Language Embedded in the Environment

Michael J. Spivey
Dept. of Psychology
Cornell University
Ithaca, NY 14853
spivey @cornell.edu

Daniel C. Richardson
Dept. of Psychology
Stanford University
Stanford, CA 94305
richardson@psych.stanford.edu

"A language fractionated from all its referents is perhaps *something*, but whatever it is, it is neither a language nor a model of a language."

- Robert Rosen (2000)

Introduction

While driving in England recently, the second author was listening to a play on the radio. A Roman senator was telling another how Cicero had insulted him in court. In moments like those, it seems that language could not be *less* 'situated in the world'. The senators were in Rome, not the English countryside, the events took place in the distant past; even within the play, the senator listening had not been in the courtroom nor had he ever met Cicero. Despite these many degrees of separation between the words of the Roman senator and the world of the radio listener, the play was perfectly understandable. Indeed, at one further degree of separation, you as a reader are perfectly able to understand what happened in the car.

The ability of language to convey information about objects, people and ideas that have no immediate physical presence, or have never even existed before, would seem to mark it out as a prime example of a 'representationally hungry' (A. Clark, 1997) cognitive task that is *not* situated in the world. Even if the mental representations underlying language are grounded in past perceptual or motor experiences of the world (Barsalou, 1999; Glenberg, 1997; Zwaan & Kaschak, this volume), this does not require

that when it is spoken, an utterance has any deep connection to the environment in which it is heard, as the case of the radio play clearly demonstrates.

Such unembedded communication, about objects and events that are not in the here-and-now, is obviously possible, but it is not always successful. After all, how many times have you heard someone try to tell a funny story, and end up saying, "I guess you had to be there." *Being there* is exactly what embedded language is about, and it is arguably our primary form of language use. It is somewhat ironic, of course, that we are using the medium of *unidirectional scripted language* to make our arguments, but.in this chapter we will argue that, for the bulk of everyday language use by most people, language processing is inextricably embedded in the world. Production and comprehension display many of the hallmarks of situated cognition discussed in this volume, and can often be understood best in terms of agents acting within a particular environment.

The situated and non-situated aspects of language map onto a distinction made by H. Clark (1992) between two separate traditions within language research. The 'language as product' tradition has lauded the non-situated virtues of language: spoken language is an evanescent auditory signal, yet it can be used to refer to objects people and events that are not present. This tradition, which has dominated psycholinguistics, studies the processes by which listeners form internal linguistic representations. Experimentally, this tradition seeks to decontextualize linguistic stimuli where possible in order to isolate processes and representations. In contrast, the 'language as action' tradition has characterized language use as a form of joint action that is embedded in the world. This tradition stems from school of 'ordinary language' philosophy (Grice, 1989; Austin,

1962; Searle, 1969). Experimentally, this tradition has observed and analyzed the conversations of individuals in real word social interactions and tasks.

In the last decade, attempts have been made to bridge these two traditions (cf. Trueswell & Tanenhaus, 2005). Following early work by Cooper (1974), eye tracking technology has been used during task-orientated language use in rich contexts, following the language as action tradition, and has provided a fine grained measurement of linguistic processing in the language as product tradition (e.g.,Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995). We use the term 'embedded language' to refer to this bridge between the language as product and language as action traditions. In this chapter, we will present evidence that the external world is involved in a wide range of linguistic processes, from parsing syntax to encoding facts, to understanding figurative phrases. In recent work, conversation appears as 'jointly situated cognition', in which conversants coordinate each others' visual attention around the external world. Our review of the literature suggests that fine-grained behavioural measures such as eye-tracking convincingly demonstrate language use to be a prime example of situated cognition.

Embedded language processing involves complex situational variables imposing immediate influences on word recognition, syntactic parsing, and discourse comprehension. This entails that language processing is an interdisciplinary activity in the brain, involving visual perception, auditory perception, motor processing, and reasoning, in addition to linguistic computations. Importantly, however, the purview of embedded language processing is not only coextensive with other internal cognitive processes, it extends beyond the language user's brain. That is, the very cognitive

operations involved in embedded language processing are themselves coextensive among speaker, listener, and environment. We shall lay out these arguments at the computational level before offering empirical evidence at the behavioural level.

Computational arguments

Language has been held up as the best case in the argument for cognitive modules that are isolated from each other and from external variables (Fodor, 1983). In contrast, situated cognition is about the inextricability of cognition from key situational variables in the environment. A particular mathematical distinction becomes especially important when discussing the separability of cognition from the environment, as well as the separability of individual cognitive processes from one another. Cognitive psychology's traditional information-processing approach (cf. Neisser, 1967; Sternberg, 1969) relies heavily on an assumption of *component-dominant dynamics*, whereas a more continuous and interactive situated cognition approach (cf. Greeno, 1998; Neisser, 1976) steers the mathematical description more toward *interaction-dominant dynamics*. In component-dominant dynamics, most of the important functional processes are carried out inside each subsystem (e.g., inside a language subsystem, inside a vision subsystem, and inside a memory subsystem). And when those subsystems share information with one another, the transmission is extremely limited and constrained. The transmission process itself does not contribute significantly to the functional computational result. In contrast, interaction-dominant dynamics are when most of the important functional processes are carried out via the interactions *between* subsystems, not inside each of them. That is, each subsystem's behavior is substantially described by parameters external to it, and

only partly described by its own internal state-transition parameters. In such a circumstance, the openness of these subsystems requires that the level of analysis be "zoomed-out a notch," so that the larger system being described (which comprises those subsystems) is a bit more closed (not relying so much on parameters external to it), and therefore more scientifically analyzable. If internal mental processing is a sufficiently open system that proper analysis requires describing the larger system in which internal mental processing is embedded, then the appropriate level of analysis is the organism-environment dyad, not just the organism (Gibson, 1979; Turvey & Shaw, 1999).

When a system exhibits interaction-dominant dynamics, the traditional feed-forward linear-systems analysis borrowed from electrical engineering will no longer suffice. That is, the popular "divide and conquer" scientific paradigm endorsed by the modularity framework is no longer valid for such a system. Instead, what is needed is a dynamical-systems framework that applies continuous mathematical descriptions to the functional interactions that emerge between subsystems (e.g., Kelso, 1995; Port & Van Gelder, 1995; Van Orden, Holden, & Turvey, 2003; Ward, 2002). The dynamical-systems framework, which is becoming increasingly popular in ecological psychology (Turvey & Carello, 1995; Turvey & Shaw, 1999), has the necessary explanatory tools for dealing with an account of cognition that is not limited to the individual organism's brain, but can also include situational variables as part and parcel to cognition.

In what follows, we will review a variety of real-time experimental evidence from psycholinguistics that is gradually accruing support for rich interaction between cognitive subsystems as well as rich interaction between the biological substrate and the environmental substrate. These findings indicate that language processes rely heavily on

situational variables at a very fine time scale, and therefore exhibit the kind of interaction-dominant dynamics that is more consistent with the developing situated cognition approach to psychology than with the traditional information-processing approach.

Spoken Language Processing Embedded in a Situational Context

Put simply, language is not processed by "language processors," it is processed by people. Much like J. J. Gibson's (1979) ecological psychology, where the level of analysis is the organism-*cum*-environment, a proper understanding of how language processing works requires attention to the perceptual-motor patterns that interface the organism with its environment (cf. Zwaan & Kaschak, this volume) and those environmental properties in which the language use is embedded. Without the many different environmental constraints that situate language processing -- too often lumped together under the monolithic term "context" -- language as we know it would not exist.

