Toward a Situated View of Language

Michael J. Spivey
Cognitive and Information Sciences
University of California-Merced, CA, USA

Stephanie Huette
Department of Psychology
University of Memphis, TN, USA

Michael J. Spivey
Cognitive and Information Sciences
University of California, Merced
Merced, CA 95343

Introduction

In this chapter we briefly recount some of the historical motivating factors in the field of sentence processing that led it to explore the integration of visual context and language processing (especially with the Visual World Paradigm). We discuss some of the strengths and weaknesses of this experimental methodology and the implications for theories of sentence processing. We conclude that the majority of contemporary findings in sentence processing point to a richly interactive cognitive processing system in which structural constraints and content-based constraints have roughly equal timing and importance in their influence on real-time sentence comprehension. In this emerging theoretical framework, it is expected that any given linguistic process of interest will be best understood when analyzed not in isolation but when embedded in the context in which it is typically situated.

The past several decades of research in sentence processing have seen the pendulum swing between extremes in theoretical frameworks. Around the 1960s, language and communication research was driven chiefly by syntactic structure (Chomsky, 1965), and an assumption that the purpose of language is to produce an internal representation of a transmitted message. Herb Clark (1992) later dubbed this long-standing tradition the "language-as-product" approach. This framework was supported with laboratory tests on theories of transformational grammar (Miller, 1962) and clausal processing (Bever, Lackner, & Kirk, 1969). Around the 1970s, a resurgence of a psychological framework called the "New Look" (Erdelyi, 1974) helped renew an emphasis on semantics (Lakoff, 1971), pragmatics (Clark & Haviland, 1977), and their fluid interaction with syntax (Marslen-Wilson, 1975). This framework treats language not as a message-transmission device but instead as a richly interactive enterprise that is

part and parcel of coordinated action among multiple people. Clark (1992) dubbed this alternative tradition the "language-as-action" approach (see also Trueswell & Tanenhaus, 2005).

By the 1980s, the field of sentence processing returned its emphasis to structure, with syntactic parsing as the autonomous front-end processor in a staged-based modular account of sentence processing (Frazier & Rayner, 1982; Ferreira & Clifton, 1986). In the 1990s, parallel interactive constraint-based approaches rose to prominence once again, with new experimental evidence (Altmann, Garnham, & Dennis, 1992; MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995).

Coincident with those theoretical oscillations over those decades, there tended to be oscillations between the predominant experimental methods being used. With some exceptions, the studies supporting modular stage-based accounts of sentence processing generally used pared-down contexts and the earliest on-line measures available (e.g., eye-movement measures while reading isolated sentences on a computer screen in the dark). By contrast, the studies supporting interactive dynamic accounts of sentence processing tended to use rich realistic contexts and tasks and relatively off-line measures of processing (e.g., analyses of natural conversation transcripts during cooperative tasks). Consequently, there was a common assumption by the early 1990s: if an experiment showed processing interactions between structure and content, then the temporal precision of its experimental methods was probably just too coarse to detect that brief early processing stage during which syntactic processing took place autonomously and in a context-free manner.

All this changed when headband-mounted eyetracking during spoken language comprehension became one of the new prominent experimental methods in the field of sentence processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; for an underappreciated predecessor, see Cooper, 1974). In this paradigm, participants have their eye movements recorded while they look at visual objects on a table or on a computer screen, and listen to spoken instructions or stories about those objects (for a detailed methodological introduction, see Pykkönnen & Crocker, this volume). For better or worse, this new method eventually became known as the Visual World Paradigm, an approach that permeates this volume. Methodologically speaking, the Visual World Paradigm allows the best of both worlds, in that these two seemingly mutually-exclusive experimental design features were finally combined:

1) rich realistic contexts and tasks
2) the real-time recording of eye movements in response to linguistic input

What does Context Mean?

Every psycholinguist acknowledges that context is important, but some theoretical positions reserve the influence of context to a late-stage module that merely revises or corrects the output of an autonomous early-stage module (e.g., Rayner, Carlson, & Frazier, 1983; Staub, 2011; Swinney, 1979). In this type of account, just about anything could be the early-stage "process-in-question," and just about anything else could be the "context." For example, the process-in-question could be syntactic parsing and the context could be pragmatic discourse constraints (Altmann & Steedman, 1988; Ferreira & Clifton, 1986). Or the process-in-question could be word recognition and the context could be syntactic structure (Goodman, McClelland, & Gibbs, 1981; Tanenhaus, Leiman

& Seidenberg, 1979). The curious thing that happened in the field of sentence processing, in particular, is that the prevailing emphasis on the importance of syntactic structure had the effect of allowing many researchers to slip into the implicit assumption that syntactic parsing was, by default, *the* "process-in-question," and everything else was "context."

In actuality, the process-in-question can be anything one wishes to manipulate and test experimentally, be this syntax, semantics, pragmatics or phonetics. Context will always be relative to this main variable, and what we contend here is that there is absolutely nothing that cannot be context. In doing so, two implications emerge: there is a continuum of context strength ranging from very unrelated to very related, and that in principle anything can *become* context. The former could be tested by seeing if people are sensitive to degrees of relationship strength, and the latter can be thought of both intuitively, and investigated experimentally.

Intuitively, imagine we take two very unrelated words that one would never expect to hear together, such as "potato" and "sky". A potato is traditionally not thought of as related to the sky. But if every time my coauthor and I meet we say "The potato is in the sky", then after a period of time we will begin to use "sky" as context for "potato". One may think of "refrigerator" as a *better* context for a potato, but it is *better* because we have experience with potatoes being in this location. Perhaps the *best* context for a potato is "ground" because it is common knowledge that this is where potatoes grow and spend most of their time. Again, this is the *best* context because of the extent of our experience with seeing potatoes in this location, or simply by others using this as a context most often linguistically.

To rephrase this definition of context theoretically, it naturally stems from a statistical learning account where percepts and features are defined by the strength of their connections, and those connections emerge as a result of the embodied and situated character of natural language use (Louwerse, 2008). These connections are developed as a result of co-occurrence: two things in close proximity in either space or time. Thus, two words in the same sentence, or two objects sitting near one another on a table, could constitute some of this learning, by ear and by eye respectively. If certain discourse devices exhibit co-occurrences with certain syntactic structures (e.g., Crain & Steedman, 1985), then this too will be learned. Many seemingly high-level inferences can be the result of spatiotemporal proximity, for example children attributing the cause of an event based on order, rather than another causal cue (Bullock & Gelman, 1979). Proximity and probability are the core principles of this account, though their exact role in a learning mechanism still remains much debated (Levy, 2011; see also Jones & Love, 2011).

Thus, in a fully interactive dynamic process of language comprehension, no one information source can be *the* "process-in-question." Rather, every information source that is relevant (or correlated with behavioral outcomes) is combined as soon as it is available. Syntax, semantics, phonological correlations, lexical frequency effects, discourse information, and visual/situational information are all contextual constraints for each other. Context is relative.

What does a Real-Time Measure Mean?

Experimentally, those intuitions about context can be applied in the following manner. If it is indeed the case that various information sources can perform as context

for each other, then our experimental designs should try to steer toward ecologically valid tasks that situate the language user in a realistic environment where many of those potential contexts are present (and systematically controlled as much as possible). If our tasks were to continue to focus on *one* "process-in-question" and *one* contextual manipulation, while brutishly eliminating all other contextual variables from the stimulus environment, then our research field would risk producing results that do not generalize to natural situations. Importantly, there is nothing that in principle makes these richly contextualized circumstances mutually exclusive with continuous real-time measures of cognitive processing. It is merely a historical accident that the two have tended not to converge.

