based on the following reasoning. (i) If **HS** were true, the resolution strategy of θ should be the same for both maximum and relative adjectives and hence the same processing pattern is expected for both types

of adjectives. (ii) If **HP** were true, the resolution strategies of θ should be different for maximum and relative adjectives and hence their processing patterns should differ as well. (iii) Given that RCE was observed only for maximum adjectives, the empirical finding is compatible with **HP** but not **HS**.

This kind of reasoning is commonplace in experimental semantics/pragmatics and we would like to probe some of the premises involved. In this paper, we focus on (i). For probabilistic approaches, the resolution strategy of θ is specified at the computational level (Marr, 1982). Given that having the same computational mechanism does not generally guarantee the same processing pattern (e.g., it takes longer to calculate the sum of two 30-digit numbers than 3-digit numbers, even though the underlying computational mechanisms can well be the same), (i) is not valid without additional assumptions. In fact, two types of additional assumptions are needed: (a) an assumption about the computational product of the resolution strategy, e.g., the contrast manipulation will affect probabilistic beliefs about the intended referent in the same way for both types of adjectives, and (b) a linking hypothesis: an assumption about the link between the computational product (probabilistic belief about the referent) and the processing pattern (proportion of looks), e.g., the proportion of looks to an object reflects the probability that the listener assigns to the object being the intended referent (in the same way for both types of adjectives).

Given the central role that listeners' beliefs about the intended referent play in the above assumptions, we directly measured these beliefs using a novel offline paradigm (which we refer to as the *incremental decision task*). We will focus on the linking hypothesis in (b) for two main reasons. First, existing probabilistic theories mentioned above are designed for *descriptive* uses of gradable adjectives (e.g., "John is tall") and do not directly make predictions about the *referential* uses in Leffel et al.'s experiment, Therefore it is unclear whether the assumption in (a) holds. But even if it does, the linking hypothesis in (b) still needs to hold for (i) to be the case. More importantly, the linking hypothesis in (b) is widely assumed in the psycholinguistics literature and is independent of the particular theoretical debate about gradable adjectives. Testing it therefore is relevant to any area of experimental linguistics that uses visual world eye-tracking.

Experiment 1: Incremental decision task

To directly measure listeners' beliefs about the intended referent at various points in the unfolding sentence and compare them with Leffel et al.'s eye movement data, we conducted an offline incremental decision task similar to the gating task used by Allopenna et al. (1998).

Methods

Participants We recruited 100 self-identified native English speakers via Amazon Mechanical Turk.

Materials We used the same visual stimuli as Leffel et al., 2016 (examples in Fig. 2). There were 60 critical visual displays, 20 of which were constructed out of 5 maximum adjectives (*empty*, *full*, *straight*, *flat*, and *closed*) and 40 of which were constructed out of 5 pairs of relative adjectives (*tall/short*, *long/short*, *big/small*, *wide/narrow*, *thick/thin*). Half of the displays occurred in the contrast condition and the other half in the no-contrast condition.

Procedure Participants were told that they were playing a game with another Turker, who sent a message to instruct them to click on one of the objects. In addition, they were told that due to a slow internet connection, they would sometimes need to make a choice even before their partner's entire message came through. The critical sentence "Please click on the [adj] [noun]" was revealed incrementally and participants clicked on the presumed intended referent after (a) the article "the" (Fig. 4), (b) the adjective, and (c) the head noun. After each click the next word or the next display was shown. After one practice example, each participant saw 120 displays in a random order, 30 of which were critical displays (10 maximum and 20 relative).

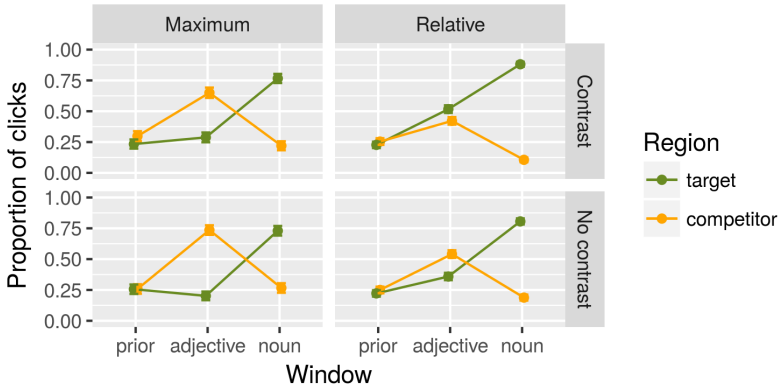


Figure 5: Proportions of clicks on the targets and competitors in the incremental decision task in different windows (x-axis), for different adjective types (columns) across contrast conditions (rows). Error bars indicate 95% CIs.