Eye movements are a particularly compelling example of how component-dominant dynamics does not provide a good description of how the brain and body interface with their environment, and that interaction-dominant dynamics may provide a more appropriate description. Considerable evidence (e.g., Gold & Shadlen, 2000; Spivey & Dale, 2004; van der Heijden, 1996) supports the observation that the brain does not achieve a stable percept, then make an eye movement, then achieve another stable percept, then make another eye movement, etc. The eyes often move *during the process* of achieving a stable percept. As a result, before perception can finish settling into a unitary interpretation of its input, oculomotor output changes the perceptual input by

placing new and different visual information on the foveas. For example, an initial eye fixation may *cause* certain dynamical perceptual processes to be set in motion, which then (before they become stable) *cause* a new eye movement, which then allows different environmental properties to *cause* different dynamics in the perceptual process, which then *cause* yet another eye movement, etc. Thus, perception is simultaneously influenced by sensory input (caused by environment properties reflecting light onto the retinas) and by oculomotor output (caused by intermediate products of perception's own analog computations). Since eye movements tend to operate at a slightly finer timescale than does perceptual stabilization, the perception-action cycle in this case becomes a circular causal loop -- for which "distinguishing the chicken from the egg" becomes moot. This kind of loop is called *impredicative* (Poincaré, 1906; Russell, 1906) because it is composed of elements that can only be defined with reference to the larger system of which they are members (Rosen, 2000; Turvey, 2004). With impredicative systems, such as a continuously flowing perception-action loop, there can be no context-independent definition of each computational element, followed by a linear, feedforward, component-wise integration of those elements. In this loop, perceptual/cognitive parameters, motor parameters, and environmental parameters are so causally intertwined with one another that the system can only be adequately studied as a whole: a mind situated in its environment.

As the time scale of eye movements brings into sharp relief this insight into the real-time situated character of cognition, it is natural that we first report on evidence for embedded language processing that comes from eye-tracking experiments. For a decade now, eye-tracking studies have been demonstrating the wide variety of ways in which the

visual world is continuously accessed and integrated with spoken language processing, at the level of word recognition (Allopenna, Magnuson, & Tanenhaus, 1998), reference resolution (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995), syntactic processing (Tanenhaus et al. , 1995), and thematic role assignment (Altmann & Kamide, 1999).

For example, even during the few hundred milliseconds it takes to hear a spoken word, situational constraints provided by the visual context can influence the activations of lexical representations that result from processing those first few phonemes. When sitting in front of a table with real objects (see Figure 1), and instructed to "Pick up the candy," about 1/3 of the time participants will initially fixate the *candle* for a couple hundred milliseconds, before then fixating the bag of candy to pick it up (Spivey-Knowlton, Eberhard, Sedivy, & Tanenhaus, 1998). They do this because that object's name shares several initial phonemes with "candy," i.e., it is a *cohort* of the word (Marslen-Wilson, 1987). In a control condition, if you deliver the same spoken instruction with a visual display that does not contain a candle, participants quickly land their eyes on the candy, and rarely fixate the other objects. (Similar results are found in the graded curvature of computer-mouse movements, with instructions like "Click the candle," cf. Spivey, Grosjean, & Knoblich, 2005.) If you ask them whether they noticed having looked at the candle, they will deny having done so. Thus, in the right visual environments, eye movements can be used as an early strategy-proof measure of partially active representations that can be sampled without interrupting normal task performance.

Insert Figure 1 about here

To further strengthen this linking hypothesis between multiple lexical activations and the probability of fixating various objects, Allopenna et al., (1998) replicated the cohort effect above, extended it to rhymes (looking at a *handle* when instructed to "pick up the candle"), and demonstrated a close correspondence between the eye-movement data and the lexical activations of McClelland and Elman's (1986) TRACE model of spoken word recognition (see alsoDahan, Magnuson, & Tanenhaus, 2001; Spivey et al. , 2005). Similar eye-movement cohort effects have been observed in French (Dahan, Swingley, Tanenhaus, & Magnuson, 2000), in Russian (Marian & Spivey, 2003), and even across two languages (Ju & Luce, 2004; Spivey & Marian, 1999).

Notably, it is not just the mere visual co-presence of an object during spoken word recognition that influences the comprehension process, but more specifically how *actionable* the situational constraints allow that co-present object to be. For example, Chambers, Tanenhaus, Eberhard, Filip, and Carlson (2002) gave participants instructions like "Put the cube inside the can," when there were two cans in the display. They found that situation-specific affordances (of one of the cans in the display being large enough to contain the cube and the other can being too small to contain the cube) immediately constrained the referential domain of "the can," such that participants looked only at the appropriately actionable object -- despite the glaring referential ambiguity in the speech signal. Thus, speech input alone does not determine spoken word recognition. The situational context, constrained by what relevant objects are visible and actionable, plays an immediate role in driving the processes that map phonemes onto lexical representations and lexical representations onto referential operations.

In fact, the temporal continuity with which the speech stream is integrated with visual/situational constraints can be systematically illustrated by tracking the comprehension of complex clause structures. Eberhard et al. (1995) gave participants instructions like "Put the queen of hearts that's below the jack of diamonds above the king of spades." In Figure 2, a schematic example of a scanpath shows the eye position starting at the central cross and jumping to the queen of clubs soon after hearing "queen." After hearing "of hearts," the eyes saccade down to the distractor queen of hearts and flit around there until hearing "that's below the jack of diamonds" causes them to fixate the target queen of hearts. After a quick check of that jack of diamonds, and once "king of spades" is heard, the eyes finally move up to the upper left corner of the display in order to beginning planning the manual action. Thus, the eyes are linking referring expressions from the spoken linguistic input to their potential referents in the world about as quickly as they arrive.

Insert Figure 2 about here

Interestingly, when such prepositional-phrase structures contain syntactic ambiguities, the visual/situational context can be used immediately to resolve them. For example, if participants are presented with a display of real objects like that in Figure 3A, and instructed to "Put the spoon on the napkin in the bowl," they first look at the spoon and then frequently fixate the upper right napkin, before finally fixating the bowl and carrying out the proper action (Tanenhaus et al., 1995; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002). This suggests that participants are often initially parsing "on the napkin"

as syntactically attached to the verb phrase, i.e., temporarily treating it as the goal for the spoon's "putting" event. The brief fixation of the irrelevant napkin is evidence of this incorrect parse because it does not happen in the control condition, where the spoken instruction is syntactically unambiguous (e.g., "Put the spoon that's on the napkin in the bowl."). The critical demonstration that visual/situational context can intervene during real-time syntactic processing comes from the condition where the same syntactically ambiguous instruction was delivered with the visual display in Figure 3B. In this context, participants hardly ever looked at the irrelevant goal (e.g., the other napkin). Since there are two spoons in this context, upon hearing "the spoon," the participant does not yet know which spoon is being referred to, and both spoons tend to get fixated briefly. Then, upon hearing "on the napkin," the referential uncertainty introduced by having two potential spoons causes the comprehension process to parse "on the napkin" as syntactically attached to the noun phrase rather than to the verb phrase, thereby discriminating which spoon is being referred to (cf. Crain & Steedman, 1985). Thus, even syntactic processing (once thought to be the "poster child" of encapsulated modularity, cf. Frazier & Clifton, 1996) is fluidly integrated in real-time with the situated context in which language comprehension takes place.