The Visual World Paradigm exploits natural eye movements to provide a continuous real-time measure of what objects/locations in the visual environment are attracting attention moment-by-moment as a result of the participant processing linguistic input in a variety of situational contexts. A great deal of research in visual cognition and cognitive neuroscience has convincingly shown that, under unrestricted viewing conditions, where the eyes move is a very useful index of where attention is being directed (Hoffman & Subramaniam, 1995). This is largely due to the fact that eye movements and visual attention have many brain areas in common (Corbetta, 1998).

Since eye movements are so tightly interwoven with cognitive processes, and they happen 3-4 times per second, recording them thus provides a rich semi-continuous measure of language and cognition. With this eyetracking methodology, one sees first-hand the fluidity with which eye movements respond to the continuous stream of spoken linguistic input, and how those eye movements then change what parts of the visual

context project onto the foveas, and how that newly foveated object changes the way the next phoneme is processed. This perception-action loop (a la Neisser, 1976) has such a continuous-in-time circular flow that the causal chain (of whether a foveated visual stimulus caused a cognitive process to begin or whether a cognitive process caused a visual stimulus to be foveated) becomes impossible to unravel into a simple linear sequence.

In this way, use of this methodological tool has profound consequences for theory development in psycholinguistics. It is actually quite common for new scientific tools to inspire new perspectives on old theories – such as when electrophysiological measurements by DuBois-Reymond and Helmholtz supplanted the comparative physiology techniques used in the 19th century, and thus dramatically shifted the study of physiology from being a qualitative science to becoming a quantitative science (Lenoir, 1986; see also Gigerenzer, 1992). Scientific tools and scientific theories are not as independent of one another as they are often treated. By collecting multiple measurements within the time span of a single experimental trial, instead of the traditional one-measurement-per-trial, the dense-sampling measurement of eye movements allows the experimenter to obtain a glimpse at the ongoing temporal dynamics of a single cognitive process – not just its end result. In the case of the perception-action loop of eye movements, what we observe is a recurrent causal loop of ongoing cognitive processes instigating eye movements that then substantially alter the trajectory of those same cognitive processes every few hundred milliseconds. Thus, each cognitive event is simultaneously caused by the sensory results of the previous eye movement, and causes the direction of the next eye movement, and then may itself be

altered mid-process due to the sensory result of that new eye movement. The new perspective on the old theory, in this case, is one in which dynamical systems theory, emergence and self-organization (Beer, 2000, Elman, 2004; Spivey, 2007; Van Orden, Holden, & Turvey, 2003) may figure prominently in the explanation of language processes in a visual context.

It is not necessary for one's own metatheoretical stance to drift toward dynamical systems theory, emergence and self-organization, as a result of exploring the Visual World Paradigm. However, when sifting through the data from this methodology, it is inevitable that the range of theoretical alternatives one considers will expand. The temporal fluidity with which different information sources seem to interact, as evidenced by the eye movement patterns, can at times be difficult to reconcile with traditional non-cascading stage-based models of real-time processing. The adaptation of headband-mounted eyetracking methods from visual cognition experiments (Ballard, Hayhoe & Pelz, 1995) into psycholinguistic experiments (Tanenhaus et al., 1995) opened up the floodgates for a wide range of experimental designs that altered the theoretical landscape not just in sentence processing (Chambers, this volume; Knoeferle, this volume), but also in referential processing (Engelhardt & Ferreira, this volume; van Gompel & Järvikivi, this volume), discourse comprehension (Kaiser, this volume), figurative language processing (Huette & Matlock, this volume), perspective-taking (Barr, this volume), and natural conversation (Brown-Schmidt, this volume; Famer, Anderson, Freeman, & Dale, this volume).

Syntactic Ambiguity Resolution in the Visual World Paradigm

A number of important insights have been obtained from the application of eyetracking (and other dense-sampling measures of motor movement, such as postural sway and computer-mouse tacking) to spoken language processing in a constraining visual context. One of the most important of these insights is that language comprehension is not simply incremental (such that words are processed upon arrival, rather than waiting for a phrase to be delivered before parsing it), but is genuinely continuous in time. To truly be "incremental," the process would need to have identifiable increments in time. However, every time we look at a potential increment (whether it be a sentence, a word, or a phoneme) we find temporal fluctuations within the processing of that putative increment – suggesting that the increment has sub-increments within it that are interacting with other information sources. Just as physics came to grips with the fact that no atom is indivisible, psycholinguistics is gradually coming to grips with the fact that no linguistic unit is indivisible.

Another important insight from the Visual World Paradigm is that the continuous cascade of processing appears to go not just in feedforward but also in feedback and through lateral connections. In the following sections, we recount this wide variety of contextual sensitivities that are observed at many levels of language processing. This richly interactive dynamic account of language encourages the field to do more than merely take the old fashioned box-and-arrow diagram of language processing and add new arrows connecting previously unconnected boxes. A dynamical systems framework of language encourages the field to move away entirely from the box-and-arrow metaphor and instead adopt an approach that combines all information sources into one

high-dimensional state space where the interaction between different formats of information is constrained in a graded statistical fashion (Elman, 2004; Gaskell & Marslen-Wilson, 2002; Onnis & Spivey, 2012), but never summarily prohibited by the architecture of the system (as argued in Forster, 1979, and Staub, 2011).

An important real-time measure of this fluid and immediate interaction between syntactic information and situational context information came from work by Tanenhaus et al. (1995), in their development of the Visual World Paradigm. They placed real three-dimensional objects on a table in front of the participant (who wore a headband-mounted eyetracker) and recorded their eye movements while they carried out instructions that were spoken live into a microphone, such as "Put the apple that's on the towel in the box." That unambiguous control sentence was juxtaposed with a syntactically ambiguous version, "Put the apple on the towel in the box," which can be expected to cause listeners to briefly consider treating "on the towel" as the destination of the *put* event. When the table had only one apple on it (resting on a towel), participants frequently looked at a second irrelevant towel, as though they were briefly considering placing the apple on that other towel. This eye movement was thus indicative of a syntactic garden-path effect in that visual context: halfway through the sentence, people temporarily considered a structural parse that involved attaching "on the towel" to the verb. By contrast, when the same instruction was delivered in a context that had *two* apples (one already on a towel and the other not), that garden-path eye movement no longer happened. Essentially, the presence of an extra apple (which was not resting on a towel) introduced a referential ambiguity for the noun phrase "the apple," such that the prepositional phrase "on the towel" had to be syntactically attached to the noun phrase to disambiguate the reference

(see also Altmann & Steedman, 1988). Thus, the syntactic garden-path was prevented by the visual/situational context.

One concern with those results was that the garden-path may have been avoided not by syntax consulting visual context information but by the simple fact that, in the two-referent context, the eyes were busy vacillating between the two apples while the disambiguating information in the sentence was eventually delivered. What was needed was a visual context in which "the apple" was not quite referentially ambiguous, but still readily accommodated parsing "on the towel" as a modifier for that noun phrase – instead of being attached to the verb to denote the destination of the action. To deal with this concern, Spivey, Tanenhaus, Eberhard, and Sedivy (2002) designed a "3-and-1-referent" context, in which the extra apple was replaced by a trio of indistinguishable apples. In this context, "the apple" clearly refers to the lone apple resting on a towel because the determiner "the" presupposes uniqueness of that referent (Heim, 1982; Spivey-Knowlton & Sedivy, 1995). As a result, participants almost never looked at the trio of apples when they heard "Put the apple…" And yet, the naturalness of "on the towel" being a noun-phrase modifier in that visual context still allowed them to avoid the syntactic garden-path.

