Reference Resolution in the Wild: On-line Circumscription of Referential Domains in a Natural, Interactive, Problem-solving Task.

Sarah Brown-Schmidt (sschmidt@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Abstract

We examined how naïve conversational participants circumscribed referential domains during the production and comprehension of referring expressions by monitoring participants' eye movements during a referential communication task. The results replicated some well-established results, e.g., incremental reference resolution, demonstrating the feasibility of studying real-time language comprehension in interactive conversation. We also observed a high proportion of underspecified referential expressions that were easily understood by addresses because of discourse and pragmatic constraints, including constraints developed as a result of the conversation.

Background

In characterizing work in language performance, Clark (1992) pointed out that the field has been largely divided into two traditions. One tradition, the language-as-action tradition, emphasizes interactive conversation as the most basic form of language use. According this tradition the principles of language performance and language design cannot be understood without taking into account the interactive collaborative processes that are embedded in conversation. A central tenet in work within this tradition is that utterances can only be understood within a particular context, which includes the time, place and participant's conversation goals. Thus researchers within this tradition have focused primarily in investigations of interactive conversation using natural tasks, typically with real-world referents.

A second tradition, the language-as-product tradition, focuses primarily on the processes by which listeners decode (and speakers encode) linguistic utterances. Psycholinguistic work on language comprehension within in this tradition typically examines moment-by-moment processes in real-time language processing using fine-grained reaction time measures. The rationale for using these measures is that comprehension processes are closely time-locked to the linguistic input which, for spoken language, unfolds over time. Until recently, the real-time response measures in the psycholinguist's toolkit required the use of de-contextualized language, typically pre-recorded sentences presented in impoverished contexts. This constraint ruled out real-time investigations of natural, interactive conversation. Moreover, a dominant theoretical perspective within the product tradition was that initial "core" processes (e.g., lexical access and syntactic processing) were informationally encapsulated from contextual influences (e.g., Fodor, 1983).

Recently, the advent of light-weight head-mounted eye-tracking systems has made it possible to investigate real-time comprehension in more natural tasks, such as tasks where participants follow spoken instructions to manipulate objects in a task-relevant "visual world" (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Fixations to task-relevant objects are closely time-locked to the unfolding utterance, providing a continuous real-time measure of comprehension processes at a temporal grain fine enough to track the earliest moments of lexical access, parsing and reference resolution (Tanenhaus, Magnuson, Dahan & Chambers, 2000).

A growing body of research employing eye-tracking techniques demonstrates clear effects of contextual constraints. For example, syntactic ambiguity resolution is influenced by referential constraints provided by the visual context, including the number of potential referents and their properties (Tanenhaus et al., 2000). Moreover, some recent work using confederates in constrained tasks suggests that under some circumstances information provided by knowledge of the speaker's perspective and intentions can affect even the earliest moments of comprehension (Hanna, 2001).

However, a major limitation of previous work is that all of the language used has come from scripted language, ruling out spontaneous collaborative processes that are likely to underlie circumscription of referential domains, and interpretation of referential expressions in natural interactive settings. For example, Clark & Wilkes-Gibbs (1986) investigated conversational partners' use of collaborative processes to refer to low-codability shapes in a referential communication task (Krauss and Weinheimer, 1966). Pairs of participants worked together to arrange different abstract shapes. Over the course of the conversation they converged on shared names for the shapes, dramatically increasing the efficiency of their communication over the course of the interaction. In Brennan's (1996) words, conversational partners develop 'conceptual pacts' during the course of a conversation. The mere action of participating in the development of conversational pacts is essential for the increase in efficiency- overhearers privy to the entirety of the conversation and it's context are unable to perform as well in these natural tasks (Schober and Clark, 1989). These results suggest two things: 1) The act of participating in a natural conversation contributes to efficient communication. 2) Extracting conversational interaction from language comprehension removes a central component of natural language.