Insert Figure 3 about here

Similar eye-movement evidence for the immediate use of situational context during syntactic parsing has been found with participants as young as 8 years old (Trueswell, Sekerina, & Logrip, 1999). Moreover, Snedeker and Trueswell (2004)

showed that these situational constraints combine with the verb's frequency-based biases in argument structure (e.g., how strongly it prefers a prepositional phrase) to determine the likelihood of a mis-parse. Most importantly, for demonstrating the importance of situated context in language processing, the relative affordances of the objects being moved around play a crucial role in defining the situated constraints in which comprehension takes place. Chambers, Tanenhaus, and Magnuson (2004) gave participants instructions like "Pour the egg in the bowl over the flour." When the second of the two referents (an egg in a glass) was compatible with the action (i.e. it was in liquid form and thus could be poured), the context functioned like the previous two-referent contexts and prevented the mis-parse associated with fixating the incorrect goal (an extra bowl that was empty). However, when that second egg was in shell form (and thus not able to be poured), participants immediately knew which egg was being referred to upon hearing "Pour the egg," and thereby fell into the trap of parsing "in the bowl" as a goal rather than a noun-phrase modifier.

Thus, as suggested by McRae, Ferretti, and Amyote (1997), individual verbs, such as *pour*, may activate rather complex sets of semantic features regarding the thematic roles (e.g., Agents, Patients, Themes, Goals, etc.) that are likely to participate in the event being described. For example, Altmann and Kamide (1999) presented participants with computer-screen displays containing pictures of a man, a woman, a newspaper, and a cake. As evidence of anticipations from a verb's thematic role preferences, when the spoken instruction was "The woman will read the newspaper," participants often fixated the newspaper shortly *before* hearing the word "newspaper" -- immediately after hearing the word "read." moreover, consistent with the notion of mental animation (Hegarty &

Just, 1993; Matlock & Richardson, 2004), when Altmann (2004) showed the pictures and then took them away before presenting the sentence, he observed the same basic results, with participants fixating the appropriate blank regions that used to contain the relevant objects. This "blank screen paradigm" illustrates how the external environment alone is not what imposes the situated constraints, but rather it is the time-dependent relationship between the environment and the internal mental representations that creates the situated character of language processing.

Even when two sentences describe identical referents in the world, aspects of *figurative* meaning are detectable in how a visual scene is inspected. When static referents in the visual scene are best understood by mentally animating them, eye movements can provide evidence of that mental animation (cf. Hegarty & Just, 1993; Rozenblit, Spivey, & Wojslawowicz, 2002). Matlock and Richardson (2004) contrasted literal descriptions of scenes (*The road is in the valley*) with descriptions using fictive motion (*The road runs through the valley*). Fictive motion descriptions are a type of figurative language because they employ a verb of motion (*runs*) but no motion takes place. Norming results show that fictive and non-fictive spatial descriptions are equivalent in meaning. Participants heard one of these descriptions while looking at a schematic drawing of a scene. Matlock and Richardson (2004) found that participants spent more time fixating the path when it was described using fictive motion.

In that experiment, it is as if the eyes are *acting out* the subtle figurative spatial dynamics implied by the road "running" (cf. Matlock, 2004), and Richardson and Matlock (2005) demonstrated that such fictive motion descriptions affect eye movements specifically by evoking mental representations of motion. When participants heard

information about terrain that would affect actual motion across the scene, it influenced how they viewed a picture if it was described with fictive motion. Looking times and eye movements scanning along the path increased during fictive motion descriptions when the terrain was first described as difficult (*The desert is hilly*) as compared to easy (*The desert is flat*); there were no such effects for descriptions without fictive motion. It appears that some form of dynamic visuospatial simulation is generated when comprehending fictive motion descriptions (Barsalou, 1999; Zwaan, 2004), and this simulation even incorporates information about terrain maneuverability.

A typical information processing account might suppose that a linguistic figurative description is processed into a set of spatial relations, and if necessary, the result is later passed on for comparison with the output of a visual process. At odds with this view is the fact that traces of non-literal content can be found in the earliest moments of visuomotor processing. Comprehension of figurative language can produce mental representations that are distinct from equivalent literal counterparts. These representations are immediately integrated with visual processing, such that various forms of dynamic spatial information can drive eye movements around a scene. In this way, even comprehension of nuances of figurative meaning appears to be embedded in the world.

Semantic memory

We have seen that during language processing, the external world is interrogated continuously throughout the course of incremental linguistic input. This embedded nature can be seen at multiple time scales of language comprehension, even in the way that semantic information is encoded and recalled. In a series of experiments, Richardson

and colleagues (Richardson & Spivey 2000; Richardson & Kirkham, 2004) found that when listening to spoken information, relevant spatial locations are encoded and accessed with a saccade when relevant. These experiments can be seen as a linguistic case of situated cognition, related to 'epistemic actions' (Kirsh & Maglio, 1994) and 'deictic pointers' (Ballard, Hayhoe, Pook, & Rao, 1997).

Figure 4 shows a schematic of Richardson and Spivey's (2000) first experiment. Participants watched a video clip of a talking head delivering a short piece of information such as, "Shakespeare's first plays were romantic comedies. His last was the Tempest." These talking heads appeared in turn in each of four ports of a 2 x 2 grid. After presentation, the participants looked at a blank grid and heard a statement that related to one of the facts (e.g., "Shakespeare's last play was the Tempest"). They answered out loud whether it was true or false. As they answered, participants' eye movements were recorded. The port that had previously contained the talking head that conveyed the relevant information was termed the 'critical port'. It was found that there were almost twice as many fixations to the critical port than of each of the other ports. This result was replicated when the video clips were replaced by four identical spinning crosses (Spivey & Richardson, 2000, Experiment 2), when the ports moved to the center of the screen during presentation (Spivey & Richardson, 2000, Experiment 5), and when the ports moved independently to different locations on the screen before the answer period (Richardson & Kirkham, 2004). Moreover, in all experiments, participants' accuracy in answering the factual question was not related to whether or not they looked at the critical port. Why then do participants continue to 'spatially index' information - encoding location of an event and re-fixating that location when recalling its properties?

Insert Figure 4 about here

A first attempt at explaining this spatial indexing behavior might cite the phenomena of context-dependent memory, in which memory is improved if the conditions that were present during encoding are re-instantiated during recall (e.g. Bradley, Cuthbert, & Lang, 1988; Godden & Baddeley, 1975; Winograd & Church, 1988). This explanation falls short on two counts, however. Memory was not improved in this paradigm, since looks to the critical port did not produce more correct answers. In several of the experiments, the conditions that were present during encoding were not re-instantiated during recall, since the ports were in different locations between presentation and test phases.

The more fitting explanation draws on the notion of 'external memory' (Brooks, 1991; A. Clark, 1997; O'Regan, 1992). In everyday life (Simons & Levin, 1997) and carefully circumscribed experimental tasks (Ballard, Hayhoe, & Pelz, 1995; Hayhoe, Bensinger & Ballard 1998), it appears that participants do not encode many properties of the visual world. Instead, as and when information is needed, it is accessed from external memory via an eye movement. As every system of information storage needs a system of information retrieval, perhaps spatial indexing stores the "addresses" for a content-addressable memory that exists in the external environment, rather than in the brain.