Another concern even with those results is the fact that on any one particular trial, the data show the subject either looking at the garden-path object or not. This complicates the parsing account that one can formulate. It could be that two syntactic parses are being simultaneously considered after the ambiguity is encountered, and context is able to quickly bias the competition process between those two parses (e.g., MacDonald et al., 1994; Spivey, Anderson, & Farmer, 2013). Alternatively, instead of a

competition process, it could be that only one parse is ever held in working memory at any one time, and context can immediately participate in determining which single parse is pursued (Van Gompel, Pickering, Pearson, & Liversedge, 2005; Van Gompel, Pickering, & Traxler, 2001). In the former scenario, individual experimental trials should comprise a continuous distribution with gradations between mild and strong magnitudes of garden-path effects. In the latter scenario, individual trials should either involve a garden-path effect or not, and thus should comprise a bimodal distribution. Since the eye-movement data cannot help but produce a binomial distribution in which each event either did or did not involve a fixation of the garden-path object, it is difficult to use those data to distinguish between these two theoretical alternatives.

An adaptation of the Visual World Paradigm that allows for the production of a normal distribution in which each event can show a gradation of garden-path magnitude (if such exists) is computer-mouse tracking. In computer-mouse tracking, the streaming x,y coordinates of mouse position over time are recorded while participants select and/or move objects on the computer screen. Partial consideration of one object followed by final selection of a different object is often realized as a curved mouse trajectory that initially moves somewhat toward the partially considered object and then directly toward the selected object. The magnitude of that curvature toward the competitor object can be treated as a graded indicator of how strongly that unchosen alternative was considered (Spivey, Grosjean, & Knoblich, 2005). In the case of syntactic ambiguity resolution, this allowed Farmer, Anderson, and Spivey (2007) to record continuous mouse trajectories when people were instructed to, "Put the apple on the towel in the box," and measure how much the movement of the apple curved toward the irrelevant towel on its way to the

box. Not only did they find that changes in visual context could make the syntactic garden-path come and go (just as in the eye-movement data), but they also found that the magnitude of that garden-path curvature was able to clear up the question of whether: a) individual trials involve a binomial option of either garden-pathing or not (Van Gompel et al., 2001, 2005), or b) parallel competition among two active syntactic parses can produce graded degrees of garden-path magnitude (MacDonald et al., 1994; Spivey et al., 2013). While the former predicts a bimodal distribution of substantially curved mouse trajectories and straight ones, the latter predicts a unimodal distribution of moderately curved trajectories. Consistent with a parallel competition account of syntactic ambiguity resolution, Farmer et al. found a clearly unimodal distribution that was generally normal (though somewhat leptokurtotic). For more in-depth discussion of computer-mouse tracking, see Farmer, Anderson, Freeman, and Dale (this volume).

It is worth noting that these results of visual context influencing the competition between two mutually exclusive syntactic parses of a sentence should not be interpreted as indicating that it is simply *the objects themselves* in the visual context that can exert that influence. In certain circumstances, it would be more appropriate to think of it as *the actions that are afforded by those objects* that are exerting the influence on syntactic ambiguity resolution (see Chambers, this volume). For example, Chambers, Tanenhaus, and Magnuson (2004) gave participants instructions like, "Pour the egg in the bowl on the flour." and then manipulated the affordances of those eggs. When participants were viewing a real 3-D table with two liquid eggs (extracted from their shells, one in a glass and one in a bowl), along with an irrelevant empty bowl and a pile of flour on wax paper, their eye-movement patterns indicated that they were parsing the syntactically ambiguous

prepositional phrase “in the bowl” as a noun-phrase modifier, and thus avoided the garden-path effect in that visual context. Essentially, both eggs were potential references of “Pour the egg” because they were both pourable. In contrast, when the visual context was subtly changed, such that there were still two eggs but the one in the glass was still in its shell and thus not pourable, all of a sudden the garden-path effect came back! Simply having two referents for “the egg” is not enough to introduce the referential uncertainty that leads to avoidance of the garden-path. There needs to be referential uncertainty for the entire phrase “Pour the egg,” so they both need to be pourable. Thus, the constraints being imposed on the syntactic ambiguity resolution process are not simply visual objects that may or may not be referred to, but rather a more complex notion of the entire situation (and the possible actions that it affords) in which the utterance is being delivered (Barsalou, 1999).

Of course, it would be naïve to think that somehow situational context was the only information source that influenced syntactic parsing. Even in the circumstance of an immersive visual/situational context that constrains the range of actions that could be carried out, more purely linguistic information sources are also playing a role in resolving syntactic ambiguity. In self-paced reading and eye-tracking reading experiments, Trueswell, Tanenhaus, and Kello (1993) already showed compelling evidence for verb-specific biases (in terms of statistical preferences for certain argument structures) having an immediate influence on the resolution of syntactic ambiguity. During reading, these verb-specific preferences can create or prevent a garden-path effect depending on what direction they bias the parsing process.

Snedeker and Trueswell (2004) used the Visual World Paradigm to show that these verb-specific biases can still influence processing even in constraining visual/situational contexts. For example, the verb "choose" does not have a strong statistical preference for an Instrument with-phrase, as in "Choose a donut with the tongs." It is much more common for a with-phrase after "choose" to be a modifier for the noun-phrase as in, "Choose a donut with pink frosting and sprinkles." By contrast, the verb "tickle" is quite frequently followed by an Instrument with-phrase, as in "Tickle the baby with the feather." Verb-specific biases like this become quite relevant when someone is instructed to "Choose the cow with the stick," or "Tickle the cow with the stick," in a visual context that has a stick, and two toy cows (one of which is holding a stick). Clearly, there is a wide variety of information sources that influence syntactic ambiguity resolution, including lexical biases (Snedeker & Trueswell, 2004), semantic biases (Trueswell, Tanenhaus, & Garnsey, 1994), discourse context (Altmann & Steedman, 1988), and visual/situational context (Tanenhaus et al., 1995), among others. Moreover, it looks as though these information sources combine as soon as they are available and their integration may involve a competition process that gradually settles somewhat toward one or another of the syntactic alternatives (MacDonald et al., 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998). Importantly, it may very well be that the various information sources that immediately influence parsing do so with different relative weights depending on the mode of language processing, such as reading versus instruction-following in a visual context versus unconstrained two-way conversation (see discussion in Spivey et al., 2002).

Semantic Comprehension in the Visual World Paradigm

Syntactic parsing is certainly not the only linguistic process that will reveal its underpinnings when tested in the Visual World Paradigm. Just like the structure, the content of a spoken sentence shows itself to be incrementally understood and sensitive to contextual biases as the speech unfolds over time. In fact, it is sometimes even faster than incremental: it is anticipatory. Altmann and Kamide (1999) presented participants with line drawings of scenes containing a potential agent (e.g., a boy) and several possible direct objects (only one of which was edible, e.g., a cake). When participants heard "The boy will move the cake," they pretty quickly moved their eyes from the boy to the cake. However, when they heard "The boy will eat the cake," many of them were already fixating the cake before the word "cake" was uttered! Thus, the verb's thematic role preferences (e.g., direct objects that are edible) were immediately combined with the situational context to make the full sentence understood before it was even finished being spoken (see also Kamide, Altmann, & Haywood, 2003, and Kamide, in this volume). One may then ask what happens when the situational context and the verb's preferred thematic role properties don't quite match up? What if your thematic role knowledge of verbs tells you that spying is typically performed by detectives and hexing is typically performed by wizards, but the visual scene shows you a detective holding a magic wand and a wizard using a pair of binoculars? Knoeferle and Crocker (2006) showed that, in situations like that, participants make anticipatory eye movements that are consistent with using the *visual context* as the guide for likely agents of spying events and hexing events. Similar to that observed in the syntactic ambiguity resolution literature, it looks as though verb-based preferences are indeed still active during spoken language comprehension in

the Visual World Paradigm, but when the visual context conflicts with them, the co-present situational information tends to outweigh the stored lexical biases (see also Knoeferle's chapter in this volume).

