PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Language processing in the natural world

Michael K Tanenhaus and Sarah Brown-Schmidt

Phil. Trans. R. Soc. B 2008 **363**, doi: 10.1098/rstb.2007.2162, published 12 March 2008

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 1105–1122
doi:10.1098/rstb.2007.2162
Published online 25 September 2007

Language processing in the natural world

Michael K. Tanenhaus[1,*] and Sarah Brown-Schmidt[2]

[1]*Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA*
[2]*Beckman Institute, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA*

The authors argue that a more complete understanding of how people produce and comprehend language will require investigating real-time spoken-language processing in natural tasks, including those that require goal-oriented unscripted conversation. One promising methodology for such studies is monitoring eye movements as speakers and listeners perform natural tasks. Three lines of research that adopt this approach are reviewed: (i) spoken word recognition in continuous speech, (ii) reference resolution in real-world contexts, and (iii) real-time language processing in interactive conversation. In each domain, results emerge that provide insights which would otherwise be difficult to obtain. These results extend and, in some cases, challenge standard assumptions about language processing.

Keywords: eye movements; language comprehension; spoken word recognition; conversation; parsing; speech perception

1. INTRODUCTION

As people perform everyday tasks involving vision, such as reading a newspaper, looking for the car keys or making a cup of tea, they frequently shift their attention to task-relevant regions of the visual world. These shifts of attention are accompanied by shifts in gaze, accomplished by ballistic eye movements known as *saccades,* which bring the attended region into the central area of the fovea, where visual acuity is greatest. The pattern and timing of saccades, and the resulting fixations, are one of the most widely used response measures in the brain and cognitive sciences, providing important insights into the functional and neural mechanisms underlying reading, attention, perception, memory (for reviews see Rayner (1998) and Liversedge & Findlay (2000)) and, most recently, visual behaviour in natural tasks (Land 2004; Hayhoe & Ballard 2005).

During the last decade, we have been using saccadic eye movements to investigate spoken-language processing in relatively natural tasks that combine perception and action (Tanenhaus *et al*. 1995; see Cooper (1974) for an important precursor). In these studies, participants' fixations are monitored, typically using a lightweight head-mounted eye tracker, as they follow instructions to manipulate objects or participate in a dialogue about a task-relevant workspace—the 'visual world'. These methods have made it possible to monitor real-time language comprehension at a grain, fine enough to reveal subtle effects of sub-phonetic processing while using tasks as natural as unscripted interactive conversation. Before describing this work in more detail, we briefly sketch the motivation for studying language processing in the natural world.

Until recently, most psycholinguistic research on spoken-language comprehension could be divided into one of two traditions, each with its roots in seminal work from the 1960s (Clark 1992, 1996), and with its own characteristic theoretical concerns and dominant methodologies. The *language-as-product* tradition has its roots in George Miller's synthesis of the then-emerging information processing paradigm and Chomsky's theory of transformational grammar (Miller 1962; Miller & Chomsky 1963). The product tradition emphasizes the individual cognitive processes by which listeners recover linguistic representations—the 'products' of language comprehension. Psycholinguistic research within the product tradition typically examines moment-by-moment processes in real-time language processing, using fine-grained reaction time measures and carefully controlled stimuli.

The motivation for these measures comes from two observations. First, speech unfolds as a sequence of rapidly changing acoustic events. Second, experimental studies show that listeners make provisional commitments at multiple levels of representations as the input arrives (Marslen-Wilson 1973, 1975). Evaluating models of how linguistic representations are accessed, constructed and integrated given a continuously unfolding input requires data that can be obtained only by response measures that are closely time-locked to the input as it unfolds over time. We can illustrate this point by noting that one consequence of the combination of sequential input and time-locked processing is that the processing system is continuously faced with temporary ambiguity. For example, the initial portion of the spoken word *beaker* is temporarily consistent with many potential lexical candidates, e.g. *beaker*, *beet*, *beep*, *beetle* and *beagle*. Similarly, as the utterance, *Put the apple on the towel into the box* unfolds, the phrase, *on the towel*, is temporarily consistent with at least two mutually incompatible possibilities; *on the towel* could introduce a goal argument for the verb *put* (the location where the apple is to be put) or it could modify the theme argument, *the apple*, specifying the location of the theme (on the towel). A similar

* Author for correspondence (mtan@bcs.rochester.edu).

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

Table 1. Excerpt of dialogue taken from Brown-Schmidt *et al.* (2005).

speaker	utterance
1	*ok, ok I got it* ele…ok
2	alright, *hold on*, I got another easy piece
1	*I got a* well wait I got a green piece right above that
2	above this piece?
1	well not exactly right above it
2	it can't be above it
1	it is to the…it' doesn't wanna fit in with the cardboard
2	it is to the right, right?
1	yup
2	w- how? *where*
1	*it is* kinda line up with the two holes
2	line 'em right next to each other?
1	yeah, vertically
2	vertically, meaning?
1	up and down
2	up and down

argument for the importance of time-locked response measures holds for studies of language production where the speaker must rapidly map thoughts onto sequentially produced linguistic forms (Levelt *et al.* 1999). The *language-as-action* tradition has its roots in work by the Oxford philosophers of language use, e.g. Grice (1957), Austin (1962) and Searle (1969), and work on conversational analysis, e.g. Schegloff & Sacks (1973). The action tradition focuses on how people use language to perform acts in conversation, the most basic form of language use. Psycholinguistic research within the action tradition typically focuses on interactive conversation involving two or more participants engaged in a cooperative task, typically with real-world referents and well-defined behavioural goals. One reason for the focus on these types of tasks and situations is that many aspects of utterances in a conversation can be understood only with respect to the context of the language use, which includes the time, place and participant's conversational goals, as well as the collaborative processes that are intrinsic to conversation. Moreover, many characteristic features of conversation emerge only when interlocutors have joint goals and when they participate in a dialogue both as a speaker and an addressee.

Table 1 illustrates some of these features using a fragment of dialogue from a study by Brown-Schmidt and colleagues (Brown-Schmidt *et al.* 2005; Brown-Schmidt & Tanenhaus in press). Pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and confirm those configurations. The excerpt contains many well-documented aspects of task-oriented dialogue, including fragments that can only be understood as combinations of utterances between two speakers, false starts, overlapping speech (marked by asterisks) and negotiated referential terms (e.g. *vertically* meaning up and down).

Detailed analyses of the participants' linguistic behaviour and actions in cooperative tasks have provided important insights into how interlocutors track information to achieve successful communication (Clark 1992, 1996). They also demonstrate that many aspects of communication, establishing successful reference, for instance, are not simply an individual cognitive process; they are arrived at as the result of coordinated actions among two or more individuals across multiple linguistic exchanges (Clark & Wilkes-Gibbs 1986). However, because they interfere with interactive conversation, researchers in the action tradition have for the most part eschewed the time-locked response measures favoured by researchers in the product tradition (but see Marslen-Wilson *et al.* 1982; Brennan 2005). For example, some of the widely used experimental paradigms for examining real-time spoken-language comprehension require: (i) asking a participant to make a metalinguistic judgement while monitoring the speech input for a linguistic unit such as a phoneme, syllable or word, (ii) measuring a response to a visual target that is presented on a screen during a sentence, or (iii) monitoring EEG activity while the participant's head remains relatively still. None of these procedures can be used without disrupting interactive conversation. Thus, little is known about the moment-by-moment processes that underlie interactive language use.

2. WHY STUDY REAL-TIME LANGUAGE PROCESSING IN NATURAL TASKS?

While it is tempting to view the product and action traditions as complementary, research in the product tradition examines the early perceptual and cognitive processes that build linguistic representations; research in the action tradition focuses on subsequent cognitive and social–cognitive processes that build upon these representations—we believe that this perspective is misguided. An increasing body of evidence in neuroscience demonstrates that even low-level perceptual processes are affected by task goals. Behavioural context, including attention and intention, affect basic perceptual processes in vision (Gandhi *et al.* 1998; Colby & Goldberg 1999). In addition, brain systems involved in perception and action are implicated in the earliest moments of language processing (Pulvermüller *et al.* 2001). Thus, studies that examine sub-processes in isolation, without regard to other subsystems, and broader behavioural context, are likely to be misleading. Moreover, it is becoming clear that at least some aspects of conceptual representations are grounded in perception and action. The language used in interactive conversation is also dramatically different than the carefully scripted language that is studied in the product tradition. The characteristics of natural language illustrated in the excerpt from Brown-Schmidt *et al.* (2005) are ubiquitous, yet they are rarely studied outside of the action tradition. On the one hand, they raise important challenges for models of real-time language processing within the product tradition, which are primarily crafted to handle fluent, fully grammatical well-formed language. On the other hand, formulating and evaluating explicit mechanistic models of how and why these conversational phenomena arise requires data that necessitate real-time methods.

Moreover, the theoretical constructs developed within each tradition sometimes offer competing explanations for phenomena that have been the

primary concern of the other tradition. For example, the product-based construct of *priming* provides an alternative mechanistic explanation for phenomena such as lexical and syntactic entrainment (the tendency for interlocutors to use the same words and/or the same syntactic structures). A priming account does not require appeal to the action-based claim that such processes reflect active construction of common ground between interlocutors (cf. Pickering & Garrod 2004). Likewise, the observation that speakers articulate lower frequency words more slowly and more carefully, which has been used to argue for speaker adaptation to the needs of the listener, has a plausible mechanistic explanation in terms of the greater attentional resources required to sequence and output lower frequency forms.

Conversely, the interactive nature of conversation may provide an explanation for why comprehension is so relentlessly continuous. Most work on comprehension within the product tradition takes as axiomatic the observation that language processing is continuous. If any explanation for *why* processing is incremental is offered, it is that incremental processing is necessitated by the demands of limited working memory, *viz*., the system would be overloaded if it buffered a sequence of words rather than interpreting them immediately. However, working memory explanations are not particularly compelling. In fact, the first-generation models of language comprehension—models that were explicitly motivated by considerations of working memory limitations—assumed that comprehension was a form of sophisticated catchup in which the input was buffered long enough to accumulate enough input to reduce ambiguity (for discussion, see Tanenhaus (2004)). There is, however, a clear need for incremental comprehension in interactive conversation. Interlocutors, who are simultaneously playing the roles of speaker and addressee, need to plan and modify utterances in midstream in response to input from an interlocutor.

Finally, the action and product traditions often have different perspectives on constructs that are viewed as central within each tradition. Consider, for example, the notion of *context*. Within the product tradition, context is typically viewed as information that either enhances or instantiates a context-independent core representation or as a *correlated constraint* in which information from higher-level representations can, in principle, inform linguistic processing when the input to lower levels of representation is ambiguous. Specific debates about the role of context include whether, when and how: (i) lexical context affects sub-lexical processing, (ii) syntactic and semantic context affect lexical processing, and (iii) discourse and conversational context affect syntactic processing. Each of these questions involves debates about the architecture of the processing system and the flow of information between different types of representations—classic information processing questions. In contrast, we have already noted that within the action tradition context includes the time, place and the participant's conversational goals, as well as the collaborative processes that are intrinsic to conversation. A central tenet is that utterances can only be understood relative to these factors. Although these notions can be conceptualized as a form of constraint, they are intrinsic to the comprehension process rather than a source of information that helps resolve ambiguity in the input. For these reasons, we believe that combining and integrating the product and action approaches is likely to prove fruitful by allowing researchers from each tradition to investigate phenomena that would otherwise prove intractable. Moreover, research that combines the two traditions is likely to deepen our understanding of language processing by opening up each tradition to empirical and theoretical challenges from the other tradition.