It could be argued that participants in Richardson and Spivey (2000) and in Richardson and Kirkham (2004) were tacitly behaving as if the factual information they had heard could be accessed from external memory. When a fact was heard, participants

associated the information with a port on the computer screen. As the information was needed during the question period, the association was activated and a saccade was launched to retrieve that information. Of course, in this case, there was no useful information there at all, and so accuracy in answering the question did not increase with fixations to the empty critical port.

This interpretation suggests that, in terms of their looking behavior, participants in Richardson and Spivey (2000) and in Richardson and Kirkham (2004) were treating pieces of evanescent, auditory, semantic information as if they were stable physical objects in the world, there to be re-inspected whenever the need arose. Richardson and Kirkham (2005) found that despite the fragility of infants' spatial abilities around that age (Columbo, 2001; Gilmore & Johnson, 1997), infants as young as six months of age showed the same spatial indexing behaviour: they encoded the location of a toy that danced to a tune inside a port, tracked the location of the port as it moved, and re-fixated the port when the tune was heard again. These results this suggests that the embedded nature of linguistic semantic memory has roots in early development (see also Smith, this volume).

Conversation

A stream of speech is processed by a listener, moment-by-moment, in reference to the external world. In this sense, language comprehension is embedded. But of course, in the main, language use is an interaction between people. What does the embedded nature of language entail for the interplay of comprehension and production that occurs during naturalistic conversation? The 'language as action' tradition has long seen conversation as a form of 'joint action' that is situated in the world (H. Clark, 1996). And recent

technological advances have allowed the time-course of conversation and linguistic processing to be studied in fine detail, and thus formed a bridge to the 'language as product' tradition (Trueswell & Tanenhaus, 2004).

In this section, we draw three broad conclusions from the current literature that speak to the embedded nature of conversation. Firstly, participants in a conversation are aware of how each other is cognitively situated in the world. H. Clark (1996) describes this as knowledge of the 'common ground', and recent evidence supports his notion that linguistic processing is intimately affected by assumptions about an interlocutor's visual perspective, past experience and beliefs. Secondly, interlocutors actively manipulate the common ground for communicative purposes. Just as an individual may alter her environment during a situated cognitive task (Kirsch & Maglio, 1994; Hutchins, 1995), or two participants may anticipate one another's task constraints during joint action (cf. Knoblich & Jordan, 2003; Sebanz, Knoblich, & Prinz, 2003), so two conversants will use actions and gestures, concurrently with their speech, to coordinate their interaction and update the common ground. Lastly, the degree to which interlocutors are able to coordinate their visual attention, moment-by-moment across a shared visual display is causally related to the success with which they communicate.

Listeners are certainly sensitive to some facts about a speaker. For example, Fitneva and Spivey (2004) showed that knowledge of the author of a spoken statement influences lexical ambiguity resolution. Though the word "case" is ambiguous between a court case and a container case, when a judge says the word, listeners immediately resolve it as meaning court case and when a store owner says it, listeners immediately resolve it as a meaning a container case.

Metzing and Brennan (2003) showed that the identity of speakers is also important when parsing novel forms of reference known as 'conceptual pacts' (Brennan & H. Clark, 1996). In their task, a participant and a confederate repeatedly referred to a novel object as, for example, the 'shiny cylinder'. This is an example of 'lexical entrainment' (Garrod & Anderson, 1987). At a later stage the object was referred to again, either by the old or a new confederate, or by the old or a new name (e.g., the 'silver pipe'). While participants fixated the correct object equally fast regardless of which confederate used the old name, participants were relatively slow when the old confederate used the new name. This momentary confusion shows that speaker identity is linked to particular conceptual pacts, and that listeners expect terms within common ground to be reliably used.

Initial results from other collaborative tasks, however, suggested that listeners do not extend very far into a consideration of the speaker's mental states (Keysar, Barr, Balin, & Paek, 1998). Keysar, Barr, Balin & Brauner (2000) used a referential communication task (Krauss & Weinheimer, 1964) in which two people are sat on either side of an array of pigeon holes containing various objects and the director (a confederate) instructs the matcher to move objects. Some of the array is blocked so that the director cannot see some of the objects. Keysar et al.(2000) found that when the matchers hear 'Pick up the smallest candle' they fixate (and sometime pick up) the smallest candle they can see, even though that candle is occluded from the directors' view, and hence could not possibly be the intended referent. It was argued that during such a conversation, the matcher initially takes an egocentric interpretation of the director's instruction, and can only take into consideration the director's knowledge state

at the end of the process, as an error correction.

Later studies (Hanna, Tanenhaus, & Trueswell, 2003) argue that the Keysar et al. (2000) result misses the pervasive effect of common ground information by swamping it with typicality effects. In short, when the matchers hear ”the smallest candle,” the object that is occluded from the speaker’s perspective happens to be the most typical referent of the statement, and hence it attracts a higher proportion of fixations than a completely unrelated item. Using a similar design, Hanna et al. (2003) deconfounded these variables. In the key condition, the director instructs the matcher to “put a triangle on top of the red one,” when two red triangles are in view. One of these red triangles is not known to the speaker, however, and so could not be the intended referent. In this case, when the target and the competitor are identical, matchers make a higher proportion of fixations to the correct target from the very moment they hear the word ‘red’. Therefore, rather than acting solely as a late source of error correction, common ground information acts as a constraint (among many) on reference resolution at the earliest stages of linguistic processing (Hanna & Tanenhaus, 2004; see also Nadig & Sedivy, 2002).

In the matching studies discussed so far, common ground was investigating by manipulating whether or not a director had knowledge of the physical presence of a particular item. Conversants appear to keep track of changes in common ground brought about by linguistic manipulations as well. In an experiment by Brown-Schmidt, Campana and Tanenhaus (2004), a speaker instructed a listener to move various blocks on a grid. Both were naïve participants. Sometimes during this task, the speaker referred to 'the red one' even though there were several red blocks in sight. The listener was able to fixate the correct block, however, because what the speaker had said previously had

implicitly identified a smaller set of objects that included only one of the red blocks. In this way, linguistic context can 'circumscribe the referential domain'.

Our second claim regarding the embedded nature of conversation is that participants are not only aware of how each other has situated their dialog in the world, but that they actively manipulate that common ground as a coordinated 'joint-activity' that utilizes pointing, placing and gestures (H. Clark, 1996, 2003; H. Clark & Brennan, 1991; Schober, 1993).). For example, Bangerter (2004) examined how participants discussed pictures when they could or could not point. At long distances, where they would be ambiguous, there were fewer pointing gestures. At closer ranges, pointing increased and linguistic description of location decreased. In this way, pointing gestures were used opportunistically as part of a composite signal with speech.

H. Clark and Krych (2004) analyzed other forms of physical action that are employed to manipulate common ground. In their task, a director participant instructed a builder participant how to construct a Lego model. When the director could not see what the builder was doing, performance suffered. When the director was able to see the builder at work, this visual common ground was exploited and continually updated as a joint activity. For example, while the director was describing the next block to pick up, the builder might find it in a pile, exhibit it to the director, who would interrupt or alter his description mid-sentence to confirm whether it was the correct piece. Similarly, the builder would poise a block just above where she believed it should be attached to the model, or turn the model to face the director, so that the director simply had to acknowledge whether the move was correct. These gestures were precisely timed and coordinated with the director's speech. When a visual common ground is available,

participants situate their dialog by engaging in complex joint activities that support and even supplant verbal communication (see also Brennan, 2004).