It is worth noting that it is not only references to Subjects and Objects of a verb that can direct participants' attention in the Visual World Paradigm. The verb itself can direct attention, even when its implication of motion is subtle and metaphorical. Take, for example, the sentence, "The road goes through the desert." The road itself doesn't actually *go* anywhere. It is made of asphalt that stays right where it was laid. However, cognitive linguistic analyses have suggested that there is a kind of imaginary form of motion, i.e., fictive motion, which is generated by the use of such action verbs in non-action descriptions. Richardson and Matlock (2007) used the Visual World Paradigm to show that people's eye movements actually provide a hint into that perceptually simulated visual motion during comprehension of fictive motion sentences. When the context sentence described the road as rocky and difficult to traverse, participants spent more time passing their eyes over the road region of the display than when the context sentence described the road as smooth and easy. It was as though listeners were mentally simulating movement on the road, and went slower when the road was difficult. Control sentences that did not contain fictive motion, such as "The road is in the desert," showed no such effect of the context sentence (see also Huette and Matlock, this volume).

Not only can the eyes be guided by a perceptual simulation of visual information (such as motion), that isn't actually present in the static visual display, but they can also be guided by a visual memory of information – after the display has become entirely blank. For example, Altmann (2004) replicated some of the anticipatory eye-movement

results from Altmann and Kamide (1999) with a display that initially presented the potential Subjects and Direct Objects and then took them away. With the screen totally blank, participants still made eye movements to the corresponding locations of the appropriate entities (which were now empty) while the spoken sentence was being understood. Knoeferle and Crocker (2007) then followed suit, showing that the demonstrated preference for depicted-event biases over thematic-role biases (Knoeferle & Crocker 2006) wanes over time after the scene has been removed, such that thematic-role biases drive processing more and more as the visual memory decays. (See also Chambers & San Juan, 2008, for evidence of the integration of immediately-perceptible constraints and more abstract thematic/conceptual constraints in real-time reference resolution). In fact, the Visual World Paradigm can even be informative when there was never any visual input provided in the first place! Rather than visual memory of a recently viewed scene, a perceptual simulation generated solely by the spoken sentence can guide the eyes to move in ways that correspond to the relative locations of entities and events in a story. Spivey and Geng (2001) delivered spoken vignettes to participants while they faced a large blank projection screen, and observed that stories about upward-moving events elicited a preponderance of upward saccades, and stories about downward-moving events elicited a preponderance of downward saccades. Even more subtle differences in the spoken input, such as grammatical aspect, can influence the eye movement pattern while participants are viewing a blank screen. Huette, Winter, Matlock, Ardell, and Spivey (2014) compared a series of sentences delivered in the past progressive form, such as "John was delivering a pizza" (which uses imperfective aspect to emphasize the ongoing nature of the event) and a series of sentences delivered in the

simple past form, such as "John delivered a pizza" (which uses perfective aspect to emphasize the completed end-state of the event). With the imperfective grammatical aspect, they found a wider dispersion of eye movements over the span of the blank display, and significantly shorter fixation durations, suggesting that the grammatical emphasis on ongoing action elicits eye movement patterns that are consistent with a perceptual simulation of visual motion – even while viewing a completely blank screen.

Spoken Word Recognition

At a finer time scale, of words instead of sentences, the Visual World Paradigm has provided some of its most well known discoveries in the real-time dynamics of spoken word recognition. Eberhard, Spivey-Knowlton, Sedivy, and Tanenhaus (1995) reported delayed mean saccade latencies to a named object (such as "candle") when a real 3-D object with a similar name was also visually present (such as a candy), as well as frequent eye movements to that object with the similar name (Spivey-Knowlton, 1996). Allopenna, Magnuson, and Tanenhaus (1998) used a computer display to extend those findings to include not just cohorts (such as looking briefly at a candy when instructed to "Click the candle") but also rhymes (such as looking briefly at a handle when instructed to "Click the candle"). Moreover, they mapped out a computational implementation of how the time course of activations of lexical representations in the brain might be mapped onto the time course of proportions of fixations on objects with those names. Using the TRACE neural network model of speech perception (McClelland & Elman, 1986), they fit the activation curves of lexical nodes onto the proportion-of-fixation curves in the eye-movement data. Thus, a linking hypothesis was computationally

fleshed out between putative activations of lexical representations and the observed behavior.

Due to priming studies, it had been generally accepted that multiple lexical representations become active in parallel during the recognition of a spoken word (Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989). However, seeing the eyes spontaneously move toward objects that have names that should be partially active was a compelling demonstration of this prediction (which stems from most theories of spoken word recognition). Nonetheless, these eye-movement results were initially met with some degree of skepticism on the grounds that the task and display might be unnatural and prone to strategic influences. For example, the apparent parallel activation of multiple lexical items during spoken word recognition in this paradigm could, in principle, be the result of a working memory buffer containing the names of the objects in the display (e.g., candy, candle, penny, spoon, etc.) It could be that -- in these less than ecologically valid circumstances involving computer-delivered instructions to move random objects -- acoustic-phonetic input is mapped onto that temporarily-constructed working memory buffer rather than onto the lexicon. If there were a cognitive module called the lexicon that was required for normal everyday spoken word recognition, and the task in those experiments didn't even use that module, then the results would indeed have little application to normal everyday spoken word recognition.

Notably, there are numerous findings that make it hard for that "working memory buffer" account to hold water. For starters, lexical frequency effects show up in the eye movement data (Dahan, Magnuson & Tanenhaus, 2001; Magnuson, Tanenhaus, Aslin & Dahan, 2003). Competitor objects with higher frequency names are more likely to attract

eye movements than competitor objects with lower frequency names. That shouldn't happen if the acoustic-phonetic input were purely being mapped onto a temporary buffer. Also, interlingual cohort effects show that bilinguals listening to one of their languages will often produce eye movements to objects whose names are phonetically similar in the other language (Ju & Luce, 2004; Marian & Spivey, 2003; Spivey & Marian, 1999; Weber & Cutler, 2004). For example, Russian-English bilinguals will often look at a stamp when instructed to "Pick up the marker," because in Russian the stamp is known as *marka*. It is perhaps unlikely that bilinguals construct a temporary buffer in *both* of their languages for all the objects that are in front of them.

The finding that partial phonological similarity in an object's name can attract an eye movement during the real-time comprehension of a spoken word has been extended in a number of ways. Dahan, Swingley, Tanenhaus, and Magnuson (2000) showed that, in French, a gendered determiner that preceded the temporarily ambiguous spoken word (e.g., *les boutons*) could prevent eye movements to the object with a similar-sounding name (e.g, *bouteilles*) simply because it has the wrong gender marking. Thus, the activation of lexical representations during incremental processing of a word's unfolding acoustic-phonetic input is constrained by the context of the determiner delivered only a couple hundred milliseconds beforehand. And it is more than phonological similarity that pulls attention and eye movements to competitor objects in the display. Semantic similarity works as well. When instructed to "click the piano," people often look at a trumpet (Huettig & Altmann, 2005). And when instructed to "click the lock," people often look at a key (Yee & Sedivy, 2001). In fact, Yee and Sedivy (2006) showed that, due to the phonological similarity, "click the logs" can activate the lexical representation

for *lock* (even though there is no lock present) and thus indirectly trigger eye movements to the key! High-dimensional state-space accounts of semantic similarity provide accurate predictions of the frequency of eye movements to these competitor objects (Huettig, Quinlan, McDonald, & Altmann, 2006), whether the state-space is based on feature norms (Cree & McRae, 2003) or on n-gram-based corpus statistics (Lund & Burgess, 1996).