We now review three streams of research. The first two are by Tanenhaus and his collaborators. The third is a new line of research, which we have conducted jointly. First, we briefly discuss work examining how fine-grained acoustic information is used in spoken word recognition. We review this work to: (i) illustrate the sensitivity and temporal grain provided by eye movements, (ii) address some methodological concerns that arise from studying language processing in a circumscribed world, and (iii) highlight the importance of studying language processing as an integrated system. Second, we review studies demonstrating that real-world context, including intended actions, perceptually relevant properties of objects, and knowledge about the speaker's perspective combine to circumscribe the 'referential domain' within which a definite referring expression, such as *the empty bowl* is interpreted. Third, we present preliminary results from studies that begin to fully bridge the product and action traditions by examining real-time processing during unscripted conversation to explore how the participants in a task-oriented dialogue coordinate their referential domains.

3. SPOKEN WORD RECOGNITION IN CONTINUOUS SPEECH

Since the seminal work of Marslen-Wilson and colleagues, an important goal of models of spoken word recognition has been to characterize how a target word is identified against the backdrop of alternatives or 'neighbours' that are temporarily consistent with the unfolding input (Marslen-Wilson & Welsh 1978). Some models, such as the neighbourhood activation model emphasize global similarity, without taking into account whether the overlap between the targets and potential competitors occurs early or late (Luce & Pisoni 1998). Some emphasize onset-based similarity by incorporating explicit bottom-up mismatch inhibition to strongly inhibit lexical candidates that have any mismatch with the input (Marslen-Wilson & Warren 1994; Norris *et al*. 2002). And some, such as the TRACE model (McClelland & Elman 1986) adopt a middle ground by avoiding explicit mismatch inhibition, but incorporating lateral inhibition at the lexical level. Similarity at any point can activate any word. However, there is an advantage for candidates overlapping at onset: since they become activated early in processing, they inhibit candidates that are activated later. Distinguishing among these alternative hypotheses about lexical neighbourhoods requires mapping out the time course of lexical processing.

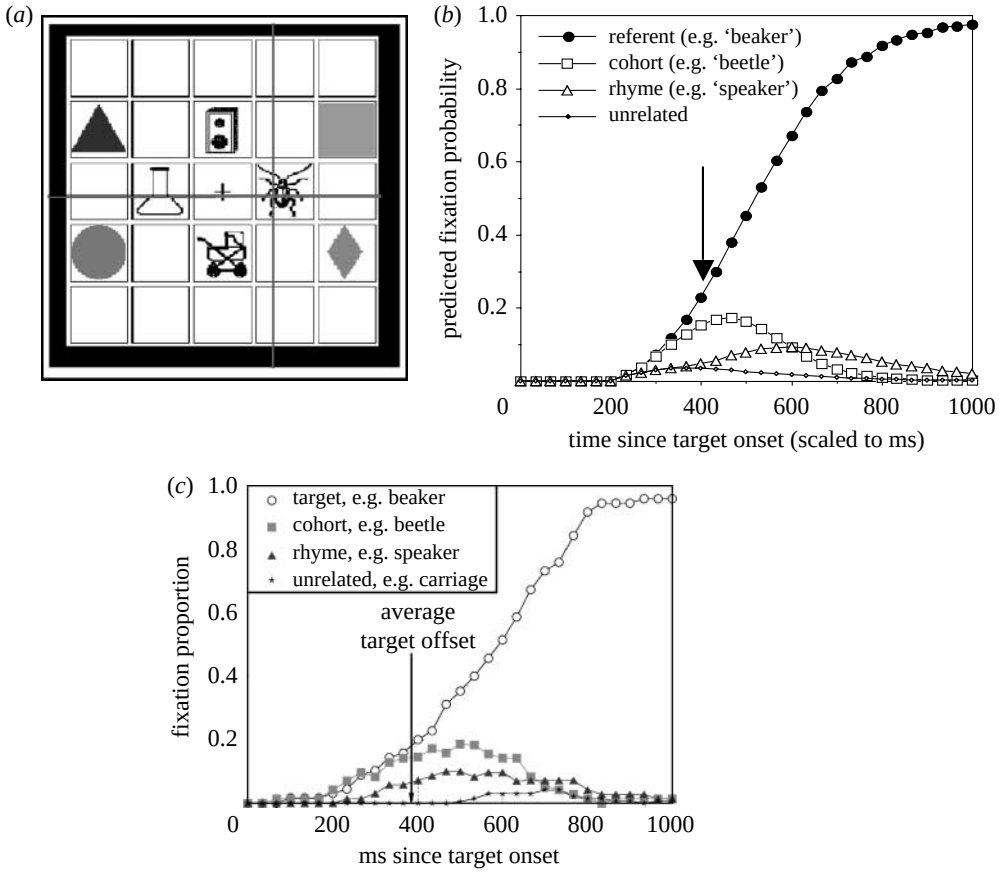

Figure 1. Shown are a sample display, simulations and data from Allopenna *et al.* (1998). (*a*) Sample display, (*b*) simulations of fixation proportions using TRACE and the linking hypothesis, and (*c*) the behavioural data. All figures are adapted from Allopenna *et al.* (1998).

Allopenna *et al.* (1998) examined the time course of activation words that share initial phonemes with a target word (e.g. *beaker* and *beetle*), which we will refer to as 'cohort competitors' that are predicted to compete by onset-similarity models, or words that rhyme with the target word (e.g. *beaker* and *speaker*), which are predicted to compete by global similarity models. Participants followed spoken instructions to move one of four objects displayed on a computer screen using the computer mouse (e.g. 'Look at the cross. Pick up the beaker. Now put it above the square'). Critical trials included cohort competitors (e.g. *beetle*) and/or rhyme competitors (*speaker*), and unrelated baseline items (e.g. *carriage*), as illustrated in figure 1*a*. The assumption linking fixations to continuous word recognition processes is that as the instruction unfolds the probability that the listener's attention will shift to a potential referent of a referring expression increases with the activation (evidence for) of its lexical representation, with a saccadic eye movement typically following a shift in visual attention to the region in space where attention has moved. Since saccades are rapid, low cost, low-threshold responses, some saccades will be generated based on even small increases in activation, with the likelihood of a saccade increasing as activation increases. Thus, while each saccade is a discrete event, the probabilistic nature of saccades ensures that, with sufficient numbers of observations, the results will begin to approximate a continuous measure. For an insightful discussion, including the strengths and weaknesses of eye movements compared with a truly continuous measure, tracking the trajectories of hand movements, see Magnuson (2005) and Spivey *et al.* (2005).

Figure 1*b* shows the proportion of looks to each of the four pictures at each of 33 ms time slices, summed across trials and participants. The proportion of fixations maps onto phonetic similarity over time: targets and cohort competitor proportions increase and separate from the rhyme and unrelated baseline of approximately 200 ms after the onset of the target word (approx. the delay required to program and launch a saccade). As the input becomes more similar to the rhyme, looks to its referent increase compared with the unrelated baseline. At approximately 200 ms after the first acoustic/phonetic information that is more consistent with the target, fixation proportions to the cohort begin to drop off, returning to the unrelated baseline sooner than rhyme fixations. Simulations of these data using TRACE, and a formal linking hypothesis between activation in the model and likelihood of fixation, account for more than 90% of the variance in the time course of fixation proportions to the target and competitors.

These results suggest that the processing neighbourhood changes dynamically as a word unfolds. Early in processing, competition will be stronger for words with many cohort competitors compared with few cohort competitors; whereas, later in processing, competition will be stronger for words with a high density of globally

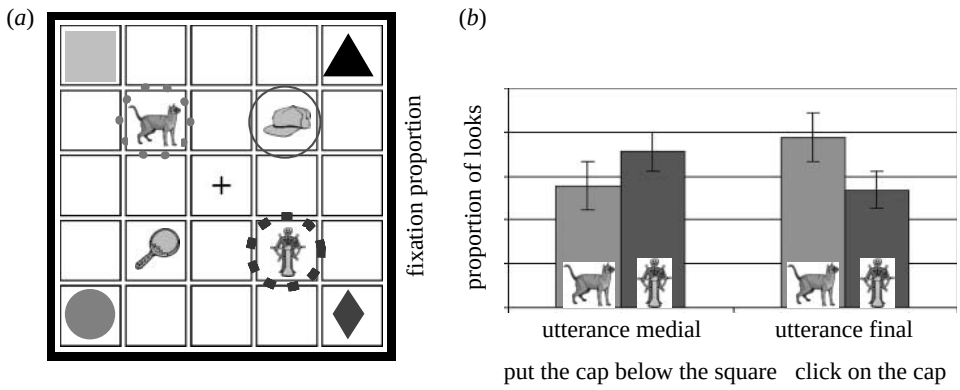

Figure 2. Sample display and proportion of looks to (*a*) the cohort competitor (picture of a cat) and (*b*) embedded carrier word competitor (picture of a captain) in utterance medial and utterance final positions.

similar competitors. Magnuson *et al.* (2007) report just this result using a display in which all of the pictures had unrelated names, and potential competitors were never pictured and never named. Magnuson *et al.* (2007) compared frequency-matched targets that differed in number of cohort competitors and neighbourhood density (words that differ from the target by one phoneme, or by adding or subtracting a phoneme). Even the earliest fixations to target words with many cohort competitors were delayed relative to fixations to targets with fewer cohort competitors. In contrast density affected only later fixations to the target.

These results also address a potentially troubling methodological concern with the visual world approach. The use of a visual world with a limited set of pictured referents and potential actions creates a more restricted environment than language processing in many, if not most, contexts. Certainly, these closed-set characteristics impose more restrictions than most psycholinguistic tasks. In the Allopenna *et al.* (1998) paradigm, the potential response set on each trial was limited to four pictured items. If participants adopted a task-specific strategy, such as implicitly naming the pictures, then the unfolding input might be evaluated against these activated names, effectively bypassing the usual activation process. However, if this were the case, one would not expect to find effects of non-displayed competitors, as did Magnuson *et al.* (2007; for further discussion and other relevant data see Dahan & Tanenhaus (2004, 2005), Salverda & Altmann (2005) and Dahan *et al.* (2007)). Most crucially, the same linking hypothesis predicts the time course of looks to targets in experiments using displayed and non-displayed competitors (for specific examples, which compare displayed and non-displayed competitors, see Dahan *et al.* (2001*a*,*b*)).

In the procedure introduced by Allopenna *et al.* (1998), the time course of lexical processing is measured to words that are embedded in utterances. This is important because the prosodic environment in which a word occurs systematically affects its acoustic/phonetic realization. Recent studies of the processing of words such as *captain,* which begin with a phonetic sequence that is itself a word, e.g. *cap,* illustrate this point. One might think that the presence of embedded words would present a challenge to spoken word recognition. However, the language processing system exploits small systematic differences in vowel duration. In particular, the vowel in a monosyllabic word such as *cap* is typically longer than the same vowel in a polysyllabic word such as *captain*. (Davis *et al.* 2002; Salverda *et al.* 2003). However, the difference in vowel duration changes with the prosodic environment; it is smallest in the middle of a phrase and largest at the end of a phrase. Consequently, the extent to which an embedded word and its carrier compete for recognition varies with position in an utterance (Crosswhite *et al.* in preparation). More generally, prosodic factors will modulate the relative degree to which different members of a neighbourhood will be activated in different environments. A striking demonstration comes from a recent study from our laboratory (Salverda *et al.* 2007). Figure 2*a* shows a sample display with pictures of a cap, captain, cat and a picture with an unrelated name, a mirror, used with instructions in which *cap* is in either medial or final position. Figure 2*b* shows that in medial position *captain* is a stronger competitor than *cat*, whereas the opposite pattern is seen in utterance-final position.