The goal of the present study was to explore the feasibility of examining real-time comprehension processes during natural, unscripted, interactive conversation. We focused on the comprehension of definite referring expressions, such as "the red block" and "the cloud". Definite reference provides an ideal domain for a first investigation for several reasons. First, definite reference is one of the most ubiquitous and central components of natural language. Second, use of definite reference assumes that a referent will be uniquely identifiable within a circumscribed referential domain. Much of the strongest evidence for the collaborative model of language processing comes from demonstrations that people collaborate to define referential domains. Third, work with restricted utterances has established two clear empirical results that allow one to track the time course of reference resolution: lexical competitor (cohort) effects and 'point of disambiguation' effects.

When listeners are instructed to pick up or move an object, such as a racket, fixations to the target object begin as early as 200 ms after the onset of the noun (Allopenna, Magnuson & Tanenhaus, 1998). Eye-movements launched at this point in the speech stream are equally likely to be directed to the eventual referent and other objects with names that are also consistent with the speech signal, such as a raccoon. However, looks to these competitors, hereafter "cohort" competitors are reduced or eliminated when context makes a cohort an implausible referent. Thus we can use cohort effects to infer the degree to which conversation restricts initial referential domains.

One of the most striking sources of evidence for rapid restriction of referential domains comes from point of disambiguation effects. For example, Eberhard, Spivey-Knowlton, Sedivy and Tanenhaus (1995) presented subjects with displays containing a variety of differently colored shapes, as subjects listened to instructions such as “Click on the red triangle”. In a subset of trials the color of the target item was not shared with any other items in the referential domain. In these trials the referentially disambiguating information was the color, which was conveyed in the prenominal adjective. In the remaining trials, the target item was the same color as another item in the referential domain. For example, the display accompanying the instruction “Click on the red triangle” might contain a red circle and a red triangle. In these trials the referentially disambiguating information came at the noun. Eye movements to the target were again closely time-locked to the speech. Looks to the target increased dramatically immediately following the point of disambiguation (POD), whether it came at the adjective or the noun.

In the present experiment we monitored eye-movements as pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and confirm those configurations. The characteristics of the blocks afforded comparison with findings from scripted experiments investigating language-driven eye-movements, specifically those demonstrating cohort effects and incremental reference resolution. We investigated: (1) whether these effects could be observed in a more complex domain during unrestricted conversation, and (2) under what conditions the effects would be eliminated, indicating that factors outside the speech itself might be operating to circumscribe the referential domain.

Method

We tested four pairs of participants from the University of Rochester, who were paid for their participation in the study. The discourse partners each had an array of blocks and a board on which to place them. The boards were partially covered, creating 5 distinct sub-areas. Initially, sub-areas contained 56 stickers representing blocks. The task was to replace each sticker with a matching block. While partners' boards were identical with respect to sub-areas, partners' stickers differed: Every place that one partner had a sticker, the other partner had an empty spot. Pairs were instructed to tell each other where to put blocks so that in the end their boards would match. No other restrictions were placed on the interaction. The entire experimental study lasted approximately 2.5 hours. For each pair we recorded the eye movements of one partner, and the speech of both partners.

The initial configuration of the stickers was such that the color, size, and orientation of the blocks would encourage the use of complex noun phrases and grounding constructions. Nineteen of the stickers (and the corresponding blocks) contained pictures similar to those used by Allopenna, Magnuson and Tanenhaus (1998), in the study described above. We used a full-color version (Rossion & Pourtois, 2001) of a large corpus of normed pictures, balanced for their linguistic codablilty (Snodgrass & Vanderwart, 1980). We selected pairs of these pictures that referred to objects which had initially acoustically consistent names (cohort competitors). Half of the cohort competitor stickers were placed such that both cohort competitor blocks would be placed in the same sub-area of the board, and half of the cohort competitor stickers were placed such that the cohort competitor blocks would be placed in different sub-areas of the board. All of the cohort competitor pairs were separated by approximately 3.5 inches. We examined the eye movements of one discourse partner with respect to the speech generated by the other discourse partner.

