The role of social information in cross-situational word learning

Kyle MacDonald (kyle.macdonald@stanford.edu)

Abstract

Social information and the statistical regularity of language are both rich sources of information for word learners. Past research shows that the presence of social cues facilitates language acquisition (Baldwin, 1995; Brooks & Meltzoff, 2008) and that, in the absence of social information, adults and children can track co-occurrence information to learn words (Blythe, Smith, & Smith, 2010; Smith & Yu, 2008). In the current study, I explore the interaction between these two information sources by including a social cue in a cross-situational word learning task. I predicted that adding eye gaze information would give learners more evidence about referential intent, allowing them to "explain away" spurious word-object mappings and reduce the need to track multiple referents for each word. Results suggest that learners were sensitive to the social cue, but did not change how they track multiple referents. Potential explanations for these results and future directions are discussed.

Keywords: word learning; eye gaze; statistical learning; language acquisition

Introduction

To learn a new word[1], the learner must solve several problems. She must simultaneously infer what is being talked about and the meaning of individual words in the speaker's utterance. This joint inference problem can be divided into two parts. The first involves making an inference about speaker's referential intent and the second making an inference about the links between words and objects. Together, these social-cognitive and mapping challenges make word learning a surprisingly difficult puzzle for the child to solve, and a great deal of research has tried to explain how children can learn words so rapidly.

To account for children's prodigious word learning skills, different theories of language acquisition emphasize different tools and information available to the learner. These proposals can be divided into two broad categories: *Social-Pragmatic* and *Associative* accounts. Social-Pragmatic theories characterize word learning as a result of children's social-cognitive skills and the highly structured learning moments created by adults (i.e. joint attentional frames). Both experimental and observational data show that children possess sophisticated intention-reading skills, which they use in the service of word learning (Baldwin, 1993) and that episodes of joint attention facilitate word learning and vocabulary growth (Brooks & Meltzoff, 2008).

In contrast, Associative accounts highlight the powerful statistical learning mechanisms that allow learner's to track the regularities of language, linking words and objects over time. In this account, word learning is best explained by domain-general pattern-finding abilities, attention, and memory. Several studies show that in the absence of social cues to word meaning, adults and children are able to rapidly learn words by tracking the co-occurrence of labels and objects across exposures (Smith & Yu, 2008; Vouloumanos, 2008). However, some researchers question the psychological plausibility of gradualist accounts, suggesting that children's rapid word learning is better described by a single hypothesis tracking mechanism (Trueswell, Medina, Hafri, & Gleitman, 2013; Medina, Snedeker, Trueswell, & Gleitman, 2011).

Recent experimental work provides evidence that both adults and children can track and recall multiple referents and that this tracking interacts with attention and memory (Yurovsky Frank, in prep). In Yurovsky and Frank's task, participants saw a set of novel objects and heard a novel word (e.g. Grink), and were asked to make a guess about the "correct"[2] word-object mapping. In subsequent test trials, participants heard the novel word again, this time paired with another set of novel objects. Critically, one of the objects in the set was either the participant's initial hypothesis (Same trials) or one of the objects that was *not* the initial hypothesis (Switch trials). On Switch trials, adults reliably selected the object that was not their initial hypothesis, even when there were eight objects in the initial exposure set, providing strong evidence that learners track multiple referents when learning new words.

However, this task did not include any of the rich social cues that typically accompany real world word learning. So it is still an open question as to how social information interacts with learners' demonstrated ability to track multiple referents. Perhaps the presence of additional evidence about a speaker's referential intent strengthens the learner's initial hypothesis, reducing the need to track alternative hypotheses. Or it could be that social information strengthens the initial hypothesis without reducing multiple referent tracking. The current study follows a recent line of research and modeling that attempts to integrate statistical and social learning(Johnson, Demuth, & Frank, 2012; Frank, Goodman, & Tenenbaum, 2009; Yu & Ballard, 2007), and asks if the presence of social information changes how learners track multiple referents when learning new words?

Methods

Participants

This experiment was posted to Amazon Mechanical Turk as a set of Human Intelligence Tasks (HITs) to be completed only by participants with US IP addresses that paid 30 cents each (for a detailed comparison of laboratory and Mechanical Turk

[1]Here I will focus on the task of mapping words to objects with the goal of learning concrete nouns and assume that the learner has already solved the problem of word segmentation.