Prosodic influences on processing neighbourhoods have broad implications for the architecture of word recognition system because pragmatic factors can have strong influences on prosody. For example, the duration of the first vowel in *captain* is similar to the typical duration of the vowel in *cap* when stress is being used to signal *contrast*, e.g. *The CAPtain was responsible for the accident*. It is possible then, that, during word recognition, the weighting of a sub-phonetic factor such as vowel duration might be modulated by a high-level property of the utterance. Evaluating this hypothesis requires examining the recognition of words embedded in an utterance at a fine temporal grain, and in a context rich enough to manipulate contrast. This can be accomplished with relatively minor extensions of the Allopenna *et al.* (1998) paradigm to create a richer discourse context. We now turn to studies that use real-world objects to focus on a particular type of context, the referential domain within which a linguistic expression is interpreted.

4. REFERENTIAL DOMAINS: EFFECTS OF ACTION-BASED AFFORDANCES

Many linguistic expressions can be understood only with respect to a circumscribed context or referential domain. Definite referring expressions are a paradigmatic

example. Felicitous use of a definite noun phrase (NP) requires reference to, or introduction of, a *uniquely identifiable* entity (e.g. Roberts 2003). For example, imagine one is playing the role of sous-chef. If two bowls were in front of you on the kitchen counter, your cooking partner could not felicitously ask you to *pour the milk into the bowl*. Instead, he would have to use the indefinite phrase, *a bowl*. He could, however, say *the bowl*, if only one bowl was on the counter, and other bowls were on a visible shelf, or if there were several bowls on the counter and pouring the milk was embedded in a sequence of actions to add ingredients to a mixture in that particular bowl. The definite expression is felicitous in these situations because the satisfaction of uniqueness takes place with respect to a relevant context, or referential domain.

Recent research demonstrates that listeners dynamically update referential domains based on expectations driven by linguistic information in the utterance and the entities in the visual world (Eberhard *et al.* 1995; Altmann & Kamide 1999; Chambers *et al.* 2002). For example, in Eberhard *et al.* (1995), the participants touched one of four blocks that differed in marking, colour or shape. With instructions such as *Touch the starred yellow square*, the participants launched an eye movement to the target block on an average of 250 ms after the end of the word that uniquely specified the target with respect to the visual alternatives. In the example, the earliest possible point of disambiguation (POD) is after *starred* when only one of the blocks is starred, and after *square* when there are two starred yellow blocks. Similar results are obtained with more complex instructions and displays. With a display of seven miniature playing cards, including *two* five of hearts, Eberhard *et al.* (1995) used instructions such as, *Put the five of hearts that is below the eight of clubs above the three of diamonds*. To manipulate the POD, we varied whether or not the competitor five of hearts had a card above it, and if so, whether it differed in denomination or suit from the card above the target five. The following is a representative sequence of fixations. As the participant heard *the five of hearts*, she successively looked at each of the two potential referents. After hearing *below the*, she immediately looked at a 10 of clubs, which was above the (competitor) 5 that she had been fixating on. By the end of *clubs*, her eyes moved to interrogate the card above the other five, the eight of clubs, thus identifying that five as the target. The eye immediately shifted down to the target card and remained until the hand began to grasp the target, at which point gaze shifted to the three of diamonds.

In collaboration with Chambers and colleagues, we asked whether referential domains take into account the affordances of potential real-world referents with respect to the action evoked by the instruction. Chambers *et al.* (2002; Experiment 2) presented the participants with six objects in a workspace, as illustrated in figure 3*a*. On test trials, the objects included a large and a small container, e.g. a large can and a small can. We manipulated whether the to-be-moved object, the cube in figure 3*a*, could fit into both of the containers, as was the case for a small cube, or only fit into the larger container, as was the

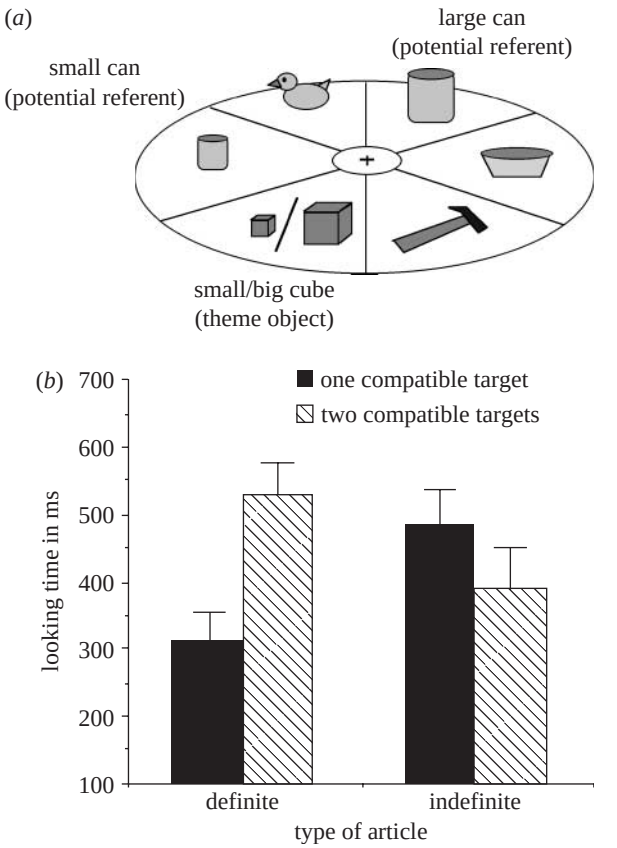

Figure 3. (*a*) Sample stimuli. The small cube will fit into both cans but the large cube will only fit into the big can. (*b*) The mean latency to launch an eye movement to the goal with definite and indefinite instructions and one and more than one compatible goal referents.

case for a large cube. Instructions for the display were: *Pick up the cube. Now put it inside a/the can*.

The size of the theme-object determined whether one or two of the potential goals (containers) were compatible referents. The instructions manipulated whether the goal was introduced with the definite article, *the*, which presupposes a unique referent, or the indefinite article, *a*, which implies that the addressee can choose from among more than one goal.

First, consider the predictions for the condition with the small cube. Here we would expect confusion when the definite article was used to introduce the goal because there is not a unique referent. In contrast, the indefinite article should be felicitous because there is more than one action-compatible goal referent. This is what we found: eye-movement latencies to fixate the goal chosen by the participant were slower in the definite condition compared with the indefinite condition. This confirms expectations derived from the standard view of how definite and indefinite articles are interpreted. Now consider predictions for the condition with the large cube, the theme-object that would fit into only one of the goal objects, i.e. the large can. If the referential domain consists of all of the objects in the visual world that meet the linguistic description in the utterance, that is both cans, then the pattern of results should be similar to that for the small cube. If, however, listeners dynamically update referential domains to include only those objects that afford the

required action, i.e. containers that the object in hand would fit into, then only the large can is in the relevant referential domain. Therefore, use of a definite description, e.g. *the can* should be felicitous, because the cube could be put into only one can, whereas an indefinite description, e.g. *a can*, should be confusing.

Figure 3*b* shows the predicted interaction between definiteness and compatibility. Eye-movement latencies to the referent following the definite referring expressions were faster when there was only one compatible referent compared with when there were two compatible referents, whereas the opposite pattern occurred for the indefinite expressions. Moreover, latencies for the one-referent compatible condition were comparable to control trials in which only a single object met the referential description in the instruction, e.g. trials with only a single large can. Thus, referential domains can be dynamically updated to take into account the real-world properties of potential referents with respect to a particular action.

To further evaluate the claim that intended actions were indeed constraining the referential domain Chambers *et al*. (2002) conducted a second experiment in which the second instruction was modified to make it into a question, e.g. *Pick up the cube. Could you put it inside a/the can?* In order to prevent the participants from interpreting the question as an indirect request, the participant first answered the question. On about half of the trials when the participant answered 'yes', the experimenter subsequently asked the participant to perform the action. Unlike following a command, answering a question does not require the participant to perform an action that brings the affordance restrictions into play. Thus, the referential domain should now take into account all the potential referents that satisfy the linguistic description, not just those that would be compatible with possible action mentioned in the question. If referential domains take into account behavioural goals then, under these conditions, definite expressions should be infelicitous regardless of compatibility, whereas indefinite expressions should always be felicitous. This is what we found. Time to answer the question was longer for questions with definite compared with indefinite referring expressions. Crucially, definiteness did not interact with compatibility (i.e. size of the theme-object). Moreover, compatibility had no effect on response times for questions with definite articles. These results demonstrate that referential domains are dynamically updated using information about available entities, properties of these entities and their compatibility with the action evoked by the utterance. This notion of referential domain is consistent with the rich view of context endorsed by researchers in the action tradition.

Assignment of reference necessarily involves mapping linguistic utterances onto entities in the world, or a conceptual model thereof. A crucial question, then, is whether these contextually defined referential domains influence core processes in language comprehension that many have argued operate without access to contextual information. In order to address this question, we examined whether action-based referential domains affect the earliest moments of syntactic ambiguity resolution.

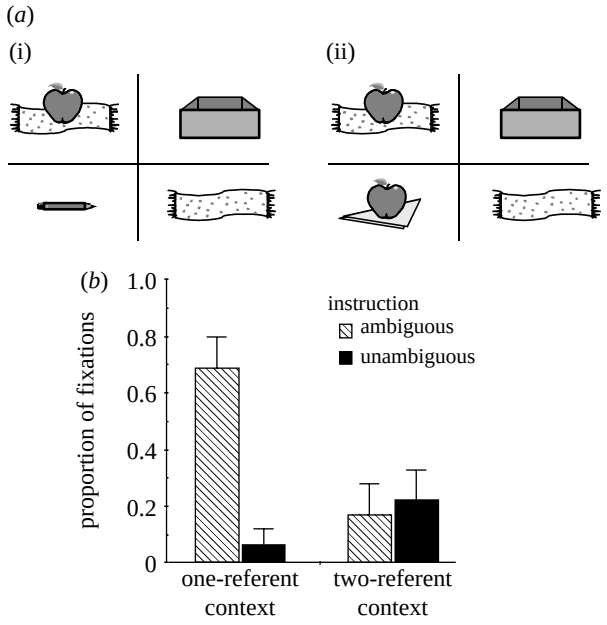

Figure 4. (*a*) Sample stimuli for (i) one-referent (pencil) and (ii) two-referent (apple on napkin) conditions. (*b*) The proportion of looks to the competitor goal (the towel) for instructions with locally ambiguous and unambiguous prepositional phrases in one-referent and two-referent contexts.

Previously, we noted the temporary ambiguity in an utterance, such as, *Put the apple on the towel…* Temporary 'attachment' ambiguities like these have long served as a primary empirical test bed for evaluating models of syntactic processing (Tanenhaus & Trueswell 1995). Crain & Steedman (1985; also Altmann & Steedman 1988) noted that in many attachment ambiguities, the ambiguous phrase could either modify a definite NP or introduce a syntactic complement (argument) of a verb phrase. Under these conditions, the argument analysis is typically preferred. For instance, in Example 1, listeners will initially misinterpret the prepositional phrase, *on the towel*, as the goal argument of *put* rather than as an adjunct modifying the NP, *the apple*, resulting in temporary confusion.

Example 1. *Put the apple on the towel into the box*

Crain & Steedman (1985) noted that one use of modification is to differentiate an intended referent from other alternatives. For example, it would be odd for Example 1 to be uttered in a context in which there was only one perceptually salient apple, whereas it would be natural in contexts with more than one apple. In the latter context, the modifying phrase, *on the towel*, provides information about which of the apples is intended. They proposed that listeners might initially prefer the modification analysis to the argument analysis in situations that provided the appropriate referential context. They also argued that referential fit to the context, rather than syntactic complexity, was the primary factor controlling syntactic preferences (also see Altmann & Steedman 1988).