As was seen with syntactic ambiguity resolution, there is a weakness with eye-movement data in that each individual trial can either show evidence of a brief misinterpretation of the spoken word (a sort of "lexical garden-path") or not. On any given trial, the participant either looks at the competitor object or doesn't. Thus, one could still adhere to an account that suggests the lexicon conducts its mapping of acoustic-phonetic input onto lexical items and completes any competition processes internally *before* sending its finalized output to other subsystems (such as reaching and eye-movement subsystems). An account like this would suggest that the reaching and eye-movement subsystems never receive the cascaded parallel output of multiple partially activated lexical representations. Rather, the lexicon gives single unitary commands to those action subsystems, sometimes quickly and sometimes slowly, and occasionally must send revision signals to instigate corrective eye movements and corrective reaching movements (van der Wel, Eder, Mitchel, Walsh, & Rosenbaum, 2009). To test an account like this, the Visual World Paradigm must extend itself to other measures that are not as ballistic and discrete as saccadic eye movements are. Recording computer-mouse movements can allow the detection of graded curvatures in the response movements.

Spivey, Grosjean and Knoblich (2005) found that when participants were instructed to "click the candle," their computer-mouse movements showed graded curvature toward the midpoint between the candle and the candy, before finally settling into the image of the candle. This curvature was reliably greater for cohort conditions (candle/candy) than for control conditions (candle/towel). Moreover, computational modeling of dynamically averaged motor commands produces remarkable fits to the mouse-tracking data (Spivey, Dale, Knoblich, & Grosjean, 2010). A theoretical comparison of the kinds of data extracted from eye-tracking and from mouse-tracking show that they have complementary strengths and weaknesses, and can easily be conducted at the same time (Magnuson, 2005). In fact, this mouse-tracking version of the Visual World Paradigm has revealed continuous real-time competition between representations that are active in parallel in other domains as well, such as color categorization (Huette & McMurray, 2010), semantic categorization (Dale, Kehoe, & Spivey, 2007), gender stereotypes (Freeman & Ambady, 2009), social attitudes (Wojnowicz, Ferguson, Dale, & Spivey, 2009), and even decision making (McKinstry, Dale, & Spivey, 2008).

Phoneme Perception in the Visual World Paradigm

As we zoom in the timescale from sentences to words to phonemes, we see that the observation of parallel partial activation of multiple representations extends even to the level of the dozens of milliseconds of acoustic-phonetic input that distinguishes one phoneme from another. For example, a mere 40 ms of delayed voicing (vibration of the vocal chords) is what chiefly discriminates the spoken syllable /pa/ from the spoken syllable /ba/. Classic findings have shown that when this voice onset time (VOT) is

varied parametrically with synthesized speech, listeners exhibit a categorical distinction in how they identify and discriminate speech tokens on the continuum between the canonical /ba/ and the canonical /pa/ (Liberman, Delattre, & Cooper, 1958). At first glance, it looked as though listeners were not even processing the within-category gradations in the acoustic-phonetic input (i.e., the sensory differences between a /ba/ with 10 ms VOT and a /ba/ with 20 ms VOT).

However, a couple decades later, Pisoni and Tash (1976) reported one early hint that the speech processing system was being somehow affected by the imperfectness of a /ba/ that has a VOT somewhat near the /pa/ range. Although participants consistently labeled /ba/ tokens near the category boundary as "ba," they produced longer reaction times when doing it. This suggested some kind of time course to the speech categorization process, during which the within-category acoustic variation was not quite being entirely discarded.

Another couple decades later, Bob McMurray extended the Visual World Paradigm to speech perception, and obtained not only reaction times during identification of stimuli from a /ba/-/pa/ continuum, but also proportions of eye fixations on the response icons (McMurray & Spivey, 1999). With canonical versions of /ba/ and of /pa/, participants would look only at their correct chosen response icon and click it with the mouse cursor. With versions of /ba/ and /pa/ that were near the category boundary, participants tended to quickly fixate both the /ba/ and /pa/ icons on the computer screen before finally clicking their consistently selected icon. McMurray and colleagues further demonstrated that this evidence for partial activation of both phonological representations (voiced and unvoiced) lasted long enough to influence spoken word recognition, such as

when hearing the word "bear" or "pear" with a VOT continuum (McMurray, Tanenhaus, & Aslin, 2002). In fact, with each additional 5 ms of VOT, participants exhibited a systematic gradient increase in their likelihood of fixating the pear image before clicking the bear image. And once the VOT was across the category boundary, each additional 5 ms of VOT caused a systematic gradient *decrease* in likelihood of fixating the bear image before clicking the pear image (McMurray, Aslin, Tanenhaus, Spivey & Subik, 2008). Thus, it would appear that about as fine-grained in temporal resolution as one can go in the stimulus -- 5 ms increments of speech sounds -- the Visual World Paradigm provides evidence that is consistent with a theoretical framework in which spoken language comprehension is continuously sensitive to the cascaded sensory, perceptual, and cognitive processes involved in turning sound waves into internal representations of meaning.

Spoken Sentence Production in the Visual World Paradigm

So far, this review has been focused on findings in language *comprehension*. However, tracking people's eye movements is also informative for understanding real-time language *production*. Soon after the Visual World Paradigm was developed, several researchers adapted it for observing what parts of a visual scene attract overt attention during the few seconds it takes to formulate and produce a spoken utterance. In fact, in the right circumstances, the eye-movement pattern can even be used to make predictions about what grammatical form the participant's upcoming spoken utterance will take!

Meyer, Sleiderink, and Levelt (1998) showed that when participants viewed two objects on the computer screen and were instructed to name the left object first and then the right object, they routinely fixated the left object and then the right object, and then

began naming them. Thus, their eyes were typically fixating the second object when they began naming the first object. Moreover, when an object was a *given* entity in the discourse, because it had already been mentioned, it tended to be fixated for briefer periods of time than when that object was a *new* entity in the discourse, because it had not yet been referred to (van der Meulen, Meyer, & Levelt, 2001).

Griffin and Bock (2000) presented participants with line drawings of two entities that were interacting with one another, such as a donkey kicking a horse, and asked participants to describe the scene any way they wanted to. Not surprisingly, the majority of participants used an active voice, as in "The donkey kicked the horse," and before they began their spoken utterance, their eye position tended to start on the donkey and then move to the horse. However, on those trials where participants wound up producing a passive voice sentence, as in "The horse was kicked by the donkey," their eye-movement pattern tended to reveal that alternative grammatical formulation even before the utterance began. Participants who were about to use the passive voice, but had not yet opened their mouths, tended to initially fixate the horse and then fixate the donkey (see also Griffin, 2004). Results like these show that, as people formulate an utterance, their eyes naturally move to the objects that they are thinking about and preparing to talk about – and in the particular sequence that the particular grammatical construction would entail.

Dialogue and Reference in the Visual World Paradigm

The research discussed so far tends to implicitly treat language use as if it were a unidirectional process. Either the participant is seeing a visual scene and then producing

a sentence to describe it, or she is comprehending a sentence spoken in the context of a visual scene, but never both. The findings described so far generally provide support for a situated approach to understanding the various processes of language. That is, when one analyzes sentence comprehension, it is crucial to pay attention to the context in which that process is situated. The system that is performing those sentence-level linguistic computations is *embedded* (a technical term from dynamical systems theory) in a larger system that is performing sensorimotor computations on the relevant properties of the physical environment. The same applies when one analyzes semantic comprehension, or spoken word recognition, or phoneme perception, or sentence production. The system of interest is always embedded (or situated, or contextualized) in a larger encompassing system that is dramatically influencing its real-time behavior. And things get even more interesting – and of course more complicated – when there are *two systems of interest*, one in each of the interlocutors! When two people are engaged in a language-mediated joint task, each of these systems of interest become not only embedded in their larger context but they also become *tightly coupled* with one another (another technical term from dynamical systems theory).