Results

Clicks Proportions of clicks on targets and competitors for different adjective types and conditions are shown in Fig. 5. In the prior window (i.e., right after the definite article *the*), participants' proportions of clicks on the targets and competitors were around .25. In the adjective window, for maximum adjectives the majority of the clicks were on the competitor, which perfectly satisfies the adjective, and fewer clicks (around .25) were on the target, which only loosely satisfies the adjective. In contrast, for relative adjectives, in the no-contrast condition about half of the clicks were on the competitor and fewer were on the target, which exhibits a lower degree of the property denoted by the adjective than the competitor, whereas in the no-contrast condition the reverse was the case: about half of the clicks were on the target and fewer were on the competitor. Finally, in the noun window, the vast majority of the clicks were on the target.[2]

[2]Not all of the clicks were on target, apparently because participants found that some of Leffel et al.'s stimuli did not unambigu-

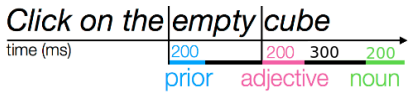


Figure 6: The 3 windows used in the eye-tracking data

Clicks vs looks To test the linking assumption that the proportion of looks to an object reflects the probability that the listener assigns to the object being the intended referent, we reanalyzed the eye-tracking data from Leffel et al., 2016 in 3 time windows: prior, adjective, and noun: the prior window is the first 200ms after the onset of the adjective, during which the information of the adjective has not yet been reflected in eye movement due to planning; the adjective window is the first 200ms after the onset of the noun; the noun window is 500–700ms after the onset of the noun (Fig. 6). The windows were chosen so that the click data and the eye-tracking data are maximally comparable: they were late enough so that the previous information had been processed as much as possible yet without the influence of the new information, making it close to the offline task that has no time limit.

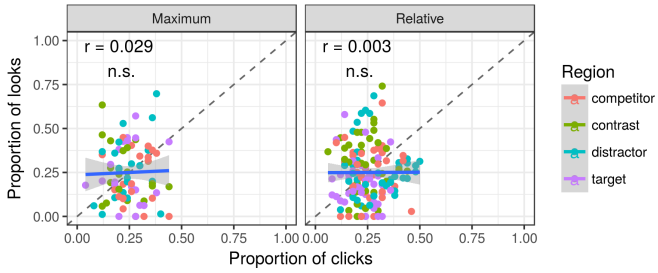


Figure 7: Correlations between click and eye movement data in the prior window for different adjective types (columns)

In the prior window (Fig. 7), we observed no significant correlations between proportion of looks and proportion of clicks on an object ($r < .03, p > .1$ for both adjective types).

In the adjective window (Fig. 8), in the contrast condition, we observed no significant correlation between proportions of clicks and looks for maximum adjectives ($r = .055, p > .1$) but a medium correlation for relative adjectives ($r = .462, p < .01$). In the no-contrast condition, we observed a weak correlation for both maximum ($r = .281, p < .01$) and relative ($r = .256, p < .01$) adjectives.

In the noun window (Fig. 9), we observed a strong correlation between proportions of clicks and looks for both maximum and relative adjectives ($r > .8, p < .001$).

Discussion

The results suggest that the degree to which an object is believed to be the referent correlates with the proportion of looks on that object to various extents depending both on the window and the adjective type.

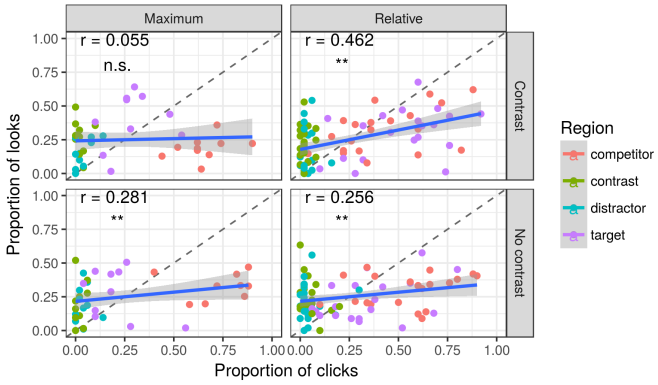


Figure 8: Correlations between click and eye movement data in the adjective window for different adjective types (columns) and contrast conditions (rows)