[2]There was actually no "correct" answers on exposure trials; rather these trials gave participants evidence about potential word-object mappings and allowed them to form an initial hypothesis.

studies see Crump, McDonnell, & Gureckis, 2013). 30 HITs were posted for each condition (Social, Non-social) for total of 60 paid HITs. If a participant completed the experiment more than once, he or she was paid each time but only data from the first HITs completion was included in the final data set. In addition, data was excluded from the final sample if participants did not give correct answers for familiar trials (excluded 1 HIT).

Stimuli

Stimuli for the experiment consisted of black and white pictures of familiar and novel objects, a schematic drawing of a human interlocutor, and audio recordings of familiar and novel words. Pictures of 32 familiar objects spanning a range of categories (e.g. squirrel, truck, tomato, sweater) were drawn from the set constructed by Snodgrass and Vanderwart (1980). Pictures of distinct but difficult to name objects were drawn from the set of 140 first used in Kanwisher, Woods, Iacoboni, and Mazziotta (1997). For ease of viewing on participants' monitors, pixel values for all pictures were inverted so that they appeared as white outlines on black backgrounds (see Figure 1). Familiar words consisted of the labels for the familiar objects as produced by ATT Natural VoicesTM (voice: Crystal). Novel words were 1-3 syllable pseudowords obeying the rules of English phonotactics produced using the same speech synthesizer.

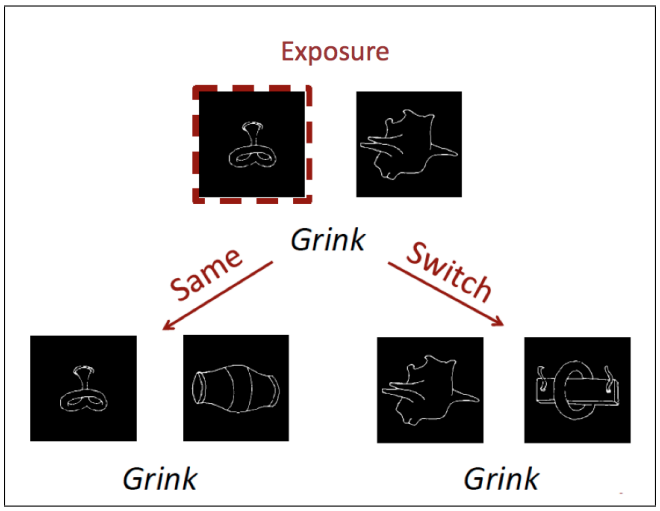

Figure 1: A schematic of the trials that participants saw in the experiment. On Exposure trials, participants saw four novel objects and heard a novel word. Participants were asked to guess it's correct referent. After the Exposure trial, participants either saw a Same or a Switch trial. On Same trials, the set of four referents contained the participant's previous hypothesis. On Switch trials, the set of referents contained one of the objects that the participants had *not* hypothesized.

A schematic drawing of a human interlocutor was chosen for ease of manipulating the direction of eye gaze, the social cue of interest in this study. Five images were created using Adobe Photoshop with the following directions of eye gaze: far left, close left, close right, far right, and eyes center. The interlocutor's eyes were made larger and green arrows were added to cue subjects' attention to the direction of eye gaze[3].

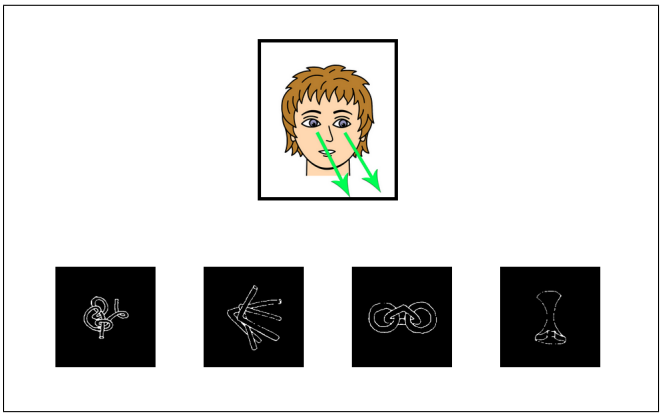

Figure 2: An example of an Exposure trial in the Social condition. In the Nonsocial condition, the interlocutor looked straight ahead during exposure. During test trials, the interlocutor looked straight ahead.

Design and Procedure

Participants were exposed to a series of trials in which they heard an interlocutor say a novel word, saw four novel objects, and were asked to indicate their guess as to which object was the referent of the word. After a written explanation of this procedure, participants were given four practice trials to introduce them to the task. On each of these trials, they heard the interlocutor say a Familiar word while looking at a line drawing of that object among a set of other familiar objects. On the first two trials, participants were asked to find the squirrel, and the correct answer was in the same position on each trial. On the next two trials, participants were asked to find the tomato, and the correct answer switched positions from the first to the second trial (in order to ensure that participants understood that the on-screen position was not an informative cue to the correct target). These trials also served to screen for participants who did not have their audio enabled or who were not attending to the task.