For several years, practitioners of the Visual World Paradigm were reticent to release the experimental controls of prepared and recorded stimuli and fixed visual displays. However, the moment one begins to study reference resolution in this paradigm, it becomes clear that there is a remarkably fluid temporal continuity with which listeners map each new speech sound onto possible matches in the visual context. This clearly would have consequences for natural interactive conversation, where

interlocutors share the visual context and often anticipate one another in ways that are impressively constructive.

This fluid continuity in reference resolution in the Visual World Paradigm was first demonstrated by Kathleen Eberhard and colleagues, when she instructed participants to "touch the starred yellow square" amid an array of several colored blocks (Eberhard et al., 1995). Some of the blocks might have stars, some might be yellow, but only one is starred, yellow and in the shape of a square. She found that participants were mapping the adjectives onto the relevant objects in the scene before the head noun was even spoken. If there was only one block with a star on it, then participants were settling their eye position on the referent block about 200 ms after hearing the word "starred" -- around the time the adjective "yellow" was being spoken. Thus, listeners were using the features of the objects in the display to dynamically restrict the referential domain of relevant objects to respond at the contextually-relevant point-of-disambiguation in the spoken noun phrase. This real-time incrementality with which these adjectives were being interpreted -- apparently without needing the head noun to which they are syntactically attached -- even led to follow-up experiments that showed how the visual system can use those adjectives to guide visual search and make it more efficient when searching for "a red vertical bar" (Spivey, Tyler, Eberhard, and Tanenhaus, 2001). Thus, not only can visual context tell language processing what to do, but linguistic context can tell visual processing what to do as well (Anderson, Chiu, Huette, & Spivey, 2011).

Keysar, Barr, Balin, and Brauner (2000) then extended this type of reference resolution paradigm into a social context with an experimental confederate, where a listener might be expected to map their understanding of a spoken instruction onto the

common ground (or mutual knowledge) shared between the two interlocutors. If a listener can see that a particular object is not visible to the speaker, then one might expect that she would not consider it as a potential referent, because the speaker is unlikely to refer to an object that he cannot see. However, Keysar et al. found that listeners frequently made eye movements to privileged objects (which the speaker could not see) when those objects had names similar to what the speaker was instructing them to pick up. This finding helped spark a flurry of research in social psychology suggesting that people are frequently egocentric in their interpretation of language and other social situations (e.g, Epley, Keysar, Van Boven, & Gilovich, 2004; Lin, Keysar, & Epley, 2010).

Interestingly, rather than interpreting this egocentrism as evidence that common ground is not accommodated among interlocutors, subsequent work suggests that common ground does indeed play an important immediate role in language comprehension, but it does so in concert with many other linguistic and perceptual factors. For example, Hanna, Tanenhaus, and Trueswell (2003) directly compared a common-ground condition – where there was a target object and a communally visible competitor object – to a privileged-ground condition where the competitor object was a "secret shape" that was not in common ground but instead only in the listener's privileged ground. Although the privileged-ground "secret shape" reliably interfered with reference resolution, indicating that common ground information was unable to summarily rule out the privileged object from attracting attention, the common ground competitor object exerted reliably more interference than the privileged competitor shape did. Thus, common ground information was clearly influencing the earliest eye

movement patterns, just in a probabilistic fashion. Essentially, when the acoustic-phonetic input maps substantially onto the name of an object in the listener's field of view, this is one constraint that will contribute to the likelihood that the eyes move to that object. And when the common ground among speaker and listener suggests that this same object is not likely to be referred to by the speaker (because the speaker cannot see that object), this is one factor that will contribute to the likelihood that the eyes *do not* move to that object. Neither of these opposing constraints is able to completely eliminate the effects of the other. Therefore, even though common ground is indeed being taken into account immediately (along with many other constraints), the listener will still occasionally look at an object that is only in her privileged ground.

Similar findings of the immediate use of common ground information to *partially* reduce the perceived relevance of a privileged object were also reported by Nadig and Sedivy (2002) with 5- and 6-year-olds. Then Hanna and Tanenhaus (2004) extended these observations in a natural collaborative task involving a cooking scenario with real physical kitchen implements and ingredients. In general, as the task and context become more natural and ecologically valid, it appears that any and all relevant information sources – from low-level lexical and syntactic constraints to high-level broadly encompassing constraints such as common ground -- are integrated into the evolving interpretation of incoming linguistic input as soon as they are available (e.g., Kaiser & Trueswell, 2008; see also Kaiser, this volume).

In fact, as the conversational context in the laboratory becomes even more realistic, an obvious component to add is natural speech disfluencies and speech repairs. As a matter of fact, eye-movement data show that listeners will interpret a brief speech

disfluency as an indicator for a given/new distinction in the conversation (Arnold, Tanenhaus, Altmann, & Fagnano, 2004), and they will partially update their real-time interpretation when a spoken verb is repaired as a different verb (Corley, 2010). But it takes a measure of bravery to truly put this claim about realistic conversation to the test, and actually allow experimental participants to engage in natural, ecologically valid, *unscripted* two-way conversation – with its spontaneous disfluencies, repairs, and general free-formedness -- while still making every effort to maintain experimental control and real-time measurements in the laboratory. Brown-Schmidt, Campana, and Tanenhaus (2005) did exactly that with a large array of blocks and pictures of various objects that two participants used in an unscripted interactive problem-solving task. After analyzing the transcripts of the conversations, they found a couple hundred instances where complex noun phrases were temporarily ambiguous with respect to the set of objects to which they could refer – a bit like Eberhard et al.'s (1995) reference to "the starred yellow square" amidst a set of colored blocks of various shapes. They found that even in this unscripted natural conversation situation, listeners would dynamically restrict the referential domain to look at objects referred to in the speech stream very soon after the contextually-relevant point-of-disambiguation – just as observed in the scripted instruction task used by Eberhard et al. Interestingly, however, this ebb and flow of dynamic restricting of the referential domain was so ubiquitous that whenever the transcript provided an opportunity to test for spoken word cohort effects (e.g., Allopenna et al., 1998; Eberhard et al., 1995; Spivey-Knowlton, 1996), which consisted of 75 adventitious references to pictures that had cohort competitors also in the display at the time, there wasn't a single instance where a listener looked at a cohort competitor.

Essentially, realistic unscripted conversation naturally tends to restrict the domain of reference, via shared goals and shared attention, such that it is rare for two objects with cohort names to be situationally relevant at the same time (for further discussion, see Brown-Schmidt's chapter in this volume).

The shared goals and shared attention of a natural unscripted dialogue tend to induce a shared common experience of the conversational situation that is supported by a wide variety of coordinated behaviors. Not only do interlocutors tend to unintentionally mimic each other's syntactic choices in production (e.g., Dale & Spivey, 2006; Pickering & Garrod, 2004), they also unintentionally slip into a wide variety of emergent behavior-matching actions (Shockley, Richardson, & Dale, 2009; see also Clark, 2012). For example, their eye-movement patterns on a shared visual display become coordinated (Richardson, Dale & Kirkham 2007). Their manual and facial movements become coordinated (Louwerse, Dale, Bard, & Jeuniaux, 2012). Even the subtle postural sway patterns around the two bodies' centers of gravity become coordinated (Shockley, Santana, & Fowler, 2003). Essentially, as two people become engaged in a natural dialogue, with numerous references to their shared situational context, their various subsystems of linguistic, perceptual, and motor processes become tightly coupled across the two people. For brief periods of time, they may even function more like one system than two.