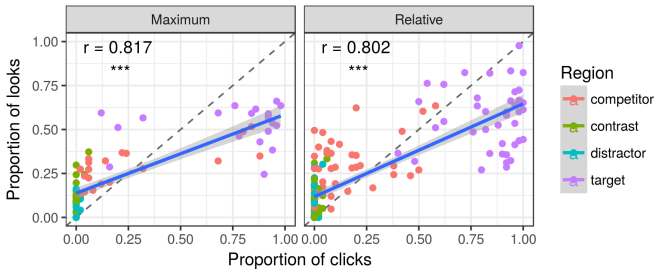
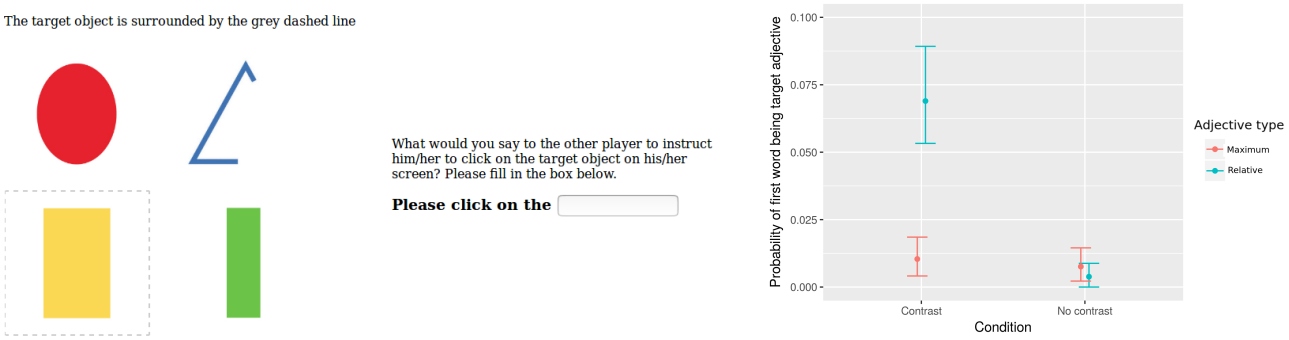


Figure 9: Correlations between click and eye movement data in the noun window for different adjective types (columns)

As the sentence unfolds, the correlation between clicks and looks generally increases. This is likely due to a tradeoff between *exploration* and *exploitation*: In earlier windows, participants were less familiar with the objects. Thus, they were likely mainly exploring the scene, resulting in eye movements that were not signal-driven, and consequently their proportions of looks did not necessarily correlate with their belief about the intended referent. In contrast, in later windows participants were more familiar with the objects and might have had more resources available for exploiting their signal-driven beliefs.

We also observed that in the adjective window, adjective type affects the correlation between clicks and looks and the correlations are relatively low compared with Allopenna et al.'s results. Note that in Allopenna et al.'s studies, participants were trained to name all the items so they had no uncertainty about how the target would be called, whereas the adjectives in Leffel et al.'s experiments were probably less expected because modification was not necessary in the no-contrast conditions and presumably color adjectives were more likely modifiers. Therefore we hypothesized that the differences in correlations were due to different expectations of hearing the adjectives. If participants hear a less expected adjective, they will need to explore the scene more, e.g., to

ously pick out the intended referent. For example, some participants clicked on a perfectly straight arrow when they saw it alongside a slightly bent line with the request "please click on the straight line". Several participants commented that their partner seemed to be using "line" to descibe arrows, and similarly for certain other shapes.



(a) Sample stimuli

(b) Proportions of first word being the target adjective

Figure 10: The free production experiment and results (error bars indicate 95% CIs)

evaluate whether each object satisfies the adjective, and hence the correlation between proportions of looks and beliefs will be lower. If participants hear a more expected adjective, they can directly exploit the signal, and hence the correlation between proportions of looks and beliefs will be higher. Given that it is easy to shift the standard θ in light of the local comparison between the target and the contrast objects for relative adjectives but difficult to do so for maximum adjectives (Syrett, Kennedy, & Lidz, 2010), and given that there is no need to use an adjective in the no-contrast condition, we hypothesized that adjectives were most expected in the contrast condition for relative adjectives.

To test this hypothesis, we conducted a free production experiment on Amazon Mechanical Turk to measure the likelihood of the participants describing the target using the adjective in Leffel et al.'s original experiment and Exp 1.

Experiment 2: Free production task

Methods

Using the same stimuli as in Exp. 1, 100 self-identified native English speakers were told that they were playing a game with a partner and their task was to instruct their partner to click on the target object, which was surrounded by the grey dashed line that their partner could not see. To familiarize the participants with the task, they first played 8 listener trials similar to Exp 1 but only with full sentences, then practiced 1 speaker trial, and finally played 30 speaker trials where they completed the sentence "please click on the __" for either the target or the competitor in the critical trials in Exp 1. They were told not to mention colors or locations to make the game more challenging. (In a pilot study without this restriction, almost all the adjectives in the responses were color terms.)