Tanenhaus *et al*. (1995) and Spivey *et al*. (2002) compared the processing of temporarily ambiguous sentences such as *Put the apple on the towel in the box* and unambiguous control sentences, such as *Put the apple*

that's on the towel in the box, in contexts such as the ones illustrated in figure 4.

The objects were placed on a table in front of the participants. Eye movements were monitored as they followed the spoken instruction. The results, which are presented in figure 4*c*, provided clear evidence for immediate use of the visual context. In the one-referent context, the participants frequently looked at the false (competitor) goal, indicating that they initially misinterpreted the prepositional phrase, *on the towel,* as introducing the goal. Looks to the competitor goal were dramatically reduced in the two-referent context. Crucially, the participants were no more likely to look at the competitor goal with ambiguous instructions compared to the unambiguous baseline (also see Trueswell *et al.* 1999).

Clearly, then, referential context can modulate syntactic preferences from the earliest moments of syntactic ambiguity resolution. But is the relevant referential domain defined by all of the salient entities that meet the referential description in the utterance or can it be dynamically updated based on real-world constraints, including action-based affordances of objects? Chambers *et al.* (2004) addressed this question using temporarily ambiguous instructions such as, *Pour the egg in the bowl over the flour*, and unambiguous instructions such as, *Pour the egg that's in the bowl over the flour*, with displays such as the one illustrated in figure 5.

The display for test trials included the goal (the flour), a competitor goal (the bowl), the referent (the egg in the bowl), and a competitor referent (the egg in the glass). The referent was always compatible with the action evoked by the instruction, e.g. the egg in the bowl was liquid and therefore could be poured. We manipulated whether the competitor referent was also compatible with the action evoked by the instruction, e.g. one can pour a liquid egg, but not a solid egg. In the compatible condition, the other potential referent, the egg in the glass, was also in liquid form. In the incompatible condition, it was an egg in a shell. The crucial result was the time spent looking at the competitor goal, which is presented in figure 5*c*.

When both potential referents matched the verb (e.g. the condition with two liquid eggs, as in figure 5*a*), there were few looks to the false goal (e.g. the bowl) and no differences between the ambiguous and unambiguous instructions. Thus, the prepositional phrase was correctly interpreted as a modifier, replicating the pattern found by Spivey *et al.* (2002) for two-referent contexts. However, when the competitor was incompatible, as in figure 5*c* (e.g. the condition where there was a liquid egg and a solid egg), we see the same data pattern as Spivey *et al.* (2002) found with one-referent contexts. Participants were more likely to look to the competitor goal (the bowl) with the ambiguous instruction than with the unambiguous instruction. Thus, listeners misinterpreted the ambiguous prepositional phrase as introducing a goal only when a single potential referent (the liquid egg) was compatible with a pouring action.

Unlike the Spivey *et al.* (2002) study, which used the verb *put,* all of the relevant affordances were related to properties that might be plausibly be attributed to the lexical semantics of the verb. For example, *pour* requires

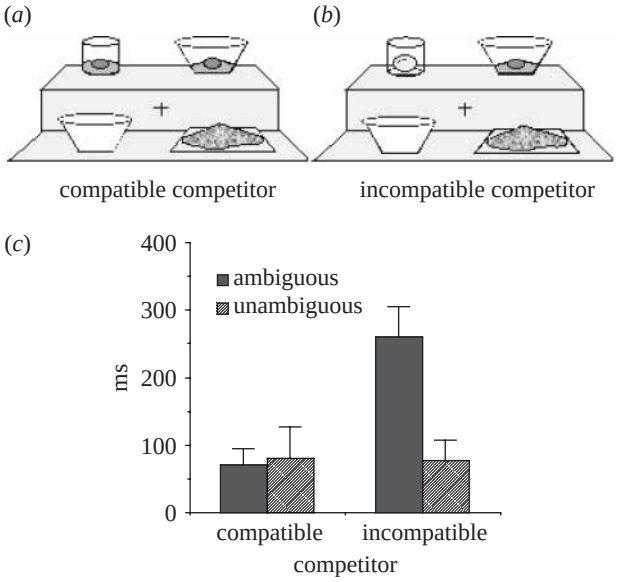

Figure 5. Sample stimuli for trials with (*a*) action-compatible competitor (two liquid eggs) and (*b*) action-incompatible competitor (one solid egg). (*c*) The mean proportion of time spent looking at the competitor goal (the empty bowl) for instructions with locally ambiguous and unambiguous prepositional phrases with action-compatible and action-incompatible competitors.

its theme to have the appropriate liquidity for pouring. There is precedent going back to Chomsky (1965) for incorporating a subset of semantic features, so-called 'selectional restrictions', into lexical representations when those features have syntactic or morphological reflexes in at least some languages. Thus, it could be argued that only real-world properties referred to with selectional restrictions can influence syntactic processing.

Chambers *et al.* (2004) addressed this issue in a second experiment. The critical instructions contained *put* (e.g. *Put the whistle* (*that's*) *on the folder in the box*), a verb that obligatorily requires a goal argument. Figure 6*a* shows a corresponding display, containing potential referents that are whistles, one of which is attached to a loop of string. Importantly, *put* does not constrain which whistle could be used in the action described by the instruction.

The compatibility of the referential competitor was manipulated by varying whether or not the participants were provided with an instrument. The experimenter handed the instrument to the participant without naming it. For example, before the participants were given the instruction described earlier, they might be given a small hook. Critically, this hook could not be used to pick up the competitor whistle without a string. Thus, upon hearing *put the*... the competitor could be excluded from the referential domain based on the affordances of the object with respect to the intended action, i.e. using the hook to move an object. If so, the participants should misinterpret *on the folder* as the goal only when ambiguous instructions are used *and* when an instrument is provided. If, however, the relevant referential domain is defined using only linguistic information, then a goal misanalysis should occur regardless of whether an instrument is supplied beforehand. Figure 6*b* shows the mean time spent fixating the false goal object within the

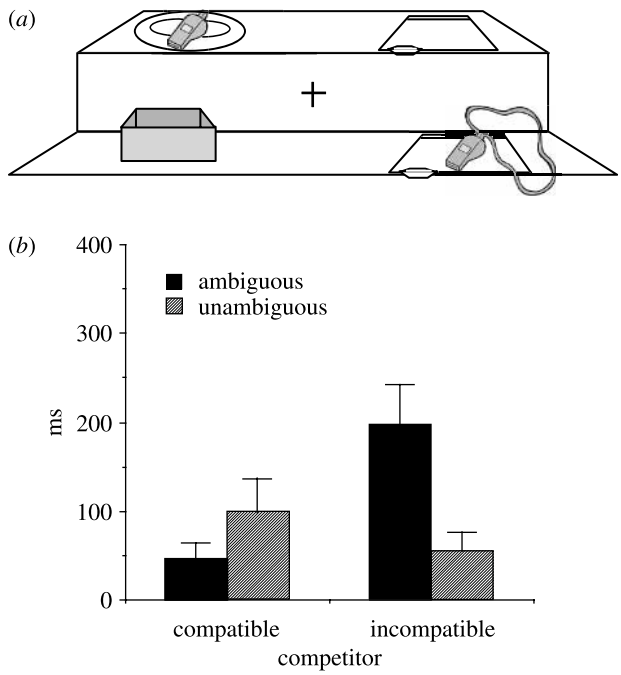

Figure 6. (*a*) Sample stimuli. Both whistles can be moved by hand, but only the whistle with the string attached can be picked up with a hook. (*b*) The proportion of time spent looking at the competitor goal when the presence or absence of an instrument makes the competitor action-compatible or action-incompatible.

2500 ms after the first prepositional phrase. The false goal is most often fixated when ambiguous instructions are used and the competitor cannot afford the evoked action. The remaining conditions all show fewer fixations to the false goal.

Thus, the syntactic role assigned to a temporarily ambiguous phrase varies according to the number of possible referents that can afford the action evoked by the unfolding instruction. The same results hold regardless of whether the constraints are introduced linguistically by the verb, or non-linguistically by the presence of a task-relevant instrument. Thus, the referential domain for an initial syntactic decision was influenced by the listener's consideration of how to execute an action—an information source that cannot be isolated within the linguistic system. This action itself can be partially determined by situation-specific factors such as the presence of a relevant instrument. The syntactic role assigned to an unfolding phrase in turn depends on whether these factors jointly determine a unique referent without additional information. These results add to the growing body of literature indicating that multiple constraints affect even the earliest moments of syntactic processing. They are incompatible with the claim that the language processing includes subsystems (modules) that are informationally encapsulated, and thus isolated from high-level expectations (Fodor 1983; Coltheart 1999).

So far, we have described experiments that extend investigations of real-time language processing into more natural real-world tasks. However, it was not clear that this approach could be used to study unscripted interactive conversation. We now turn to some work in progress, which shows that such studies are both tractable and informative.

5. REAL-TIME LANGUAGE PROCESSING IN INTERACTIVE CONVERSATION

Since Stalnaker's pioneering work on mutual knowledge (Stalnaker 1978, 2002), formal theories of discourse in both computational linguistics and pragmatics and semantics have assumed that keeping track of what is known, and not known, to the individual participants in a discourse is fundamental for coordinating information flow (Clark 1992, 1996; Brennan & Hulteen 1995). For example, a speaker making a statement (such as *The coffee's ready*) is expected to contribute information that is not known to the listener, a choice that involves judgments about the information states of the participants. If the addressee is already holding a fresh cup of coffee, this is not likely to be an informative contribution. The speaker's statement reflects what the speaker takes to be not yet commonly known at that point in the discourse. Asking a question (*Is the coffee ready?*) would similarly seem to reflect both the speaker's assessment of the addressee's information state—i.e. that the addressee is in a position to provide the answer—and the speaker's own state of ignorance or uncertainty.

However, conversational partners may not continuously update their mental representations of each other's knowledge. Building, maintaining and updating a model of a conversational partner's beliefs could be memory intensive (Keysar *et al.* 1998). In addition, many conversational situations are constrained enough that an individual participant's perspective will provide a sufficient approximation of the knowledge and beliefs shared between interlocutors. Moreover, information about another's beliefs can be uncertain at best. For these reasons, Keysar and colleagues propose that whereas considerations of an interlocutor's perspective might control language performance at a macro level, the moment-by-moment processes that accomplish production and comprehension could take place relatively egocentrically. Indirect supporting evidence comes from a growing number of studies demonstrating that speakers typically do not adapt the form of their utterances to avoid constructions that are difficult for listeners (Brown & Dell 1987; Bard *et al.* 2000; Ferreira & Dell 2000; Keysar & Barr 2005; but see Metzing & Brennan 2003). More direct evidence comes from studies showing that addressees often fail to reliably distinguish their own knowledge from that of their interlocutor when interpreting a partner's spoken instructions (Keysar *et al.* 2000, 2003). For example, in Keysar *et al.* (2000), participants were seated on opposite sides of a vertical grid of squares, some of which contained objects. Most of the objects were in 'common' ground because they were visible from both sides of the display, but a few objects in the grid were hidden from the director's view, and thus were in the matcher's privileged ground. On critical trials, the director, a confederate, referred to a target object in common ground using an expression that could also refer to a hidden object in privileged ground, which was always the more prototypical referent for the expression. Matchers initially preferred to look at the hidden object, and on some trials even picked it up and began to move it. Subsequent studies that equate the typicality of potential referents in common and

privileged ground also find intrusions of information from privileged ground. However, under these conditions, addressees seem to make partial use of common ground from the earliest moments of processing (Nadig & Sedivy 2002; Hanna *et al.* 2003).