Conclusion

In this chapter, we have walked through a progression of numerous subfields in psycholinguistics where the Visual World Paradigm has assisted in important advances in

our understanding of how linguistic and perceptual information interact immediately to conjure up an evolving understanding of what an utterance means in the context of the situation. All of these applications of the paradigm are currently active areas of research, as can be seen in the other chapters in this volume. The common *methodological* thread among these research areas is that they have all derived their unique insight into the online processing of linguist input by employing a dense-sampling method that provides multiple measures (usually eye movements) within the time course of each trial. If this wide variety of findings share one common *theoretical* thread, it is this: The temporal continuity in the uptake and processing of linguistic input and of perceptual input is exactly what allows these partially-processed portions of information to be mapped onto each other in real time.

The real-time moment-by-moment delivery of spoken language is often likened to "beads on a string" delivered incrementally, one at a time, and the language user's task is to comprehend the full pattern of the necklace. This is a useful metaphor, but it has one misleading characteristic inherent to it. Whenever one looks at the fine grain temporal dynamics of the delivery of a putative "bead" of language (be it a clause, or a word, or a phoneme), it becomes clear that the bead is made of several smaller beads that are processed incrementally. In actuality, there are no beads. Rather than "beads on a string," a more apt metaphor might be water flowing down a river, or maybe Cantor dust sliding through an hourglass. In fact, the term "incremental" doesn't quite do justice to this incredibly fluid process. There appears to be a temporally continuous cascading of multiple partially active representations as linguistic information flows through the language processing systems. Indeed it may be that at no point does any particular

information source (e.g., phonological, syntactic, semantic, pragmatic) hold back from sharing its activation patterns with other information sources.

This observation of "processes in cascade" (McClelland, 1979) has important consequences for our understanding of the architecture of the language processing system. Not only must we let go of the *information encapsulation* once proposed by Fodor (1983) for lexical and syntactic modules, but if that information permeability is constantly flowing in cascade between the various subsystems, then even the *domain specificity* of these putative modules becomes somewhat compromised. That is, if a syntax module is continuously receiving semantic and pragmatic input (on the time scale of milliseconds) that it uses to modify the syntactic structures it is in the process of forming, then the rules and constraints it is following are obviously not purely *specific* to the *domain* of syntax. In such a scenario, there is no point in time during which a measurement of that syntax module's internal computations would reveal representations that had been constructed by purely syntactic forces. There would always be some detectable influence from non-syntactic constraints on those representations that are inside the syntax module.

Importantly, the resulting compromise of the domain specificity of the syntax module should not be taken as an argument for syntax simply not existing. Even advocates of encoding syntax and semantics inside the same computational substrate (e.g., Elman, 1990; Tabor & Hutchins, 2004) would not themselves interpret the tight coupling of these two information sources as evidence that one of them doesn't exist. Let's take an example from vision research. Vision scientists have been discovering that their visual modules are more interactive and less domain-specific than once thought. As

a result, findings of motion perception interacting with color information (Møller and Hurlbert, 1997) and with transparency information (Trueswell & Hayhoe, 1993) are generally interpreted as evidence that there is still a visual subsystem that processes mostly visual motion information, but it also processes some other sources of information a little bit. Similarly, psycholinguistics is slowly coming to grips with the idea that any given linguistic module is promiscuous enough with its information flow to process some sources of information that are not what it is primarily known for. In such an account, these modules are partially specialized, but they are not quite domain-specific and certainly not informationally encapsulated.

From phoneme recognition all the way up to natural unscripted conversation, and everywhere in between, the Visual World Paradigm has provided a treasure trove of important insights into how various linguistic processes are immediately influenced by the contextual processes in which they are situated or embedded. As a result, the modular view of language is slowly giving way to a general situated view of language, which is arguably on its way to becoming mainstream in the field of experimental psycholinguistics. A dynamical systems theory approach to situated language, which is well stocked with mathematical tools for understanding how situatedness may be an embedding of one system inside a larger system, is however still in its infancy. The findings of interactivity between various linguistic processes and the context in which they are embedded make it difficult for the field to continue with its implicit adherence to the old modular box-and-arrow model of language comprehension, where phonology is a domain-specific processor that sends its output to syntax, which is a domain-specific processor that sends its output to semantics, which is a domain-specific processor that

sends its output to pragmatics (see Onnis & Spivey, 2012). However, the field has not yet settled on what formalism, or schematic diagram, will replace that old chestnut. Nonetheless, one thing seems for sure: You don't have to go dynamical, but you can't stay modular.

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*, 419-439.

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

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

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30,* 191-238.

Altmann, G., Garnham, A., & Dennis, Y. (1992). Avoiding the garden-path: Eye movements in context. *Journal of Memory and Language*, *31*, 685-712.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, *137*, 181-189.

Arnold J. E., Tanenhaus, M.K., Altmann R. J., & Fagnano, M. (2004). The old and thee, uh, new. *Psychological Science,15*, 578–582.

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

Barr, D. (this volume). Visual world studies of conversational perspective taking.

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

Beer, R.D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences, 4,* 91-99.

Bever, T.G., Lackner, J.R., & Kirk, R. (1969). The underlying structures of sentences are the primary units of immediate sentence processing. *Attention, Perception, & Psychophysics, 5,* 225-234.

Brown-Schmidt, S. (this volume). Visual environment and interlocutors in situated dialogue.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2005). Real-time reference resolution by naïve participants during a task-based unscripted conversation. In: Trueswell, J. & Tanenhaus M. (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions.* pp. 153–171. Cambridge, MA: MIT Press.

Bullock, M. & Gelman, R. (1979). Preschool children's assumptions about cause and effect: Temporal ordering. *Child Development, 50,* 89-96.

Chambers, C. (this volume). The role of affordances in visually-situated language comprehension.

Chambers, C. G., & San Juan, V. (2008). Perception and presupposition in real-time

language comprehension: Insights from anticipatory processing. *Cognition, 108,* 26–50.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 687–696.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R.O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1-40). Hillsdale, NJ: Erlbaum.

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

Clark, H. H. (2012). Spoken discourse and its emergence. In Spivey, M., McRae, K., & Joanisse, M. (Eds.), *Cambridge Handbook of Psycholinguistics*. (pp.541-557). NY, NY: Cambridge University Press.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6,* 84–107.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences, 95,* 831-838.

Corley, M. (2010). Making predictions from speech with repairs: Evidence from eye movements. *Language and Cognitive Processes, 25*, 706–727.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological syntax processor. In D.R. Dowty, L. Karttunnen, & A.M. Zwicky (Eds.), *Natural Language Parsing* (pp. 320-345). Cambridge University Press.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General, 132*, 163–201.

Dahan, D., Magnuson, J., & Tanenhaus, M. (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.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition, 35, 15-28.*

Dale, R. & Spivey, M. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*, 391-430.

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.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Elman, J.L. (2004). An alternative view of the mental lexicon. *Trends in cognitive sciences, 8,* 301-306.

Engelhardt, P. & Ferreira, F. (this volume). Reaching sentence and reference meaning.

Epley, N., Keysar, B., Van Boven, L., Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327-339.

Erdelyi, M.H. (1974). A new look at the new look: Perceptual defense and vigilance. *Psychological Review, 81,* 1-25.

Farmer , T., Anderson, S., Freeman, J. & Dale, R. (this volume). Coordinating action and language.