Results

Fig. 10b shows the proportions of descriptions of the target object in which the first word was the adjective used in Leffel et al.'s original experiment and Exp 1 (referred to as the target adjective). The target adjective was used the most in the contrast condition for relative adjectives, and in the other three cases the target adjective was used less. Note that the target adjective was not likely to be used right after the definite article (probabilities $<.1$ in all four cases). Instead, participants often used comparative forms (e.g., *wider rectangle*), modifiers (e.g., *almost empty cube*), and sometimes different adjectives (e.g., *big* instead of *tall*/*wide*).

Discussion

The results provide some initial support for our hypothesis that expectations of the adjectives also play a role in the correlation between looks and clicks. The target adjective was more likely to be used right after the definite article *the* in the contrast condition for relative adjectives than in the other 3 cases, therefore it was most expected by the listener and indeed the correlation between looks and clicks was the highest for relative adjectives in the contrast condition. However, note that in the other 3 cases since the target adjective was very unlikely to be mentioned (probabilities $< .02$), we do not have enough evidence to determine whether expectations of adjectives can account for their different correlations.

General discussion

Our results suggest that at least in this dataset, the linking hypothesis stated previously is only partially supported: the degree to which an object is believed to be the referent is only one factor that affects eye movements in referential tasks in which participants' goal is to interact with the intended referent; utterance expectations also play a role in this referential task. This has methodological implications for how proportions of looks in visual-world eye-tracking experiments should be interpreted.

Experimental results that manipulation X induces more looks on the target are often characterized as the manipulation facilitating reference resolution, which in turn often implies that manipulation X induces a stronger belief that the target is the intended referent (e.g., Leffel et al., 2016; Mulders & Szendroi, 2016; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2015; Salverda & Tanenhaus, 2017, among many others). However, this interpretation is valid only if the correlation between beliefs and looks is constant across manipulation X. Our results show that this may not always be the case. Therefore, additional caution is needed to make sure that empirical measures such as proportions of looks actually

track the theoretical constructs that researchers are interested in. For example, without testing the linking hypothesis, one might look at Leffel et al.'s results in Fig 3 and conclude from more looks on the target in the adjective window in the contrast condition for maximum adjectives that participants preferred the imprecise interpretation of maximum adjectives in the presence of contrast.[3] This is at odds with the results from our offline incremental decision task that directly measured participants' beliefs (Fig. 5).

We note that our study is only a first stab at testing explicit linking hypotheses used in visual-world eye-tracking studies. Further research is required to assess to what extent these results are robust and generalizable across eye movement datasets in experimental linguistics, but we believe that the offline incremental decision task provides a promising way to start investigating such problems. We are focusing on referential tasks where participants are instructed to interact with the referent. Such tasks can be straightforwardly adapted to offline incremental decision tasks and the correlations between clicks and looks can be tested. It would also be interesting to extend and apply the offline incremental decision task to passive-listening tasks to test correlations between looks and listeners' beliefs about the current or upcoming referent. Finally, the offline incremental decision task might be useful to test other linking hypotheses.

Conclusion

In this paper, we supplemented the eye-tracking data from Leffel et al. (2016) with an offline *incremental decision task* to measure participants beliefs about the intended referent at various points in the unfolding sentence, and tested a simple, explicit linking hypothesis: the proportion of looks to an object correlates with the probability that the listener assigns to the object being the target. Our results suggest that the degree to which an object is believed to be the referent is only one factor that affects eye movements in referential tasks. Preliminary free production data we have collected for the scenes suggests that utterance expectations also play a role in determining the correlation between clicks and looks. When the adjective is most expected, we observed the highest correlations, i.e., beliefs were a better predictor of eye-movements. Based on these results, we argue that proportions of looks in visual-world eye-tracking experiments should be interpreted with more caution and suggest stating the linking hypothesis explicitly and test it using the incremental decision task.

Acknowledgments

We are very grateful to Tim Leffel, Ming Xiang, and Chris Kennedy for generously sharing their data and stimuli.

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 and Language*, *38*(4), 419–439. doi: https://doi.org/10.1006/jmla.1997.2558

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2015). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of SALT 23*.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

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

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Mulders, I., & Szendroi, K. (2016). Early Association of Prosodic Focus with alleen only ': Evidence from Eye Movements in the Visual-World Paradigm. *Frontiers in Psychology*, *7*(March), 1–19.

Qing, C., & Franke, M. (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In J. Grieser, T. Snider, S. D'Antonio, & M. Wiegand (Eds.), *Proceedings of SALT* (pp. 23–41). elanguage.net.

Qing, C., & Franke, M. (2014b). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of cogsci* (p. 1204-1209). Austin, TX: 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 Blackwell.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

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.

Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1-35.

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.

[3] Note that Leffel et al. (2016) did not make this claim.