Thus, whether and, if so, when interlocutors seek and use information about each other's probable intentions, commitments and probable knowledge remain open questions. The answers will determine which classes of theoretical constructs can be imported from semantics, pragmatics and computational linguistics. They will also determine the extent to which production and comprehension can be viewed as encapsulated from social/cognitive processes. However, the standard approaches that have been used to address these questions have serious problems. One problem is that those studies that have shown the strongest support for use of common ground may invite the subject to adopt the speaker's perspective, e.g. by drawing attention to the mismatch in perspective by mislabelling objects. The second problem is that using a confederate to generate instructions eliminates many of the natural collaborative processes that occur in interactive conversation and dramatically changes the form of the language. More seriously, in studies with confederates, all of the instructions are simple declarative commands, which carry the presupposition that there is a unique action that the addressee can perform, in effect attributing omniscience to the hidden speaker. In addition, asking a participant to follow instructions from a director may create a weak version of the suspension of skepticism that can occur in situations where there is an authority giving directions (e.g. an experimenter, a health professional, etc.). The addressee aims to do what he is told on the assumption that the person generating the instruction has the relevant knowledge.

The solution, which is to examine interactive conversation in unscripted joint tasks, is rife with methodological challenges. The experimenter gives up a substantial degree of control because trials cannot be scripted in advance. Rather, they have to emerge from the conversational interaction. Thus, tasks have to be carefully crafted to generate appropriate trials and baseline control conditions. Data analysis is time consuming because the conversation and state of the workspace have to be transcribed in order to identify relevant trials for subsequent data analysis. In addition, as we have seen, the form of the language differs from the sanitized language used in the typical psycholinguistic experiment with pre-recorded materials. With these challenges in mind, we conducted a series of preliminary investigations to determine the feasibility of monitoring real-time language processing with naive participants during task-oriented interactive dialogue.

Our approach adopts a 'targeted language games' methodology. Pairs of naive participants complete a type of language game—a referential communication task (Krauss & Weinheimer 1966)—while gaze and speech are monitored. These language games are 'targeted' in that the parameters of the game are carefully designed to elicit specific types of utterances, without explicitly restricting what the participants say. The task is structured so that the conditions of interest that would be manipulated in a typical within-subjects design, including control conditions, emerge during the game. Each utterance is naturally produced assuring that it is contextually appropriate. Conversations are lengthy enough to generate sufficient trials in the conditions of interest to approximate a standard factorial design. We then compare characteristics of these utterances and the corresponding eye movements across conditions. In what follows, we test the validity of the methodology by replicating some standard findings. We then extend the methodology to investigate aspects of language production and comprehension that have been difficult or impossible to study using standard techniques. In doing so, we make novel observations that clarify how interlocutors generate and comprehend referential expressions and how and when they take into account their partner's perspective—an ability that comprises part of our theory of mind (see Keysar *et al.* 2003).

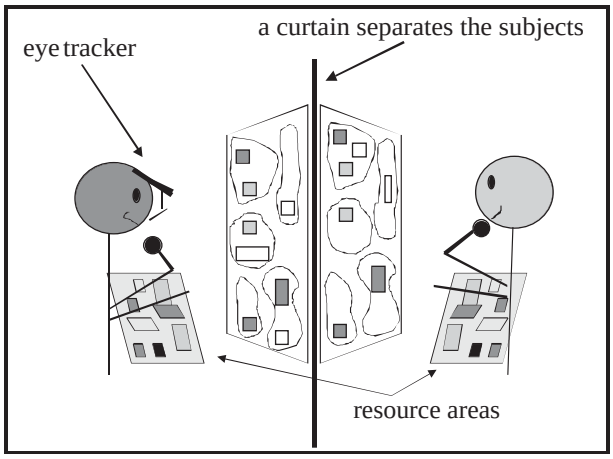

Figure 7. Schematic of the setup used in the referential communication task. Shaded squares and rectangles represent blocks and unshaded squares and rectangles represent stickers (which will eventually be replaced with blocks). The scene pictured is midway through the task, so some portions of the partners' boards match, while other regions are not completed yet.

(a) *Language interpretation*

In our initial experiments, pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and confirm those configurations (Brown-Schmidt *et al.* 2005; Brown-Schmidt & Tanenhaus in press). The characteristics of the blocks afforded comparison with findings from scripted experiments investigating language-driven eye movements, specifically those demonstrating POD effects during reference resolution. We investigated: (i) whether these effects could be observed in a more complex domain during unrestricted conversation and (ii) under what conditions the effects would be eliminated, indicating that factors outside of the speech itself might be operating to circumscribe the referential domain. Figure 7 presents a schematic of the experimental setup. We divided participants' boards into five physically distinct subareas, within which the blocks were arranged. Most of the blocks were of assorted shapes (square or rectangle) and colours (red, blue, green, yellow, white or black). The configuration of the blocks was such that their

colour, size and orientation would encourage the use of complex NPs. Other blocks contained pictures that could be easily described by naming the picture (e.g. 'the candle'). We included pairs whose names were cohort competitors, e.g. *clown* and *cloud*.

Partners were highly engaged and worked closely with one another to complete the task. Each pair worked through the game board in a different way, and none used overt strategies, such as establishing a grid system. All pairs made frequent references to the game pieces. Referential expressions were indefinite NPs, definite NPs and pronouns. We designed the task to focus exclusively on the interpretation of definite references, and this is what we turn to now.

The POD for each of the non eye-tracked partner's definite references to coloured blocks was defined as the onset of the word in the NP that uniquely identified a referent, given the visual context at the time. Just over half of these NPs (55%) contained a linguistic POD. The remaining 45% were technically ambiguous with respect to the sub-area that the referent was located in (e.g. *the red one* uttered in a context of multiple red blocks). Eye movements elicited by NPs with a unique linguistic POD were analysed separately from those that were never fully disambiguated linguistically. The eye-tracking analysis was restricted to cases where at least one competitor block was present. Eye movements elicited by disambiguated NPs are pictured in figure 8*a*. Before the POD, subjects showed a preference to look at the target block. Within 200 ms of the onset of the word in the utterance that uniquely specified the referent (POD), looks to targets rose substantially. This POD effect for looks to the target is similar to that seen by Eberhard *et al.* (1995), demonstrating that we were successful in using a more natural task to investigate online language processing. The persistent target bias and lack of a significant increase in looks to competitors are probably due to additional pragmatic constraints that we will discuss shortly.

For ambiguous utterances (figure 8*b*), fixations were primarily restricted to the referent, and there were very few requests for clarification. Thus, the speaker's underspecified referential expressions did not confuse listeners, indicating that referential domains of the speaker and the listener were closely coordinated. These results suggest that: (i) speakers systematically use less specific utterances when the referential domain has been otherwise constrained, (ii) the attentional states of speakers and addresses become closely coordinated, and (iii) utterances are interpreted with respect to referential domains circumscribed by contextual constraints. In order to identify what factors led speakers to choose underspecified referring expressions, and enabled addressees to understand them, we performed a detailed analysis of all of the definite references, focusing on factors that seemed likely to be influencing the generation and comprehension of referential expressions. We hypothesized that speakers would choose to make a referential expression more specific when the intended referent and at least one competitor block were each salient. We focused on recency, proximity and compatibility with task constraints—factors similar to those identified by

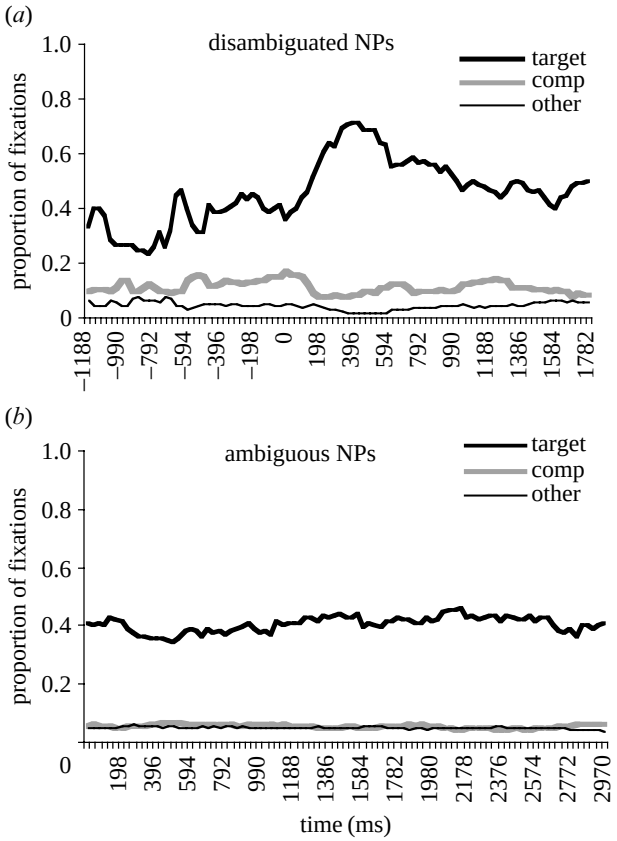

Figure 8. (*a*) The proportion of fixations to targets, competitors, and other blocks by time (ms) for linguistically disambiguated definite NPs. The graph is centred by item with 0 ms = POD onset. (*b*) The proportion of fixations for the linguistically ambiguous definite NPs.

Beun & Cremers (1998), who employed a task in which the participants, separated by a screen, worked together in a mutually co-present visual space to build a structure out of blocks.

(i) *Recency*

We assumed that recency would influence the salience of a referent, with the most recently mentioned entities being more salient than other (non-focused) entities. Thus, how recently the target block was last mentioned 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* was uttered in the context of a set of 10 blocks, 2 of which were green, recency would predict that the referent should be the green block that was most recently mentioned.

(ii) *Proximity*

We examined the proximity of each block to the last mentioned block, because partners seemed to adopt a strategy of focusing their conversation on small regions within each sub-area. Table 2 presents a segment of discourse where the referent of an otherwise ambiguous NP is constrained by proximity. The underlined referring expression is ambiguous given the visual context; there are approximately three green 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

Table 2. Excerpts from dialogue illustrating proximity and task compatibility constraints.

speaker	utterance
proximity constraint	
2	ok, so it is 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 *the one to the left of it, vertically up*?
task compatibility	
1	ok, you're gonna line it up… it is 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, take it two 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 *thuuh square*?

ambiguity because he considers only the block closest to the last mentioned block.

(iii) *Task compatibility*

Task compatibility refers to constraints on block placement due to the size and shape of the board, as well as the idiosyncratic systems that partners used to complete the task. In the exchange in table 2, compatibility circumscribes the referential domain as the partners strive to determine where the clown block should be placed. Again, the underlined referring expression is ambiguous given the visual context. While the general task is to make their boards match, the current sub-task is to place the clown piece (which they call *the green piece with the clown*). In order to complete this sub-task, Speaker 2 asks whether the clown should be lined up with the target, *thuuh square*. The listener does not have difficulty dealing with this ambiguous reference 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 due to the position of the other blocks or the design of the board.

For all ambiguous and disambiguated trials, each coloured block in the relevant sub-area was coded for recency (number of turns since last mention), proximity (ranked proximity to last mentioned item) and task constraints (whether or not the task predicted a reference to that block). Target blocks were more recently mentioned and more proximal than competitor blocks, and better fit the task constraints, establishing the validity of these constructs. However, recency, proximity and task compatibility of the target blocks did not predict speaker ambiguity. Ambiguity was, however, determined by the proximity and task constraints associated with the *competitor* blocks. When a competitor block was proximate and fit the task constraints, speakers were more likely to linguistically disambiguate their referential expression. A logistic regression model supported these observations: ambiguity was significantly predicted by a model that included task and proximity effects, with no independent contribution of recency.