In a classic case of situated cognition, Kirsch and Maglio (1994) found that expert Tetris players would rotate shapes using a button press, as it was a faster way to view different orientations than using mental rotation. They termed this button press an 'epistemic action' because participants take action in the world to manipulate knowledge, rather than using purely mental operations. The external world was acted upon in order to process information. Similarly, in situated communication, conversants take joint-action in the world, rather than using purely linguistic communication. In this way, the physical gestures and actions that conversants use can be thought of as 'joint epistemic actions'.

But is it the case that these situated strategies and dynamic coordination of common ground are actually more efficient than unembedded one-way communication (as in reading or listening to the radio)? Research quantifying the degree of coordination between conversants' visual attention suggests that this is true. Richardson and Dale (in press) analyzed the statistical patterns of eye movements as a fine-grained index of how speakers and listeners deployed their attention within a visual common ground. Rather then studying speakers and listeners separately, asking them to produce or comprehend short sentences, they eye-tracked both speakers and listeners who were engaged in a spontaneous, complex discourse. They quantified the temporal coupling (or entrainment) between conversants' eye movements, and examined its relationship to the success of the discourse.

First, the speech and eye movements of one set of participants were recorded as they looked at pictures of six cast members of a TV sitcom (either 'Friends' or 'The

Simpsons'). They spoke spontaneously about their favourite episode and characters, or described what had happened in a scene they had just watched. One-minute segments were chosen and used unedited, with all the deviations hesitations and repetitions of just a minute of normal speech. These segments were then played back to a separate set of participants. The listeners looked at the same visual display of the cast members, and their eye movements were recorded as they listened to the segments of speech. They then answered a series of comprehension questions about what was said by the speaker.

Listener and speaker eye movements were coded as to which of the six cast members (if any) was being fixated during every 33ms time slice. Cross-recurrence analysis (Zbilut, Giuliani, & Webber, 1998; Zbilut & Webber, 1992) was used to quantify to degree to which the speaker and listener eye positions overlapped at successive time lags. This speaker X listener distribution of fixations could be compared to a -speaker X randomized-listener distribution, that was produced by shuffling the temporal order of each listener's eye movement sequence and then calculating the cross recurrence with the speakers. This randomized series serves as a baseline of looking 'at chance' at any given point in time, but with the same overall distribution of looks to each picture as the real listeners.

From the moment a speaker looks at a picture, and for the following six seconds, a listener was more likely than chance to be looking at that same picture. The breadth of this timeframe suggests that speakers and listeners may keep track of a subset of the depicted people who are relevant moment-by-moment, just as listeners in Brown-Schmidt et al. (2004) were able to linguistically circumscribe the referential domain. Richardson and Dale (in press) found that the overlap between speaker and listener eye movements

peaked at about 2000ms. In other words, two seconds after the speaker looked at a cast member, the listener was most likely to be looking at the same cast member. The timing of this peak roughly corresponds to results in the speech production and comprehension literatures. Speakers will fixate objects 800-1000ms (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998) before naming them, and listeners will typically take 500-1000ms to fixate an object from the word onset (Allopenna et al., 1998).These figures include cases of proper noun production and comprehension, but this peak in overlap between speaker and listener eye movements was still present in the data when all names of the cast members, and associated speaker fixations, were removed from the analysis. (Of course, no such peak was observed in any of the speaker X randomized-listener analyses.) The coupling between speaker and listener eye movements was pervasive, suggesting that planning diverse types of speech will influence the speaker's eye movements, and a few seconds later, hearing them will influence the listener's eye movements.

Insert Figure 5 about here

Importantly, this entrainment of eye-movement patterns between speaker and listener was not merely an epiphenomenal by-product of the conversation process. It played a functional role in comprehension. When the overall proportion of cross-recurrence between individual speaker-listener pairs was quantified, the strength of the relationship between speaker and listener eye-movement patterns reliably predicted how many of the comprehension questions the listener answer correctly. This correlation was

supported by a follow-up study that experimentally manipulated the relationship between speaker and listener eye movements. When a low-level perceptual cue made the eye movements of a listener more or less like the speaker's, the listener's performance on comprehension questions was correspondingly affected.

Despite the fact that conversants could not interact with each other in Richardson and Dale's (in press) experiments, their visual attention was coupled at the millisecond resolution of eye movements. Moreover, this coupling determined listeners' comprehension performance. Thus, cross-recurrence analysis shows that looking around the common ground in step with one another appears to drive the process of conversants' mutual understanding, and so provides a quantitative data visualization of the notion of embedded language and situated discourse.

Conclusion

Perhaps the case of a BBC radio play about Roman senators is actually a misleading place to start when analyzing language. Perhaps cognitive psychology's long-standing preoccupation with circumstances of unidirectional unembedded language use, such as reading experiments and listening to de-contextualized pre-recorded oration, was a misleading way to begin developing theories about language (cf. Spivey et al., 2002). The rarified ability to represent things that are not present or ideas not thought before, in the absence of immediate environmental support, might be better understood as the special peculiarity of language, not the core of its everyday function. For example, across developmental time, children take much longer to understand references to objects when they are absent, gaining competence in comprehension and then production

throughout the second year of life (Huttenlocher & Smiley, 1987; Saylor, 2004; Swingley & Fernald, 2002). Interestingly, caregivers usually anchor their references to absent objects via objects that are present, asking, for example, 'Where's Daddy?' and gesturing to his briefcase (Huttenlocher, 1974; Saylor & Baldwin, in press). Similarly, one could speculate that across evolutionary time, language might well have first emerged situated in cooperative activities such as hunting and tool use (Barsalou, 1999; Corballis, 1992), rather than as a way to refer to abstract concepts and objects that were not present.

Regardless of the ontogeny and phylogeny of language, it is clear that when adult language use is studied in naturalistic contexts, a rich interaction with the world is revealed. Language processing is not something that happens just in the individual brain of the speaker or of the listener. In many circumstances, the proper analysis of language processing is at the level of the organisms and the environment in which they situated. The phenomena that arise at such a level of analysis clearly exhibit interaction-dominant dynamics, rather than component-dominant dynamics, thus obviating the linear module-based analysis common to the information-processing framework. As part of the linguistic process of recognizing words or parsing syntax, participants make saccades to referents in the world that are phonetically, semantically, or pragmatically appropriate. In conversation, the degree to which a listener follows a speaker's gaze around the world is an indication of their understanding, and in face-to-face conversation, eye movements can serve as linguistic cues. Language comprehension could even be called 'stubbornly situated' cognition, since several paradigms have found that participants will look systematically at entirely blank regions of space (that used to contain the referents) during linguistic processing.

Though they are logically separable, the perspectives of embedded language and embodied language (see Zwaan & Kaschak, this volume) have an interesting synergy. For example, these ideas have been combined in a conversational robot called Ripley (Roy & Mukherjee, 2004; Roy 2005), which represents the meanings of objects by multimodal sensory expectations at certain locations (similar ideas have been developed in theories of perception, Noe 2005; O'Regan & Noe, 2001, O'Regan, this volume). That is to say, for the robot, the meaning of the word *apple* is not merely a LISP-based list of conceptual features, but rather is composed of the sensorimotor experiences that it expects to be situated in the world upon hearing "apple." Perhaps, so with humans, our understanding of language is composed not of amodal logical symbols that are divorced from the real world, but instead of perceptual-motor simulations and of situated actions in the environment and with other language users.