After these Familiarization trials, participants were informed that they would now hear novel words, and see novel objects, and that they should continue selecting the correct referent for each word. Participants each heard eight novel words twice. Participants saw four referents on each trial, and the two trials for each word occurred back-to-back. Four of these follow-up trials were Same trials in which the referent that participants selected on the exposure trial appeared again amongst the set of objects. The other four were Switch trials in which one of the referents in the set was selected randomly

[3]Because this task was performed over the internet and participants' screen sizes might be small, it was important to make the eye gaze cue clear.

from the objects a participant did not select on the previous Exposure trial. All other referents were completely novel on each trial. Critically, on Exposure trials the interlocutor's eye gaze was directed and thus informative, but on Same/Switch trials her eye gaze was undirected and thus uninformative.

Participants were randomly assigned to one of two conditions: Social or Nonsocial. In the Social condition, eye gaze was informative on exposure trials. In the Nonsocial condition, eye gaze was uninformative throughout the task, i.e. the speaker always looked straight ahead.

Because participants performed this task over the internet, it was important to indicate to them that their click had been registered. Thus, a red dashed box appeared around the object they selected for 1 second after their click was received. This box appeared around the selected object whether or not it was the "correct" referent.

Results and Discussion

First, I compared the distribution of correct[4] responses to the distribution expected if participants were randomly selecting (defined by a Binomial distribution with four trials and a probability of success of $(\frac{1}{\#Referents})$. Figure 3 shows participants? accuracies in identifying the referent of each word in both conditions for both kinds of trials (Same and Switch). For both Same and Switch trials in both the Social and Nonsocial conditions, participants? responses differed from those expected by chance (smallest $\chi^2(4) = 12.07, p < .01$), replicating Yurovsky and Frank's finding that adults encode more than a single hypothesis in ambiguous word learning situations.

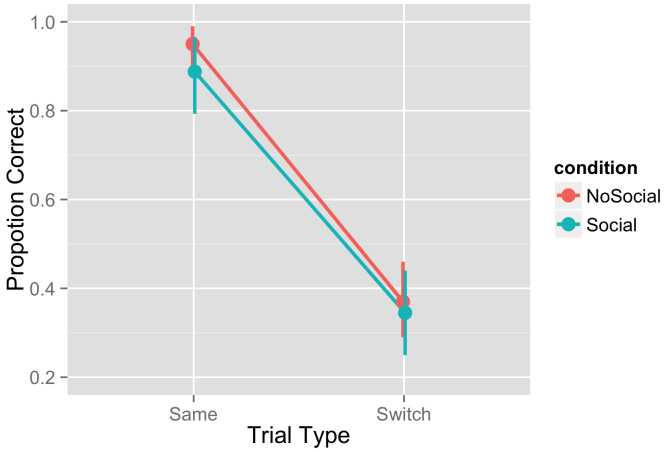

Figure 3: Proportion of repeated referents selected by participants for each trial type (Same/Switch) and condition (Social/Nonsocial). Error bars indicate 95% confidence intervals. Learning in all conditions differed from chance. There was no difference between the Social and Nonsocial conditions.

Next, I compared participants' performance on Same/Switch trials across conditions to see if the presence of a social cue during exposure changed participant's tracking of multiple referents. On both trail types, participants responses did not differ from each other: for Same trials, $t(42) = -1.19, p > .05$, for Switch trials, $t(42) = -0.38, p > .05$. Thus, I did not find evidence that the presence of social information during exposure trials changed how participants tracked and recalled multiple referents on subsequent test trials.

To check if participants were actually sensitive to the speaker's eye gaze, I compared the distribution of correct responses [5] on Exposure trials to the distribution expected if participants were selecting randomly. Participants' responses differed from those expected by chance, $t(239) = 16.89 p < .001$. Thus, there was evidence that participants were sensitive to the social cue and chose to use it, selecting the target of the interlocutor's eye gaze.

Finally, I explored participants' reaction times, comparing across conditions to see if the presence of social information altered the rate or participants' responses. Figure 4 shows participants' reaction time by trial and condition. Participants' responses did not differ on Same trials, $t(51) = 1.65, p > .05$. However, participants in the Social condition were slower on Switch trials, $t(51) = 2.63, p < .05$. This reaction time difference provides additional evidence that the presence of eye gaze had some effect on participants' behavior.