These results suggest that the relevant referential domain for the speakers and addressees were restricted to a small task-relevant area of the board. Striking support comes from an analysis of trials in which there was a cohort competitor for the referent in the addressee's referential domain. Brown-Schmidt *et al.* (2005) found that looks to cohort competitors were no more probable than looks to competitors with unrelated names. This is not simply a null effect. Owing to the length of the experiment, the participants occasionally needed to take a bathroom break. Following the break, the eye tracker had to be recalibrated. The experimenter did so by instructing the participant to look at some of the blocks. The referential domain now consists of the entire display because there is no constraining conversational or task-based goal. When the intended referent had a cohort competitor, the participant frequently looked at the competitor, showing the classic cohort effect. A follow-up experiment replicated this finding, focusing exclusively on cohort trials that were in and outside the context of the conversation, by systematically incorporating simulated calibration checks (Brown-Schmidt & Tanenhaus in press).

In summary, these results demonstrate that it is possible to study real-time language processing in a complex domain during unrestricted conversation. When a linguistic expression is temporarily ambiguous between two or more potential referents, reference resolution is closely time-locked to the word in the utterance that disambiguates the referent, replicating effects found in controlled experiments with less complex displays and pre-scripted utterances. Most importantly, our results provide a striking demonstration that participants in a task-based or 'practical dialogue' (Allen *et al.* 2001), closely coordinate referential domains as the conversation develops.

(b) *Language Production*

In conversation, speakers often update messages on the fly based on new insights, new information and feedback from addressees, all of which can be concurrent with the speaker's planning and production of utterances. Thus, message formulation and utterance planning are interwoven in time and must communicate with one another at a relatively fine temporal grain. During the last two decades, detailed models of utterance planning have been developed to explain a growing body of evidence about how speakers retrieve lexical concepts, build syntactic structures and translate these structures into linguistic forms (Dell 1986; Levelt 1989; Bock 1995; Levelt *et al.* 1999; Indefrey & Levelt 2004). However, much less is known about how speakers plan and update the non-linguistic thoughts (messages) that are translated into utterances during language production, or about how message formulation and utterance planning are coordinated (but cf. Griffin & Bock 2000; Bock *et al.* 2003).

We created situations in which the speaker, while in the process of planning or producing an utterance, encounters new information that requires revising the message. If eye movements can be used to infer when the speaker first encounters that information, then the timing between the uptake of the new information and the form of the utterance might shed light on the interface between message formulation and utterance planning. In one study (Brown-Schmidt & Tanenhaus 2006), we explored this interface by exploiting properties of scalar adjectives.

Speakers typically use a scalar adjective, such as *big* or *small*, only when the relevant referential domain contains both the intended referent and an object of the same semantic type that differs along the scale referred to by the adjective (Sedivy 2003). Pairs of naive participants took turns describing pictures of everyday objects situated in scenes with multiple distractor objects. For example, on one trial the target was a picture of a large horse. In a scene with multiple *unrelated* objects, we would expect the speaker to refer to the target as *the horse*. Sometimes, however, our displays also contained a *contrast* object that differed from the target only in size (in this example, a small horse). On these trials, we expected that speakers would use a scalar adjective to describe the target, as in *the large horse*.

This paradigm allowed us to control what speakers referred to (the target was predetermined by us and highlighted on the screen), but the choice of how to refer to that entity, including whether a scalar adjective was used, was entirely up to the speaker. When a contrast object was present, the speaker's first fixation to the contrast should provide an estimate of when the speaker first encountered information that size must be included in the message.

Speakers tailored their messages to the referential context, rarely using a size adjective when there was not a contrast in the display; when there was a size contrast, approximately three-fourths of utterances included a size adjective. These proportions are consistent with those found by Sedivy and colleagues (Gregory *et al.* 2003; Sedivy 2003) in experiments that used simpler scenes with fewer objects.

Speakers typically gazed first at the highlighted target and then looked around the scene, sometimes fixating the contrast, and then returned to the target before speaking. When speakers looked at the contrast, over 80% of the referring expressions included a size modifier compared with less than 20% on trials when the contrast was not fixated. Thus, fixation on the contrast indexes whether size was included in the message. The position of the size adjective was variable (e.g. *the big horse*, versus *the horse… big one*). We reasoned that timing of the first fixation on the contrast would be linked to the planning of the message elements underlying the size adjective. Consistent with this prediction, we found that earlier size adjectives were associated with earlier first fixations on the contrast. For example, first fixations on the contrast for utterances like *the horse* (pause) *big one* were delayed by more than a second compared to first fixations for utterances like *the big horse*.

Having demonstrated that one can monitor real-time processes in both comprehension and production using unscripted interactive language, we can now examine how and when interlocutors take account of each other's knowledge, intentions and beliefs. In work in progress, we examine how having information about what an addressee knows shapes the form of the speaker's utterances, and how the addressee's knowledge guides the interpretation of these utterances (Brown-Schmidt *et al.* in press).

If speakers can distinguish their own knowledge from an estimate of their partner's knowledge during utterance formulation, they should use declarative and interrogative forms in complementary situations, asking questions when the addressee knows more, and vice-versa for declaratives. Declarative questions, which are marked with special intonation, share features of declarative statements and interrogatives; consistent use of these forms requires even more fine-grained distinctions between the relative knowledge of speaker and addressee. Unlike typical declaratives, which contribute information to the discourse, declarative questions do not commit the speaker to the propositional content of what is said (see C. Gunlogson 2001, unpublished data), and unlike an interrogative, declarative question do not preserve the speaker's neutrality. For example, a common use of declarative questions is to indicate surprise or skepticism, as seen in:

Example 2. *A: This is a painting by Chuck Close.*
B: That's a painting?

Appropriate use of different utterance forms would seem to crucially depend on the ability of a speaker to distinguish, if even at a coarse grain, information that is shared with a specific interlocutor from information that is not shared. Successful communication may also require estimating what private information an interlocutor might have. To examine these issues, we designed a targeted language game that elicited spontaneously produced questions and statements. Some task-relevant information was mutually known to the participants, and some was privately known. As in Keysar *et al.* (2000, 2003), the distinction between mutual and private knowledge was initially established by whether an object was visually available to one or both partners.

Pairs of naive participants rearranged game-cards as they sat on either side of a game-board made up of cubbyholes. A card with a picture of an identical animal figure on either side stood in each cubbyhole (figure 9). Some cubbyholes were blocked-off on one side making them visible to only the eye-tracked partner or only the non eye-tracked partner; the remaining cubbyholes were visible to both partners. Each card featured a clip art picture of an animal (pig, cow or horse), with an accessory (glasses, shoes or a hat).

At the beginning of the task, the cards were randomly arranged. The participants' task was to rearrange the cards such that no two adjacent squares matched with respect to type of animal or type of accessory (e.g. neither two horses nor two animals wearing hats could be adjacent). Participants were required to avoid matches in adjacent squares, some of

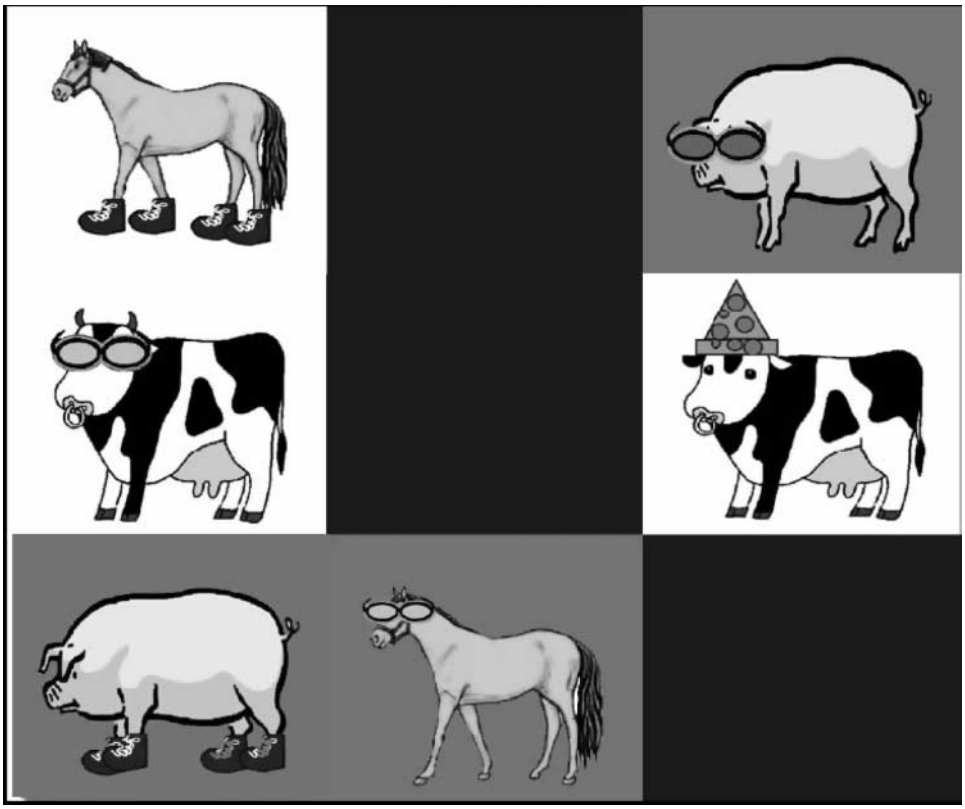

Figure 9. Schematic of part of the game board for the 'questions' experiment from the perspective of one of the participants. The animals in grey squares are in that participants privileged ground. The animals in the white squares are in common ground, that is, visible to both participants. The black squares contain animals that are only visible to the participant's partner.

which were hidden from each participant. The task therefore required extensive interaction.

We report a preliminary analysis of a subset of utterances produced by four pairs. We examined *responses to questions*, *declarative questions*, simple *declaratives*, and *wh-questions* that inquired about or made a statement about the identity of a card. Example 3 shows excerpts illustrating examples of each utterance type.

Example 3.

3.1 (*partner: What's below that?*)…*It is a horse with a hat*
3.2 *oh ok … You have a pig?*
3.3 *And then I got a pig with shoes next to that.*
3.4 *What's under the cow with the hat?*

Production analyses compared responses to questions (3.1), and declarative questions (3.2). These constructions share a similar grammatical form, but differ in communicative intent. *Comprehension* analyses compared declaratives (3.3) and wh-questions (3.4). An addressee who is sensitive to information state might take an interrogative form at the beginning of a question (e.g. *What*) as a cue that the speaker would be asking about cubbyholes that were in the addressee's private ground. In contrast, a declarative would indicate that the information was in common ground or the speaker's private ground.