Acknowledgements: We are grateful to Rolf Zwaan and the editors for helpful comments that improved the exposition, and to the two Sams for giving their fathers enough free time to write this chapter. Work on this chapter was supported by NIMH-R01-63961 to MJS.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language. 38(4), 419-439.*

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

Altmann, G. T. M. (2004) Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition. 93*, 79-87.

Austin, J.L. (1962). *How to Do Things with Words.* Cambridge: Harvard University Press.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: its origins and role in development*. Hillsdale, NJ: Lawrence Erlbaum.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*, 66-80.

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723-767.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*(6), 415-419.

Barsalou, L. W. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes, 28(1)*, 61-80.

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1988). Perceptually driven movements as

contextual retrieval cues. *Bulletin of the Psychonomic Society*, 26, 541-553.

Brennan, S. (2004). How conversation is shaped by visual and spoken evidence. In J. Trueswell, & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (95-129). Cambridge, MA: The MIT Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*(6), 1482-1493.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139-159.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2004). Real-time reference resolution by naïve participants during a task-based unscripted conversation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *World-situated language processing: Bridging the language as product and language as action traditions*. Cambridge: MIT Press.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory & Language*, 47(1), 30-49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 30(3), 687-696.

Clark, A. (1997). *Being there: Putting brain, body, and the world together again.* Cambridge: MIT press.

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

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

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243-268). Mahwah, NJ: Lawrence Erlbaum.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory & Language, 50*(1), 62-81.

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, 52., 337-367.

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. *Cognitive Psychology, 6*(1), 84-107.

Corballis, M. C. (1992). On the evolution of language and generativity. *Cognition, 1992*, 197-226.

Crain, S., & Steedman, M. (1985). On not being led up the garden path. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing*. Cambridge, MA: Cambridge University Press.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42, 317-367.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender

and spoken-word recognition in French. *Journal of Memory and Language*, *42*, 465-480.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409-436.

Fitneva, S., & Spivey, M. (2004). Context and language processing: The effect of authorship. In J. Trueswell & M. Tanenhaus (Eds.), *World situated language use: Psycholinguistic, linguistic and computational perspectives on bridging the product and action traditions*. Cambridge, MA: The MIT Press.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge: MIT Press.

Frazier, L. & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 181-218.

Gibson, J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gilmore, R. O., & Johnson, M. H. (1997). Egocentric action in early infancy: spatial frames of reference for saccades. *Psychological Science*, 8, 224-230.

Glenberg, A. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1-55.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66(3), 325-331.

Gold, J., & Shadlen, M. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*, 390-394.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*, 5-26.

Grice, P. (1989) *Studies in the Way of Words*, Cambridge: Harvard University Press

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274-279.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science, 28*(1), 105-115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory & Language, 49*(1), 43-61.

Hayhoe, M., Bensinger, D., & Ballard, D. (1998). Task constraints in visual working memory. *Vision Research, 38*, 125-137.

Hegarty, M. and Just, M.A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, *32*, 717-742.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: The MIT Press.

Huttenlocher, J., & Smiley, P. (1987). Early word meanings: The case of object names. *Cognitive Psychology*, 19(1), 63-89.

Ju, M. & Luce, P. (2004). Falling on sensitive ears: constraints on bilingual lexical activation. *Psychological Science*, *15*, 314-318.

Kelso, J. (1995). *Dynamic patterns: The self-organization of brain and behavior.* Cambridge, MA: The MIT Press.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in

conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32-38.

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language, 39(1)*, 1-20.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science,18*, 513-549.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*, 513-549.

Knoblich, G., & Jordan, J. (2003). Action coordination in groups and individuals: Learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1006-1016.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 113-114.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism: Language and Cognition*, *6*, 97-115.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71-102.

Matlock, T.M & Richardson, D.C. (2004). Do eye movements go with fictive motion? Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.

Richardson, D.C. & Matlock, T.M (2004). The integration of figurative language and static depictions: An eye movement study of fictive motion (under review)

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & Cognition*, 32, 1389-1400.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

McRae, K., Ferretti, T., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*, 137-176.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory & Language, 49*(2), 201-213.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*(2), B25-B33.

Neisser, U. (1967). *Cognitive psychology*. East Norwalk, CT: Appleton-Century-Crofts.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco, California: W. H. Freeman.

Noe, A. (2004). *Action in Perception*. Cambridge: MIT Press

O'Regan, J. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology, 46*, 461-488.

O'Regan, J., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*, 939-1031.

Poincaré, H. (1906). Les mathématiques et la logique. *Revue de Métaphysique et de Morale*, *14*, 294-317. (Translated in W. Ewald, ed., 1996, *From Kant to Hilbert. A*

Source Book in the Foundations of Mathematics, vol. 2, Oxford: Oxford University Press.)

Port, R., & van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: The MIT Press.

Richardson, D., & Dale, R. (in press). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*.

Richardson, D., & Kirkham, N. (2004). Multi-modal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General*, 133, 46-62.

Richardson, D., & Spivey, M. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition*, *76*, 269-295.

Rosen, R. (2000). *Essays on life itself*. New York: Columbia University Press.

Roy, D. (2005) Grounding words in perception and action: computational insights, *Trends in Cognitive Sciences*, 9, 389-396.

Roy, D., & Mukherjee, N. (2004). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, *19*, 227-248.

Rozenblit, L., Spivey, M., & Wojslawowicz, J. (2002). Mechanical reasoning about gear-and-belt systems: Do eye movements predict performance? In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 223-240). Berlin: Springer-Verlag. Russell 1906

Saylor, M. M. (2004). Twelve- and 16-month-old infants recognize properties of

mentioned absent things. *Developmental Science*, 7(5), 599-611.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*(1), 1-24.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. New York, Cambridge University Press

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*, B11-B21.

Simons, D. J., & Levin, D. (1997). Change blindness. *Trends in Cognitive Science*, 1, 261-267.

Snedeker, J. & Trueswell, J. (2004). T he developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*, 238–299.

Spivey, M. & Dale, R. (2004). On the continuity of mind: Toward a dynamical account of cognition. In B. Ross (Ed.), *The Psychology of Learning and Motivation.* Vol. 45. (pp. 87-142) Elsevier.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, *10*, 281-284.Spivey, Tanenhaus, Eberhard, & Sedivy, 2002

Spivey, M., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*, 10393-10398.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., Sedivy, J. (1998). Integration of visuospatial and linguistic information in real-time and real-space. In P. Olivier & K. Gapp (Eds.), *Representation and Processing of Spatial Expressions*.

(pp.201-214). Mahwah, NJ: Erlbaum.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276-315.

Swingley, D., & Fernald, A. (2002). Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory & Language*, 46(1), 39-56.

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*(5217), 1632-1634.

Trueswell, J. & Tanenhaus, M., Eds., (2004). *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. Cambridge, MA: The MIT Press.

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

Turvey, M. T. (2004) Impredicativity, dynamics, and the perception-action divide. In V. K. Jirsa & J. A. S. Kelso (Eds.), *Coordination Dynamics: Issues and Trends. Vol.1 Applied Complex Systems* (pp. 1-20). New York: Springer Verlag.