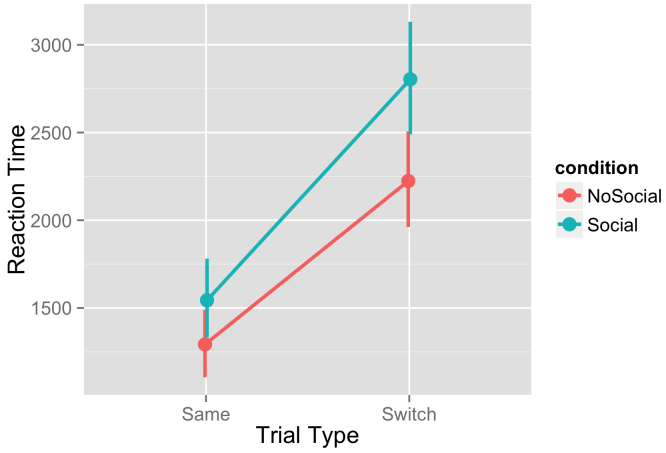

Figure 4: Reaction time for each trial type (Same/Switch) and condition (Social/Nonsocial). Error bars indicate 95% confidence intervals. There was no difference between the Social and Nonsocial conditions on Same trials, but participants in the Social condition were slower to respond on Switch trials.

Together, these results suggest that participants were sensitive to the interlocutor's eye gaze, selecting the target of her gaze on Exposure trials and responding more slowly on

[4] A correct response for Same/Switch trials was selecting the referent that was present during Exposure trials.

[5] A correct response on Exposure trials is selecting the referent that was the target of the speaker's eye gaze.

Switch trials in the Social condition. However, this sensitivity did not did not change how participants performed on accuracy in either Same/Switch trials. Potential explanations for these results and future directions are discussed below.

Conclusion and future work

In this study, I investigated whether the presence of social information would change how adults track multiple word meanings. I predicted that eye gaze would provide additional evidence about the speaker's referential intent, strengthening the learner's initial hypothesis, potentially at a cost to tracking alternative hypotheses. The results did not provide evidence in support of this prediction. Adults tracked multiple hypotheses regardless of the presence of a social cue, but were slower to respond on Switch trials when social cues were present.

The reaction time finding is somewhat interesting when you consider the structure of this cross-situational learning task. During Exposure trials, participants follow the speaker's gaze, often selecting the target as the "correct" referent. Then, participants are immediately confronted with a Switch trial in which the target of the interlocutor's gaze is no longer present. Perhaps participants slow reaction times are evidence of surprise that their initial hypothesis, which was weighted more strongly because of the presence of eye gaze, was no longer present.

So why might participants show sensitivity to social information but still show evidence of tracking multiple hypotheses? The analysis of participants' initial hypotheses rules out the alternative explanation that participants did not pay attention to the social cue during exposure. However, it is possible that the combination of a static/schematic interlocutor and the lack of a natural sentence frame (e.g. "There's a X") caused the task to be pragmatically strange, which might have led participants to discount the interlocutor as a source of evidence about word-object mappings.

Another possibility is that limiting the social information to eye gaze resulted in a weak social cue that didn't cause participants to change how they tracked the other referents. Recent research suggests that eye gaze might be a noisy and unreliable cue to reference (Frank, Tenenbaum, & Fernald, 2013). Future versions of the task could include stronger cues such as pointing or holding the object during Exposures trials to see if stronger evidence would alter learning. Future versions could also include more complex Exposure trials and more trials between exposure and test (similar to Yurovsky Frank, in prep) to see how social information interacts with attention and memory demands.

Language is a powerful tool that once mastered allows us to share our beliefs and learn from others. But learning a language is a surprisingly difficult challenge at several different levels. Thus, it is likely that language learners use several tools to take advantage of the sources of information available to them, both social and statistical. This study follows recent work that attempts to integrate social and statical learning with the hope that future studies will increase both our understanding of the unique contribution of each and how they interact to support learning.

References

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of child language*, *20*(02), 395–418.

Baldwin, D. A. (1995). Understanding the link between joint attention and language.

Blythe, R. A., Smith, K., & Smith, A. D. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(01), 207–220.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Johnson, M., Demuth, K., & Frank, M. (2012). Exploiting social information in grounded language learning via grammatical reductions. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 883–891).

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, *9*(1), 133–142.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126–156.

Vouloumanos, A. (2008). Fine-grained sensitivity to statisti-

cal information in adult word learning. *Cognition*, *107*(2), 729–742.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.