Figure 10*a* shows the distribution of referent types for declaratives that were used either to respond to a question, or to ask a question. The referent of the utterance (e.g. the referent of *a horse with a hat* in 3.1 and *a pig* in 3.2) was identified and categorized in terms

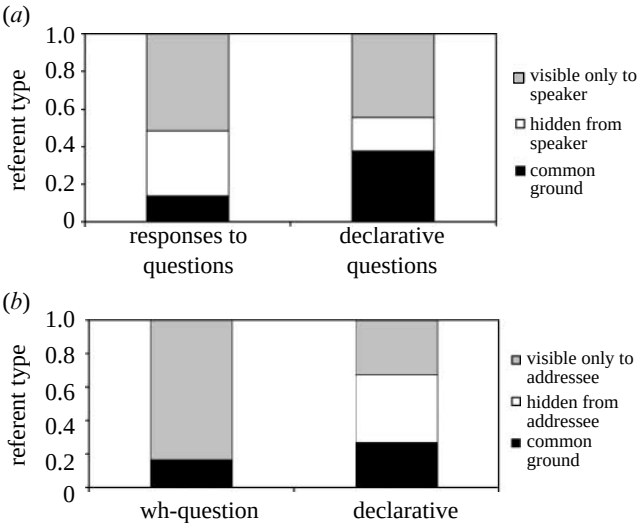

Figure 10. The proportion of referents for responses to questions, and declarative questions, categorized by mutuality of the referent, from the viewpoint of (*a*) the speaker and (*b*) the addressee.

of whether it was visible to just the speaker (private), just the addressee (hidden) or to both (shared). When speakers responded to a question, most of the time they referred to entities that only they could see (e.g. private), and to a lesser extent hidden and shared entities. In contrast, when asking a declarative question, the referent was likely to be either private or shared. Figure 10*b* shows the distribution of referent types for declaratives and wh-questions, analysed from the perspective of the addressee. When interpreting a question, the referent (e.g. the object of *What* in 3.4) is

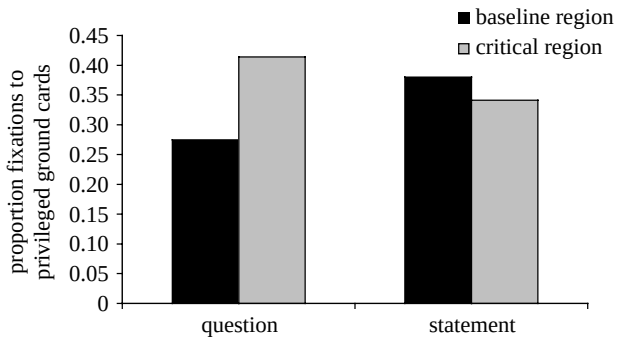

Figure 11. The average proportion of fixations to privileged ground cards (addressee's perspective) during interpretation of wh-questions and declaratives. Baseline region = 1000 ms before the onset of each expression + 200 ms; Critical region = 1000 ms after expression onset + 200 ms.

almost always in one of the addressee's private cubbyholes, indicating that the speaker asked about something she did not have direct information about. In contrast, when interpreting a declarative, the referent is most likely to be hidden from the addressee, and to a lesser extent shared or private.

We also examined addressee's eye movements as they heard wh-questions and declaratives. During wh-questions, approximately 42% of fixations were to private entities, and 43% to shared entities. In contrast, during declaratives, only approximately 30% of fixations were to private cubbyholes compared with 55% to shared cubbyholes. Most strikingly, as shown in figure 11, addressee looks to objects in private ground increase after the onset of interrogative form, e.g. after *what's* in a question such as, *What's under thee-uh, horse with the hat*. Following declaratives, looks to private entities decrease. Clearly, then, interlocutors in goal-oriented communicative tasks can use representations of their partner's perspective to guide real-time reference resolution.

6. CONCLUSIONS AND IMPLICATIONS

The studies that we have reviewed demonstrate the feasibility of examining real-time language processing in natural tasks. We showed that monitoring eye movements can be used to examine spoken word recognition in continuous speech, and why doing so is important. We then showed that actions, intentions and real-world knowledge circumscribe referential domains, and that these context-specific domains affect processes such as syntactic ambiguity resolution. Example 4 illustrates how this notion of a context-specific referential domain alters the standard view obtained from studies of decontextualized language.

Example 4. *After putting the pencil below the big apple, James put the apple on top of the towel.*

A traditional account would go something like this. When the listener encounters the scalar adjective *big*, interpretation is delayed because a scalar dimension can only be interpreted with respect to the noun it modifies (e.g. compare a big building and a big pencil). As *apple* is heard, lexical access activates the apple concept, a prototypical red apple. The apple concept is then modified resulting in a representation of a *big apple*. When *apple* is encountered in the second clause, lexical access again results in activation of a prototypical *apple* concept. Because *apple* was introduced by a definite article, this representation would need to be compared with the memory representation of the *big apple* to decide whether the two corefer (see Tanenhaus *et al.* (1985) for an outline of such an approach).

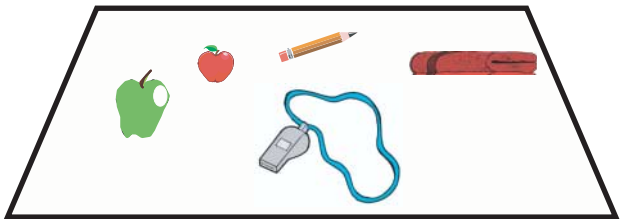

Figure 12. Hypothetical context for utterance *After putting the pencil below the big apple, James put the apple on top of the towel* to illustrate the implausibility of standard assumptions about context-independent comprehension. Note that the small (red) apple is intended to be a more prototypical apple than the large (green) apple.

This account of moment-by-moment processing seems reasonable when we focus on the processing of just the linguistic forms. However, let us reconsider how the real-time interpretation of Example 4 proceeds in the context illustrated in figure 12, taking into account results we have reviewed. At *big*, the listener's attention will be drawn to the larger of the two apples because a scalar adjective signals a contrast among two or more entities of the same semantic type. Thus, *apple* will be immediately interpreted as the misshapen (green) apple, even though a more prototypical (red) apple is present the display. And, when *the apple* is encountered in the second clause, the red apple would be ignored in favour of the large green apple.

This account is incompatible with the view that the initial stages of language processing create context-independent representations. However, it is compatible with increasing evidence throughout the brain and cognitive sciences that (i) behavioural context, including attention and intention affect even basic perceptual processes (Gandhi *et al.* 1998; Colby & Goldberg 1999) and (ii) brain systems involved in perception and action are implicated in the earliest moments of language processing (Pulvermüller *et al.* 2001). An important goal for future research will be integrating action-based notions of linguistic context with perceptual and action-based accounts of perception and cognition (cf. Barsalou 1999; Glenberg & Robertson 2000; Spivey & Dale 2004).

Finally, we showed that it is possible to examine real-time processing in unscripted interactive conversation at the same temporal grain as in tightly controlled studies with scripted utterances. When we do so, we find that interlocutors make immediate use of information about each other's probable perspective, knowledge, and intentions that is closely tied to the pragmatic constraints on different utterance types. This result is compatible with the growing evidence that social pragmatic cues such as joint attention and intentionality are critical in early language development (Bloom 1997; Sabbagh & Baldwin 2001), as well as evidence showing that non-linguistic gestures contribute to the understanding of speech (Goldin-Meadow 1999; McNeill 2000). By using natural tasks, and moving towards a methodological and theoretical union of the

action and product traditions, work in language processing can more fruitfully identify points of contact with these areas of research. Most generally, investigating real-time language processing in situations in natural tasks reveals a system in which multiple types of representations are integrated remarkably quickly, and a system in which processing language and using language for action are closely intertwined.

The research presented here has been generously supported by NIH grants HD 27206 and DC 005071. The first author would like to thank the numerous students and colleagues whose work is presented here.

REFERENCES

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L. & Stent, A. 2001 Towards conversational human–computer interaction. *AI Mag.* **22**, 27–35.

Allopenna, P. D., Magnuson, J. S. & Tanenhaus, M. K. 1998 Tracking the time course of spoken word recognition: evidence for continuous mapping models. *J. Mem. Lang.* **38**, 419–439. (doi:10.1006/jmla.1997.2558)

Altmann, G. T. M. & Kamide, Y. 1999 Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* **73**, 247–264. (doi:10.1016/S0010-0277(99)00059-1)

Altmann, G. T. M. & Steedman, M. J. 1988 Interaction with context during human sentence processing. *Cognition* **30**, 191–238. (doi:10.1016/0010-0277(88)90020-0)

Austin, J. L. 1962 *How to do things with words*. Cambridge, MA: Harvard University Press.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G. & Newlands, A. 2000 Controlling the intelligibility of referring expressions in dialogue. *J. Mem. Lang.* **42**, 1–22. (doi:10.1006/jmla.1999.2667)

Barsalou, L. 1999 Language comprehension: archival memory or preparation for situated action? *Discourse Process.* **28**, 61–80.

Beun, R.-J. & Cremers, A. H. M. 1998 Object reference in a shared domain of conversation. *Pragm. Cogn.* **6**, 121–151.

Bock, J. K. 1995 Sentence production: from mind to mouth. In *Handbook of perception and cognition*, vol. 11 (eds J. Miller & P. Eimas). Speech, language, and communication, pp. 181–216. New York, NY: Academic Press.

Bock, J. K., Irwin, D. E., Davidson, D. J. & Levelt, W. J. M. 2003 Minding the clock. *J. Mem. Lang.* **48**, 653–685. (doi:10.1016/S0749-596X(03)00007-X)

Bloom, P. 1997 Intentionality and word learning. *Trends Cogn. Sci.* **1**, 9–12. (doi:10.1016/S1364-6613(97)01006-1)

Brennan, S. E. 2005 How conversation is shaped by visual and spoken evidence. In *Approaches to studying world-situated language use: bridging the language-as-product and language-as-action traditions* (eds J. C. Trueswell & M. K. Tanenhaus), pp. 95–129. Cambridge, MA: The MIT Press.

Brennan, S. E. & Hulteen, E. 1995 Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Syst.* **8**, 143–151. (doi:10.1016/0950-7051(95)98376-H)

Brown, P. & Dell, G. 1987 Adapting production to comprehension: the explicit mention of instruments. *Cogn. Psychol.* **19**, 441–472. (doi:10.1016/0010-0285(87)90015-6)

Brown-Schmidt, S. & Tanenhaus, M. K. 2006 Watching the eyes when talking about size: an investigation of message formulation and utterance planning. *J. Mem. Lang.* **54**, 592–609. (doi:10.1016/j.jml.2005.12.008)

Brown-Schmidt S. & Tanenhaus, M. K. In press. Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cogn. Sci.*

Brown-Schmidt, S., Campana, E. & Tanenhaus, M. K. 2005 Real-time reference resolution in a referential communication task. In *Processing world-situated language: bridging the language-as-action and language-as-product traditions* (eds J. C. Trueswell & M. K. Tanenhaus), pp. 153–172. Cambridge, MA: The MIT Press.

Brown-Schmidt, S., Gunlogson, C. & Tanenhaus, M. K. In press. Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H. & Carlson, G. N. 2002 Circumscribing referential domains in real-time sentence comprehension. *J. Mem. Lang.* **47**, 30–49. (doi:10.1006/jmla.2001.2832)

Chambers, C. G., Tanenhaus, M. K. & Magnuson, J. S. 2004 Action-based affordances and syntactic ambiguity resolution. *J. Exp. Psychol. Learn. Mem. Cogn.* **30**, 687–696. (doi:10.1037/0278-7393.30.3.687)

Chomsky, N. 1965 *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

Clark, H. H. 1992 *Arenas of language use*. Chicago, IL: University of Chicago Press.

Clark, H. H. 1996 *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H. & Wilkes-Gibbs, D. 1986 Referring as a collaborative process. *Cognition* **22**, 1–39. (doi:10.1016/0010-0277(86)90010-7)

Colby, C. L. & Goldberg, M. E. 1999 Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22**, 97–136. (doi:10.1146/annurev.neuro.22.1.319)

Coltheart, M. 1999 Modularity and cognition. *Trends Cogn. Sci.* **3**, 115–120. (doi:10.1016/S1364-6613(99)01289-9)

Cooper, R. M. 1974 The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* **6**, 84–107. (doi:10.1016/0010-0285(74)90005-X)

Crain, S. & Steedman, M. 1985 On not being led up the garden path: the use of context by the psychological parser. In *Natural language parsing: psychological, computational, and theoretical perspectives* (eds D. Dowty, L. Karttunen & A. Zwicky), pp. 320–358. Cambridge, UK: Cambridge University Press.

Crosswhite, K., Masharov, M., McDonough, J. M. & Tanenhaus, M. K. In preparation. Phonetic cues to word length in the online processing of onset-embedded word pairs.

Dahan, D. & Tanenhaus, M. K. 2004 Continuous mapping from sound to meaning in spoken-language comprehension: evidence from immediate effects of verb-based constraints. *J. Exp. Psychol. Learn. Mem. Cogn.* **30**, 498–513. (doi:10.1037/0278-7393.30.2.498)

Dahan, D. & Tanenhaus, M. K. 2005 Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychol. Bull. Rev.* **12**, 455–459.