Turvey, M., & Carello, C. (1995). Some dynamical themes in perception and action. In Port, R., & van Gelder, T., *Mind as motion: Explorations in the dynamics of cognition* (pp. 373-401). Cambridge, MA: The MIT Press.

Turvey, M., & Shaw, R. (1999). Ecological foundations of cognition: I. Symmetry and specificity of animal-environment systems. *Journal of Consciousness Studies*, *6*, 111-123.

van der Heijden, A. (1996). Perception for selection, selection for action, and action for

perception. *Visual Cognition*, *3*, 357-361.

Van Orden, G., Holden, J., & Turvey, M. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*, 331-350.

Ward, L. (2002). *Dynamical cognitive science*. Cambridge, MA: The MIT Press.

Winograd, E., & Church, V. (1988). Role of spatial location in learning face-name associations. *Memory and Cognition*, 16(1), 1-7.

Zbilut, J. P., & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A, 171*, 199-203.

Zbilut, J. P., Giuliani, A., & Webber, C. L., Jr. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters, 246*, 122-128.

Zwaan & Kaschak, this volume

Zwaan, R. A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 35-62). New York: Academic Press.

Figure Captions

Figure 1. When instructed to "Pick up the candy," participants frequently look first at the (similar sounding) *candle,* and then quickly move their eyes to the candy to pick it up. At the timescale of the delivery of individual phonemes, linguistic and visual information is being integrated to drive saccadic eye movements (Tanenhaus et al., 1995), as well as continuous hand movements (Spivey et al., 2005).

Figure 2. While listening to the instruction, "Put the Queen of Hearts that's below the Jack of Diamonds above the King of Spades," participants concurrently make anticipatory fixations of various cards based on partial input. See text for details.

Figure 3. When instructed to "Put the spoon on the napkin in the bowl," participants often mis-parse the syntactic attachment of the initial prepositional phrase when there is only one visible referent for "spoon" (panel A), but not when there are two such referents (panel B).

Figure 4. Design and Results of Richardson and Spivey (2000, Experiment 1).

Figure 5. Average cross-recurrence of eye position at different time lags for 49 speaker-listener pairs (Richardson & Dale, in press). See text for details.

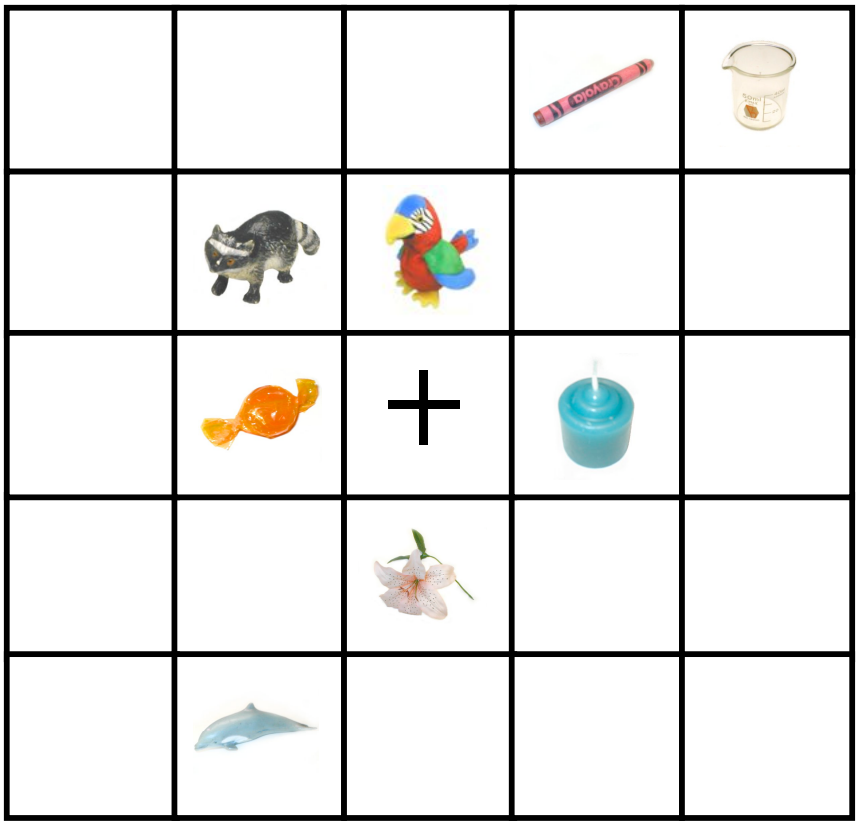

Figure 1

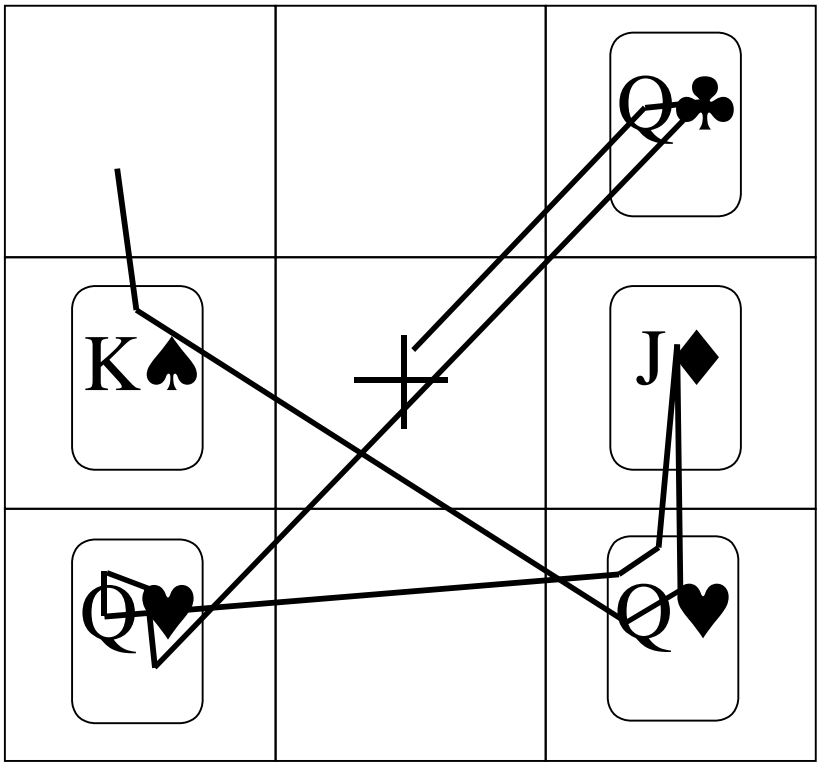

Figure 2

A.

B.