Results

The conversations for each of the four pairs were transcribed. We present eye-tracking analyses for two of the pairs; we are still analyzing the data from the remaining two pairs. The non-eye-tracked partner of each pair generated approximately 100-150 definite references to blocks during the course of the conversation. While the length of the conversation prevented us from initially analyzing more than 4 pairs, the large number of 'trials' generated by each pair gave us enough statistical power to circumvent this problem.

An ISCAN eyetracking visor was used (for details see Trueswell, Sekerina, Hill & Logrip, 1999). The image of the eye-tracked partner's board, and their superimposed eye position, along with the entirety of the conversation (both participants' voices) were recorded using a frame-accurate digital video recorder (a SONY DSR-30). Eye movements were analyzed at the onset of the definite reference, and continued 2000ms after the NP was complete. There was a high degree of variability in the length of utterances, especially those to color blocks.

References to blocks which had cohort competitors (approximately 75 references per pair) were expected to reveal similar cohort effects as observed in more restricted experimental paradigms. To our surprise, we observed only one look to a cohort competitors during both 2 1/2 hour conversations we have analyzed thus far. We do not think this null result is due to poor stimulus design, as we did observe looks to cohort competitors under special circumstances. Periodically, subjects needed to remove the eye-tracker to take a break. When we put the tracker back on and re-calibrated, we tested the calibration by asking the subjects to look at different items on the board. Under these circumstances the referential domain is not restricted by conversational constraints. Here we saw clear cases of subjects initially looking at the cohort competitor before looking at the intended referent.

Each pair provided us with approximately 75 trials of data for eye-movements elicited by definite references to colored blocks. Two researchers coded the definite NPs for their point-of-disambiguation, and resolved any coding differences. The POD was the point at which the NP uniquely identified a referent, given the visual context at the time. Average POD was 858ms (26 frames) following NP onset. Eye-movement analyses for NPs with a unique linguistic point of disambiguation (50%) were analyzed separately from those which were never fully disambiguated linguistically. The eye-tracking analysis was restricted to cases where at least one competitor block was present. As a result, the number of ambiguous trials used for the analysis was close to 50, while there were only approximately 20 for the disambiguated trials.

Eye-movement analyses (planned comparisons) were performed on 800ms epochs for both ambiguous and disambiguated NPs. Eye movements elicited by disambiguated references showed clear point of disambiguation effects; within 200ms of the disambiguation point, looks converged on the target block: we found a main effect of condition F(2,20)= 64.03, p<.0001, and Bonferroni post-tests revealed significantly more looks to the 'target' than 'competitor' and 'other' unrelated blocks. (see Figure 1). Before the disambiguation point, subjects were looking equally at the target and competitor block(s), the main effect of condition, F(2,19)= 6.77, p<.01, was due to significantly more looks to target than other blocks (p<.001); looks to target and competitor were equivalent at this region. At the 522ms baseline region (-1122 to -600ms) there was no effect for condition. This replicates the point-of-disambiguation effect seen by Eberhard, et al. (1995), demonstrating that we were successful in using a more natural task to investigate on-line language processing. Upon inspection of the eye-movement plot, one should note that we do observe a pre-POD target bias. We will argue that this initial bias is due to additional pragmatic constraints that are operating during the task.

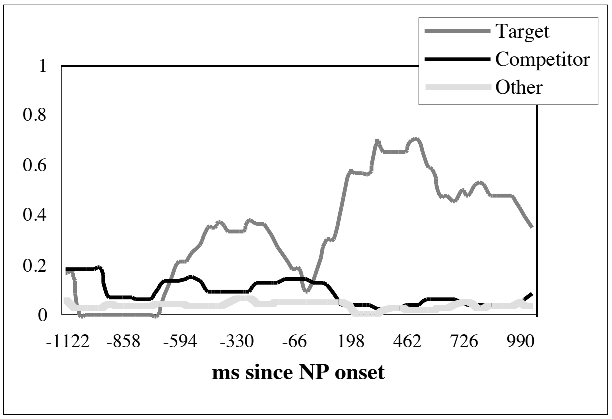

Figure 1: Proportion of fixations to Targets, Competitors, and Other blocks by time (ms) for Disambiguated NPs. Graph is centered by item with 0 ms = POD onset.