Dahan, D., Magnuson, J. S. & Tanenhaus, M. K. 2001*a* Time course of frequency effects in spoken word recognition: evidence from eye movements. *Cogn. Psychol.* **42**, 317–367. (doi:10.1006/cogp.2001.0750)

Dahan, D., Magnuson, J. S., Tanenhaus, M. K. & Hogan, E. 2001*b* Subcategorical mismatches and the time course of lexical access: evidence for lexical competition. *Lang. Cogn. Process.* **16**, 507–534. (doi:10.1080/01690960143000074)

Dahan, D., Tanenhaus, M. K. & Salverda, A. P. 2007 How visual information influences phonetically-driven saccades to pictures: effects of preview and position in display.

In *Eye movements: a window on mind and brain* (eds R. P. G. van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill), pp. 471–486. Oxford, UK: Elsevier.

Davis, M. H., Marslen-Wilson, W. D. & Gaskell, M. G. 2002 Leading up the lexical garden-path: segmentation and ambiguity in spoken word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 218–244. (doi:10.1037//0096-1523.28.1.218)

Dell, G. S. 1986 A spreading activation theory of retrieval in language production. *Psychol. Rev.* **93**, 283–321. (doi:10.1037/0033-295X.93.3.283)

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C. & Tanenhaus, M. K. 1995 Eye-movements as a window into spoken language comprehension in natural contexts. *J. Psychol. Res.* **24**, 409–436. (doi:10.1007/BF02143160)

Ferreira, V. S. & Dell, G. S. 2000 The effect of ambiguity and lexical availability on syntactic and lexical production. *Cogn. Psychol.* **40**, 296–340. (doi:10.1006/cogp.1999.0730)

Fodor, J. A. 1983 *Modularity of mind*. Cambridge, MA: Bradford Books.

Gandhi, S. P., Heeger, M. J. & Boyton, G. M. 1998 Spatial attention affects brain activity in human primary visual cortex. *Proc. Natl Acad. Sci. USA* **96**, 3314–3319. (doi:10.1073/pnas.96.6.3314)

Glenberg, A. M. & Robertson, D. A. 2000 Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* **43**, 379–401. (doi:10.1006/jmla.2000.2714)

Goldin-Meadow, S. 1999 The role of gesture in communication and thinking. *Trends Cogn. Sci.* **3**, 419–429. (doi:10.1016/S1364-6613(99)01397-2)

Gregory, M. L., Joshi, A., Grodner, D. & Sedivy, J. C. 2003 Adjectives and processing effort: so, uh, what are we doing during disfluencies? Paper presented at the *16th Annual CUNY Sentence Processing Conf.*, March, Cambridge, MA.

Grice, H. P. 1957 Meaning. *Phil. Rev.* **66**, 377–388. (doi:10.2307/2182440)

Griffin, Z. M. & Bock, J. K. 2000 What they eyes say about speaking. *Psychol. Sci.* **11**, 274–279. (doi:10.1111/1467-9280.00255)

Hanna, J. E., Tanenhaus, M. K. & Trueswell, J. C. 2003 The effects of common ground and perspective on domains of referential interpretation. *J. Mem. Lang.* **49**, 43–61. (doi:10.1016/S0749-596X(03)00022-6)

Hayhoe, M. & Ballard, D. 2005 Eye movements in natural behavior. *Trends Cogn. Sci.* **9**, 188–194. (doi:10.1016/j.tics.2005.02.009)

Indefrey, P. & Levelt, W. J. M. 2004 The spatial and temporal signatures of word production components. *Cognition* **92**, 101–144. (doi:10.1016/j.cognition.2002.06.001)

Keysar, B. & Barr, D. J. 2005 Coordination of action and belief in communication. In *Approaches to studying world situated language use: bridging the language-as-product and language-as-action traditions* (eds J. C. Trueswell & M. K. Tanenhaus), pp. 71–94. Cambridge, MA: The MIT Press.

Keysar, B., Barr, D. J., Balin, J. A. & Paek, T. S. 1998 Definite reference and mutual knowledge: process models of common ground in comprehension. *J. Mem. Lang.* **39**, 1–20. (doi:10.1006/jmla.1998.2563)

Keysar, B., Barr, D. J., Balin, J. A. & Brauner, J. S. 2000 Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychol. Sci.* **11**, 32–38. (doi:10.1111/1467-9280.00211)

Keysar, B., Lin, S. & Barr, D. J. 2003 Limits on theory of mind use in adults. *Cognition* **89**, 25–41. (doi:10.1016/S0010-0277(03)00064-7)

Krauss, R. M. & Weinheimer, S. 1966 Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *J. Person. Social Psychol.* **4**, 343–346. (doi:10.1037/h0023705)

Land, M. 2004 Eye movements in daily life. In *The visual neurosciences*, vol. 2 (eds L. Chalupa & J. Werner), pp. 1357–1368. Cambridge, MA: The MIT Press.

Levelt, W. J. M. 1989 *Speaking: from intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, W. J. M., Roelofs, A. P. A. & Meyer, A. S. 1999 A theory of lexical access in speech production. *Behav. Brain Sci.* **22**, 1–37. (doi:10.1017/S0140525X99001776)

Liversedge, S. P. & Findlay, J. M. 2000 Saccadic eye movements and cognition. *Trends Cogn. Sci.* **4**, 6–14. (doi:10.1016/S1364-6613(99)01418-7)

Luce, P. A. & Pisoni, D. B. 1998 Recognizing spoken words: the neighborhood activation model. *Ear Hear.* **19**, 1–36.

Magnuson, J. S. 2005 Moving hand reveals dynamics of thought. *Proc. Natl Acad. Sci. USA* **102**, 9995–9996. (doi:10.1073/pnas.0504413102)

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K. & Aslin, R. N. 2007 The dynamics of lexical competition during spoken word recognition. *Cogn. Sci.* **31**, 1–24. (doi:10.1207/s15516709cog000_90)

Marslen-Wilson, W. D. 1973 Linguistic structure and speech shadowing at very short latencies. *Nature* **244**, 522–523. (doi:10.1038/244522a0)

Marslen-Wilson, W. D. 1975 Sentence perception as an interactive parallel process. *Science* **189**, 226–228. (doi:10.1126/science.189.4198.226)

Marslen-Wilson, W. & Warren, P. 1994 Levels of perceptual representation and process in lexical access. *Psychol. Rev.* **101**, 653–675. (doi:10.1037/0033-295X.101.4.653)

Marslen-Wilson, W. D. & Welsh, A. 1978 Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* **10**, 29–63. (doi:10.1016/0010-0285(78)90018-X)

Marslen-Wilson, W., Levy, E. & Tyler, L. K. 1982 Producing interpretable discourse: the establishment and maintenance of reference. In *Speech, place and action* (eds R. J. Jarvella & W. Klein), pp. 339–378. New York, NY: Wiley.

McClelland, J. L. & Elman, J. L. 1986 The TRACE model of speech perception. *Cogn. Psychol.* **18**, 1–86. (doi:10.1016/0010-0285(86)90015-0)

McNeill, D. (ed.) 2000 *Language and gesture*, Cambridge, UK: Cambridge University Press.

Metzing, C. & Brennan, S. E. 2003 When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* **49**, 201–213. (doi:10.1016/S0749-596X(03)00028-7)

Miller, G. A. 1962 Some psychological studies of grammar. *Am. Psychol.* **17**, 748–762. (doi:10.1037/h0044708)

Miller, G. A. & Chomsky, N. 1963 Finitary models of language users. In *Handbook of mathematical psychology* (eds R. D. Luce, R. R. Bush & E. Galanter), pp. 421–491. New York, NY: Wiley.

Nadig, A. & Sedivy, J. 2002 Evidence for perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* **13**, 329–336. (doi:10.1111/j.0956-7976.2002.00460.x)

Norris, D., McQueen, J. M. & Cutler, A. 2002 Bias effects in facilitatory phonological priming. *Mem. Cogn.* **30**, 399–411.

Pickering, M. J. & Garrod, S. C. 2004 Towards a mechanistic theory of dialog. *Behav. Brain Sci.* 7, 169–190.

Pulvermüller, F., Härle, M. & Hummel, F. 2001 Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain Lang.* **78**, 143–168. (doi:10.1006/brln.2000.2390)

Rayner, K. 1998 Eye movements in reading and information processing: twenty years of research. *Psychol. Bull.* **124**, 372–422. (doi:10.1037/0033-2909.124.3.372)

Roberts, C. 2003 Uniqueness in definite noun phrases. *Linguist. Philos.* **26**, 287–350. (doi:10.1023/A:1024157132393)

Sabbagh, M. A. & Baldwin, D. A. 2001 Learning words from knowledgeable versus ignorant speakers: links between preschoolers' theory of mind and semantic development. *Child Dev.* **72**, 1054–1070. (doi:10.1111/1467-8624.00334)

Salverda, A. P. & Altmann, G. T. M. 2005 Cross-talk between language and vision: interference of visually-cued eye movements by spoken language. Poster presented at the *Architectures and Mechanisms in Language Processing (AMLaP) Conf.*, September 2005, Gent.

Salverda, A. P., Dahan, D. & McQueen, J. M. 2003 The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* **90**, 51–89. (doi:10.1016/S0010-0277(03)00139-2)

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M. & McDonough, J. M. 2007 Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition.* **105**, 466–476. (doi:10.1016/j.cognition.2006.10.008)

Schegloff, E. A. & Sacks, H. 1973 Opening up closings. *Semiotica* **8**, 289–327.

Searle, J. R. 1969 *Speech acts. An essay in the philosophy of language*. Cambridge, UK: Cambridge University Press.

Sedivy, J. C. 2003 Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguis. Res.* **32**, 3–23. (doi:10.1023/A:1021928914454)

Spivey, M. J. & Dale, R. 2004 On the continuity of mind: toward a dynamical account of cognition. In *The psychology of learning and motivation*, vol. 45 (ed. B. Ross), pp. 87–142. Amsterdam, The Netherlands: Elsevier.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M. & Sedivy, J. C. 2002 Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. *Cogn. Psychol.* **45**, 447–481. (doi:10.1016/S0010-0285(02)00503-0)

Spivey, M. J., Grosjean, M. & Knoblich, G. 2005 Continuous attraction toward phonological competitors. *Proc. Natl Acad. Sci. USA* **102**, 10 393–10 398. (doi:10.1073/pnas.0503903102)

Stalnaker, R. C. 1978 Assertion. In *Syntax and semantics: pragmatics*, vol. 9 (ed. P. Cole), pp. 315–332. New York, NY: Academic Press.

Stalnaker, R. C. 2002 Common ground. *Linguist. Philos.* **25**, 701–721. (doi:10.1023/A:1020867916902)

Tanenhaus, M. K. 2004 On-line sentence processing: past, present and, future. In *On-line sentence processing: ERPS, eye movements and beyond* (eds M. Carreiras & C. Clifton), pp. 371–392. London, UK: Psychology Press.

Tanenhaus, M. K. & Trueswell, J. C. 1995 Sentence comprehension. In *Speech, language, and communication* (eds J. Miller & P. Eimas), pp. 217–262. San Diego, CA: Academic Press.

Tanenhaus, M. K., Carlson, G. & Seidenberg, M. S. 1985 Do listeners compute linguistic representations? In *Natural language parsing: psychological, computational, and theoretical perspectives* (eds D. Dowty, L. Kartunnen & A. Zwicky), pp. 359–408. Cambridge, UK: Cambridge University Press.

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. (doi:10.1126/science.7777863)

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. (doi:10.1016/S0010-0277(99)00032-3)