Figure 3

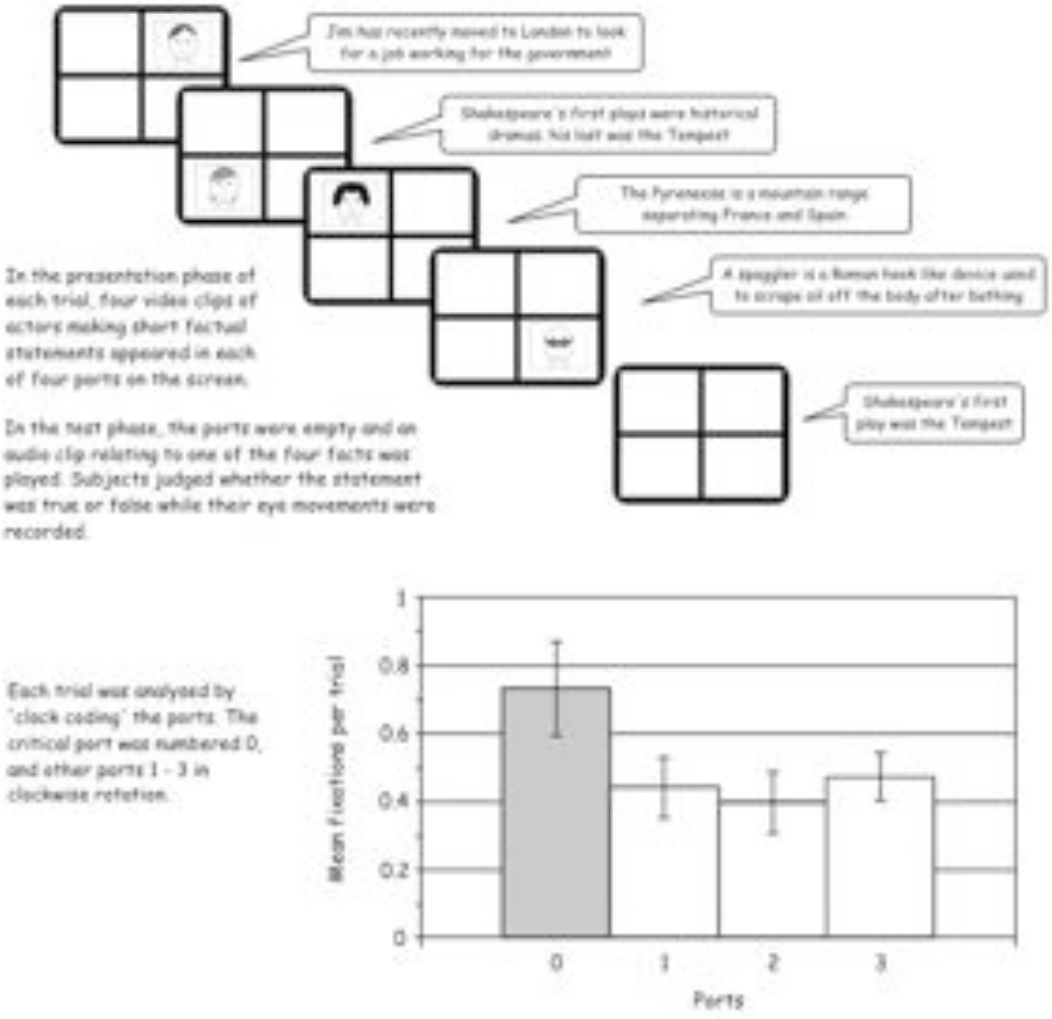
Jim has recently moved to London to look for a job working for the government
Shakespeare's first plays were historical dramas, his last was the Tempest
The Pyrenees is a mountain range separating France and Spain
A spaggler is a Roman hook like device used to scrape oil off the body after bathing
Shakespeare's first play was the Tempest
In the presentation phase of each trial, four video clips of actors making short factual statements appeared in each of four ports on the screen.
In the test phase, the ports were empty and an audio clip relating to one of the four facts was played. Subjects judged whether the statement was true or false while their eye movements were recorded.
Each trial was analysed by 'clock coding' the ports. The critical port was numbered 0, and other ports 1 - 3 in clockwise rotation.
Mean Fixations per trial
1
0.8
0.6
0.4
0.2
0
0
1
2
3
Ports

Figure 4

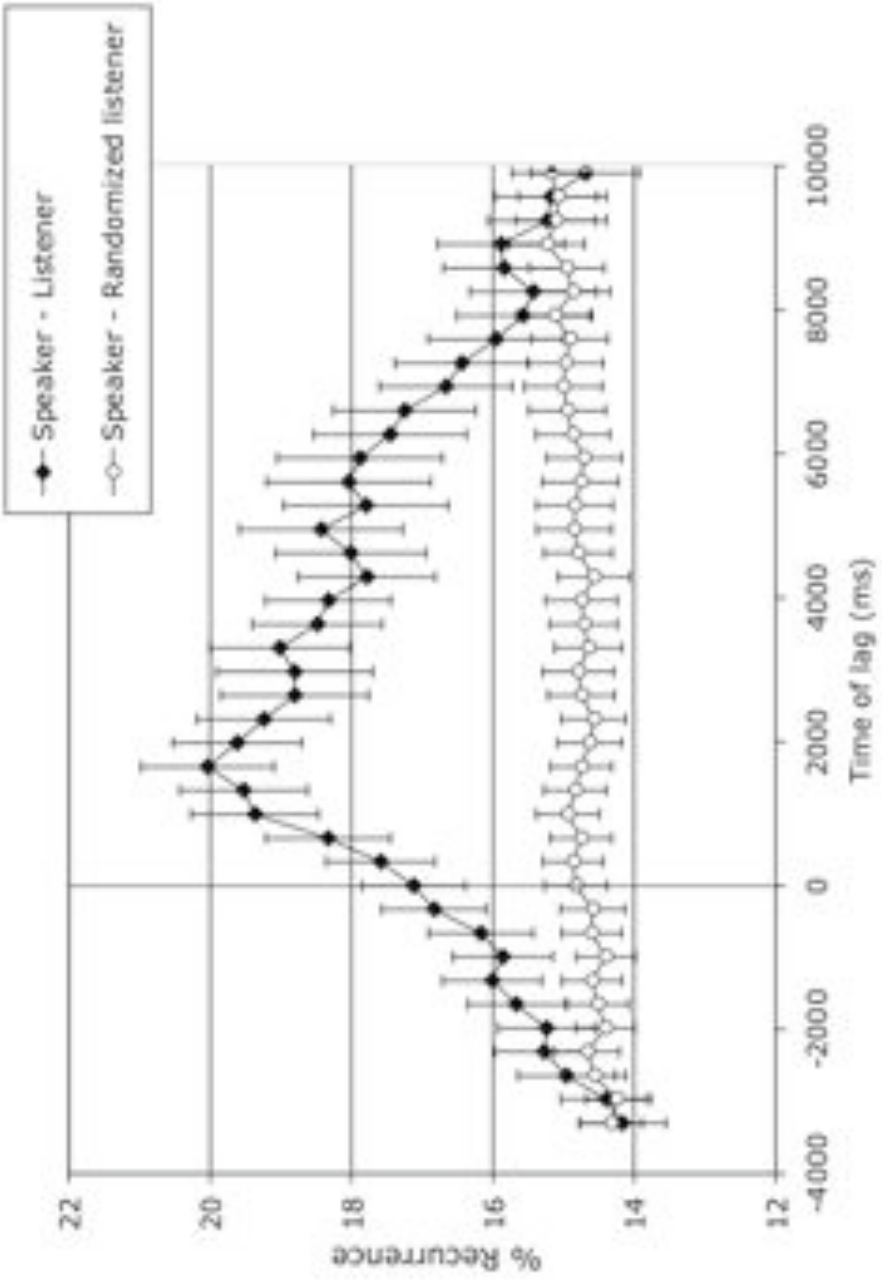
Speaker - Listener
Speaker - Randomized listener
% Recurrence
22
20
18
16
14
12
-4000
-2000
0
2000
4000
6000
8000
10000
Time of lag (ms)

Figure 5.