Most remarkably, ambiguous utterances elicited significantly more looks to the target than unambiguous utterances (see Figure 2). Moreover, fixations were primarily restricted to the referent shortly after onset of the definite reference; we observed significantly more looks to targets than competitors within the first 200 ms of NP onset, a significant effect of condition, F(2,53)= 18.37, p<.0001, was due to significant differences between target and both competitor and other blocks (p's<.0001). This effect persisted in the second 800ms window.

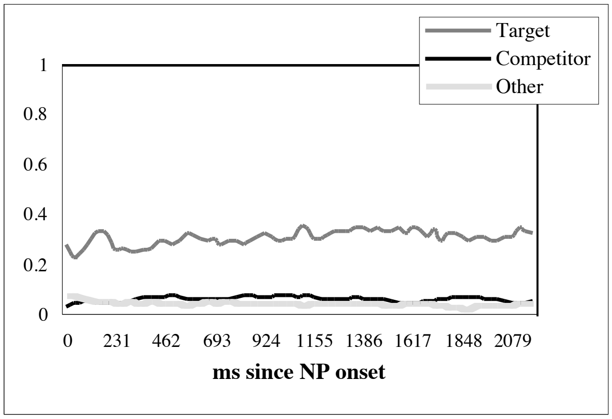

Figure 2. Proportion of fixations to Targets, Competitors, and Other blocks by time (ms) for Ambiguous NPs. Zero ms = NP onset.

These results suggest that 1) speakers systematically use less specific utterances when the referential domain has been otherwise constrained; 2) the attentional states of speakers and addresses become closely tuned; and 3) utterances are interpreted with respect to referential domains circumscribed by contextual constraints.

In order to identify the factors that led speakers to choose underspecified referring expressions, and that enabled addresses to understand them, we performed a detailed analysis of all of the definite references.

Recency is one factor that is likely to influence the form of a referring expression, with the most recently mentioned entity being more salient than other (non-focused) entities. (We are assuming that references to the most focused entity would typically result in use of a pronoun, a hypothesis that we are currently evaluating in the data set.) Thus, recency of last mention of the target block should predict the degree of specification, with references to the most recently mentioned block of a type, resulting in ambiguous referring expressions For example, if "the green block" is uttered in the context of a set of 10 blocks, 2 of which are green, 'recency' would predict that the referent should be to the green block that was most recently mentioned. Indeed, this is what we found: Ambiguous utterances were more likely to refer to the most recently mentioned block of its type, while this effect was the opposite for Disambiguated utterances. This effect was substantiated by a significant chi-square for independence, χ^2=7.389, p=.01.

While recency of mention is clearly an important factor constraining interpretation of ambiguous references, approximately 36% of these references did *not* refer to the most recently mentioned block. Additionally, the recency analysis alone does not explain why speakers sometimes chose to fully specify a reference when referring to the most recently mentioned block of its type; in fact 30% of disambiguated utterances referred to the most recently mentioned block of its type. Two questions arise from these observations: 1) Why are addresses not confused when a speaker uses an underspecified expression to refer to something other than the most recently mentioned thing; 2) What factors determine when speakers will add extra information when referring to the most recently mentioned block?

In answer to the first question, we propose that additional pragmatic and task-based factors function to constrain the referential domain, allowing speakers to underspecify, addresses to interpret ambiguous references, and may explain the early target advantage in the disambiguated trials. In addition to recency of mention, we identified three factors which contributed to the intelligibility of referring expressions: 1) Proximity of target referent to the last mentioned block; 2) Task-based constraints (such as limitations on block placement due to the size and shape of the board); 3) Spatial Terms which focus attention to a subset of the work area. We are currently performing a detailed analysis of the interaction of these constraints, and a comparison of how the predictions made by the constraints compare with the predictions of the addressee.