Farmer, T., Anderson, S., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory and Language*, *57*, 570-595.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348-368.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Forster, K. (1979). Levels of processing and the structure of the language processor. In W. Cooper & E. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. (pp.27-850). Erlbaum Press: Hillsdale, NJ.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178-210.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science, 20*, 1183–1188.

Gaskell, M., & Marslen-Wilson, W. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*, 220-266.

Gigerenzer, G. (1992). Discovery in cognitive psychology: New tools inspire new theories. *Science in Context, 5*, 329–350.

Goodman, G. O., McClelland, J. L. & Gibbs, R. W. (1981). The role of syntactic context in visual word recognition. *Memory and Cognition, 9,* 580-586.

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 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*, 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 and Language*, *49*, 43–61.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases.* Amherst, MA. GLSA.

Hoffman, J.E. & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Attention, Perception, & Psychophysics, 57,* 787-795.

Huette, S., & McMurray, B. (2010). Continuous dynamics of color categorization. *Psychonomic Bulletin & Review*, *17*, 348-354.

Huette, S. & Matlock, T. (this volume). Figurative language processing.

Huette, S., Winter, B., Matlock, T., Ardell, D. H., & Spivey, M. (2014). Eye movements during listening reveal spontaneous grammatical processing. *Frontiers in Psychology*, *5*, 410.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, *96*, B23-B32.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica, 121*, 65-80.

Jones, M. & Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition, *Behavioral and Brain Sciences, 34*, 169-231.

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

Kaiser E. & Trueswell J. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes,* 23, 709–748.

Kaiser, E. (this volume). Discourse level processing.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–159.

Kamide, Y. (this volume).

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Sciences, 11*, 32–38.

Knoeferle, P. (this volume). Accounting for visual context effects on situated language comprehension.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking. *Cognitive Science, 30*, 481–529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*, 519–543.

Lakoff, G. (1971). On generative semantics. In D. Steinberg & L. Jacobovits, (Eds.), *Semantics.* (pp. 232-296). Cambridge: Cambridge University Press.

Lenoir, T. (1986). Models and Instruments in the Development of Electrophysiology, 1845-1912. *Historical studies in the Physical and Biological Sciences, 17*, 1-54.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* 1055-1065.

Liberman, A., Delattre, P., & Cooper, F. (1958). Some rules for the distinction between voiced and voiceless stops in initial position. *Language and Speech, 1*, 153–167.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556.

Louwerse, M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review, 15*, 838-844.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science*, *36*(8), 1404-1426.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods: Instruments and Computers*, *28(2)*, 203-208.

MacDonald, M., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676-703.

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought: Commentary on Spivey, Grosjean, & Knoblich (2005). *Proceedings of the National Academy of Sciences*, *102*, 9995-9996.

Magnuson, J., Tanenhaus, M., Aslin, R., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*, 202-227.

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. (1975). Sentence perception as an interactive parallel process. *Science*, *189*, 226-228.

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

Marslen-Wilson, W. & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576-585.

McClelland, J. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287-330.

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

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*, 22-24.

McMurray, R. & Spivey, M. (1999). The categorical perception of consonants: The interaction of learning and processing. In *Proceedings of the Chicago Linguistic Society Panels, 35-2*, 205-221.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*, B33–42.

McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1609-1631.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the effects of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *37*, 283-312.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects. *Cognition, 66,* B25–B33.

Miller, G.A. (1962). Some psychological studies of grammar. *American Psychologist, 17,* 748-762.

Møller, P., & Hurlbert, A. (1997). Interactions between colour and motion in image segmentation. *Current Biology*, *7*, 105-111.

Nadig, A. & Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329-336.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* New York, NY: W.H. Freeman.

Onnis, L. & Spivey, M. J. (2012). Toward a new scientific visualization for the language sciences. *Information, 3*, 1-28.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169-226.

Pisoni, D., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, *15*, 285-290.

Pykkönnen, P. & Crocker, M. (this volume). Attention in vision and language.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 358-374.

Richardson, D.C., Dale, R. & Kirkham, N.Z. (2007) The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science,* 18 (5), 407-413.

Richardson, D.C & Matlock, T. (2007). The integration of figurative language and static depictions: An eye movement study of fictive motion. *Cognition,* 102, 129-138.

Shockley, K., Richardson, D. C. & Dale, R. (2009) Conversation and coordinative structures, *Topics in Cognitive Science,* 1, 305–319.

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 326-332.

Snedeker, J. & Trueswell. J. (2004). The 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. J. (2007). *The continuity of mind.* New York: Oxford University Press.

Spivey, M. J., Anderson, S. & Farmer, T. (2013). Putting syntax in context. In R. Van Gompel (Ed.), *Sentence Processing*. (pp.115-135). New York: Psychology Press.

Spivey, M. J., Dale, R., Knoblich, G., & Grosjean, M. (2010). Do curved reaching movements emerge from competing perceptions? *Journal of Experimental Psychology: Human Perception and Performance, 36,* 251-254*.*

Spivey, M. J. & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects*. Psychological Research, 65,* 235-241.

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

Spivey, M. J. & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, *10*, 281-284.

Spivey, M. J. & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1521-1543.

Spivey, M. J., Tanenhaus, M., Eberhard, K. & Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45,* 447-481.

Spivey, M. J., Tyler, M., Eberhard, K., & Tanenhaus, M. (2001). Linguistically mediated visual search. *Psychological Science*, *12*, 282-286.

Spivey-Knowlton, M. J. (1996). *Integration of visual and linguistic information: Human data and model simulations.* Ph.D. Dissertation, University of Rochester.

Spivey-Knowlton, M. J. & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*, 227-267.

Staub, A. (2011). Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of Experimental Psychology: General, 140,* 407-433.

Swinney, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 645-659.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 431–450.

Tanenhaus, M., & Trueswell, J. (1995). Sentence comprehension. In J. Miller, & P. Eimas (Eds.), *Handbook of Cognition and Perception*. New York: Academic Press.

Tanenhaus, M., Leiman, J.M., & Seidenberg, M.S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18,* 427-440.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, *268*, 1632-1634.

Trueswell, J., & Hayhoe, M. (1993). Surface segmentation mechanisms and motion perception. *Vision Research*, *33*, 313-328.

Trueswell, J.C. & Tanenhaus, M.K., (Eds.) (2005). *Processing world-situated language: Bridging the language-as-action and language-as-product traditions.* Cambridge, Mass: MIT Press.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language, 33*, 285-318.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 528-553.

van der Meulen, F., Meyer, A., & Levelt, W. (2001). Eye movements during the production of nouns and pronouns. *Journal of Memory and Cognition, 29*, 512–521.

van der Wel, R. P. R. D., Eder, J., Mitchel, A., Walsh, M., & Rosenbaum, D. (2009). Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: A commentary on Spivey, Grosjean, and Knoblich (2005). *Journal of Experimental Psychology: Human Perception and Performance, 35,* 588–594.

Van Gompel, R. P. G. & Järvikivi, J. (this volume). The role of syntax in sentence and referential processing.

van Gompel, R.P.G., Pickering, M.J., Pearson, J., & Liversedge, S.P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language, 52*, 284-307.

van Gompel, R. P. G., Pickering, M., & Traxler, M. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*, 225-258.

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

Weber, A., & Cutler, A. (2004). Lexical competition in non- native spoken-word recognition. *Journal of Memory and Language, 50*, 1–25.

Wojnowicz, M., Ferguson, M., Dale, R., & Spivey, M. J. (2009). The self-organization of

explicit attitudes. *Psychological Science*, *20*, 1428-1435.

Yee, E., & Sedivy, J. (2001). *Using eye movements to track the spread of semantic activation during spoken word recognition.* Paper presented to the 13th annual CUNY sentence processing conference, Philadelphia.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1–14.