Finally, analysis of the eyetracking data of the final two pairs will allow us to do a sub-analysis of the eye-movements during trials influenced by these different factors. Our hypothesis is that both linguistic factors (such as recency of mention) in addition to pragmatic factors (such as proximity and task constraints) are contributing to the ease with which subjects are identifying the intended referents of these ambiguous references.

The answer to the second question, is the reverse of the answer to the first. We would like to suggest that in cases where confusion is high, conversation is inefficient, or these additional task and pragmatic constraints may select the *wrong* referent, speakers may choose to add additional disambiguating information. In order to verify this claim, we intend to compare the referential domain circumscribed by these additional constraints, with the intended referent of the speaker. We predict that a mismatch would lead to an increase in the likelihood of extra disambiguating information.

As a part of this analysis, we have also looked at the cases in which speakers overdisambiguate. In approximately 21% of the disambiguated utterances, speakers added between 1 and 3 additional elements past the POD. In the following example, the reference was disambiguated at 'long', yet the speaker chose to continue: "the long green square that was laying down". This speaker added two extra elements, a color term, and a collaboratively defined term, which means 'horizontal'. In many cases, the speaker spent a relatively large amount of time uttering extra disambiguating information, especially lengthy collaborative terms (Collaboratively-defined terms were common in this corpus; approximately 20% of ambiguous and 40% of disambiguated references contained collaboratively-defined terms). Of the overspecifying elements used, only 50% were color terms, which are the prototypical overdisambiguating element. The other 50% included references to previous actions, the location of the object and its shape. In general, speaker overspecification may be for clear communicative purposes, rather than as a bi-product of the production system.

Conclusions

In this experiment, we investigated 1) whether it is possible to observe incremental processing effects in a complex domain during unrestricted conversation, and 2) under what conditions these effects might be absent, indicating factors outside the speech itself might be operating to circumscribe the referential domain. We were successful on both counts. We did observe incremental reference resolution in certain contexts, and in the contexts in which it was not observed we were able to identify a number of constraints that seemed to be operating to circumscribe the referential domain. These constraints included linguistic recency, pragmatic factors related to the task itself, such as physical limitations on block placement due to the size and shape of the board, and proximity of a given referent to a previously mentioned block. An example of a segment of the discourse in which recency circumscribes the referential domain is shown below:

> 2. ok. RIGHT directly next to the cloud?
> 1. mm-hmm
> 2. just throw in <u>a red piece</u>, line it up evenly
> 1. just a red, little *square*
> 2. yup
> 1. k, got it
> 2. ok
> 1. now, I got an easy one, so I wanna *give it* to you
> 2. *ok*
> 1. directly…ABOVE <u>the red</u>, grab your lamp

The first description of the target block (underlined) is unambiguous, but the second reference to the block is ambiguous given the visual context (there are 5 other red blocks in that sub-area of the board). Listeners do not have difficulty with the linguistic ambiguity in this situation because they take recency into account, unifying the referents of the two referring expressions. An example of a segment of the discourse in which task constraints circumscribe the referential domain is shown in below:

> 1. ok, you're gonna line it up… it's gonna go <pause> one row ABOVE the green one, directly next to it.
> 2. can't fit it
> 1. cardboard?
> 2. can't yup, cardboard
> 1. well, tell it too back
> 2. the only way I can do it is if I move, alright, should the green piece with the clown be directly lined up with <u>thuuh square</u>?

Again, the referring expression (underlined) is ambiguous given the visual context. In this case the listener does not have difficulty dealing with this ambiguity because, although there are a number of blocks one could line up with "the green piece with the clown", only one is task-relevant. Given the location of all the blocks in the relevant sub-area, the target block is the easiest block to line up with the clown. The competitor blocks are inaccessible because of the position of the other blocks or the design of the board. An example of a segment of the discourse in which proximity constrains the referential domain is the following:

> 2. ok, so it's four down, you're gonna go over four, and then you're gonna put the piece right there
> 1. ok…how many spaces do you have between this green piece and <u>the one to the left of it</u>, vertically up?

As before, the referring expression (underlined) is ambiguous given the visual context; there are approximately five blocks up and to the left of the previously focused block (the one referred to in the NP as "this green piece"). In this case the listener does not have difficulty dealing with the ambiguity because he considers only the block closest to the last mentioned block ("this green piece"). Finally, an example of a reference that is constrained by a spatial term is underlined in the following exchange:

> 2: ok, and then…alright, so then there is a dark green one? to thEE uh northeast of that green one?
> 1: yup
> 2: and, um, they're only overlapping...one...line and then there's a yellow one…**below** the dark green one that I just talked about and to the l- **to the RIGHT** of <u>the other dark green one</u>.

The underlined reference is constrained before the onset of the noun phrase by the spatial terms used before it (bolded). When the listener hears the target noun phrase, she is already aware that the referent is **below** 'the dark green one', and that there is a space to the left of it (as she is directed to place a yellow block **to the right** of the referent). This information narrows the interpretation of the reference down to a single block, whereas the reference was otherwise ambiguous with respect to that sub-area in general.

We are currently detailing the predictions of and the interactions between these different constraints and the degree to which they predict both speaker behavior, and the interpretation processes of the addressee. Our observations and analysis of incremental interpretation during this task suggest a view of language processing in which conversational participants coordinate a mutually aware reliance on certain discourse, pragmatic and task based constraints which facilitate efficient completion of the task at hand. Our data mark an important first step towards being able to rigorously analyze the on-line processing of interactive conversation 'in the wild'. To our knowledge, this is the first demonstration of on-line circumscription of referential domains in a natural interactive task with naïve participants. As we continue to develop more explicit models of on-line language processing, a critical part of this process should be to inform these models with observations made in these natural situations. Pairing methodologically rigorous laboratory studies with naturalistic studies such as this one is essential to an understanding of language processes that is both detailed, and ecologically valid.

To conclude, we successfully replicated a standard psycholinguistic effect, the point of disambiguation effect, in unscripted interactive conversation with naïve participants. We also obtained results suggesting that reference selection and comprehension is modulated by both discourse-based factors, such as recency, but also by task specific pragmatic constraints.

Moreover, reference resolution appeared to be affected by collaborative constraints that developed during the conversation. Our participants spontaneously created collaborative terms for troublesome words (such as ‘horizontal’ and ‘vertical’), and tuned their utterances, and comprehension systems for such details as the recency of mention of each particular kind of block, proximity of blocks to one another, and task constraints idiosyncratic to our block-game. These observations suggest that the attentional states of the speaker and listener become closely tuned during the course of interaction. An important question for future research is how these factors differentially affect speakers and addressees. The data that we have collected is rich enough to allow us to investigate this and other questions.

Acknowledgments

This research was partially supported by NIH grant HD 27206. Thank you to Jesseca Aqui, Annie Tanenhaus, and Sanjukta Sanyal for all their help.

References

Allopenna, P. D, Magnuson, J.S. & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439**.**

Brennan, S. and Clark, H. (1996) Conceptual Pacts and Lexical Choice in Conversation*. Journal of Experimental Psychology: LMC,* 22, 482-493.

Clark, H. (1992) Arenas of Language Use. Chicago: University of Chicago.

Clark, H. & Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C. & Tanenhaus, M.K. (l995). Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409-436.

Fodor (1983). Modularity of Mind. Cambridge, MS; Bradford Books.

Hanna (2001). The effects of linguistic form, common ground, and perspective on domains of referential interpretation. Unpublished doctoral dissertation. The University of Rochester.

Krauss, R. M. and Weinheimer, S. (1966) Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4, 343-346.

Rossion, B. & Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's Object database: Color and Texture improve Object Recognition. 1st Vision Conference, Sarasota, FL.

Snodgrass, J.G., & Vanderwart, M. (1980). *JEP:Human Learning and Memory*, 6:3, 174-215.

Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21, 211-232.

Trueswell, J.C., Sekerina, I., Hill, N. & Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89-134.

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

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. & Sedivy, J.E. (l995). Integration of visual and linguistic information in spoken language comprehension*. Science, 268,* 632-634.