Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm

Anuenue Kukona

Whitney Tabor

Department of Psychology, University of Connecticut

and

Haskins Laboratories, New Haven, Connecticut

Accepted Version (*Cognitive Science*)

1.28.2011

Correspondence concerning this article should be addressed to Anuenue Kukona, Department of Psychology, University of Connecticut, 406 Babbidge Road U-1020, Storrs, CT 06269. E-mail: anuenue.kukona@uconn.edu. Phone: (860) 486-3515. FAX: (860) 486-2760.

Keywords: Dynamical systems; Self-organization; Local coherence; Artificial neural networks; Connectionist modeling; Visual world paradigm; Eye tracking; Sentence Processing; Ambiguity resolution; Embodied cognition.

Abstract

The visual world paradigm presents listeners with a challenging problem: they must integrate two disparate signals, the spoken language and the visual context, in support of action (e.g., complex movements of the eyes across a scene). We present Impulse Processing, a dynamical systems approach to incremental eye movements in the visual world that suggests a framework for integrating language, vision, and action generally. Our approach assumes that impulses driven by the language and the visual context impinge minutely on a dynamical landscape of attractors corresponding to the potential eye-movement behaviors of the system. We test three unique predictions of our approach in an empirical study in the visual world paradigm, and describe an implementation in an artificial neural network. We discuss the Impulse Processing framework in relation to other models of the visual world paradigm.

1. Introduction

1.1. The visual world paradigm

The technique of eye tracking in the visual world paradigm (VWP) has provided many new insights into the temporal dynamics of language processing (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Altmann & Kamide, 1999, 2007, 2009; Chambers, Tanenhaus, & Magnuson, 2004; Knoeferle & Crocker, 2006, 2007; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). In this paradigm, listeners are presented with both a visual context and language about that context, and they are typically instructed to interact with the items in the visual context in some way. Listeners completing tasks in the VWP must integrate two informationally dense but disparate signals, the spoken language and the visual context, and execute, in response, complex patterns of eye movements across the "visual world."

To fully understand data from this paradigm, one needs a framework that specifies how the mental system interacts with both verbal and visual information, and integrates them in support of action. In line with several previous models of VWP data (e.g., Allopenna et al, 1998; Mayberry, Crocker, & Knoeferle, 2009; Spivey, 2007), we suggest that dynamical systems theory (e.g., Abraham & Shaw, 1992; Strogatz, 1994) provides a common currency that is well suited to this purpose.

Although we focus on the VWP in the present work, our interest is not simply in eye movements in highly constraining contexts. Rather, we use the VWP as a testing ground for our dynamical systems approach, with its tight experimental control, and its rich behavioral measures. Based on these results, we suggest that the approach provides a general framework for

understanding the link between cognition and action more broadly.

1.2. The dynamical systems approach

Dynamical systems theory is a general framework for studying interactions among inter-dependent variables. An often-studied type of dynamical system is organized around a set of "attractors," or stable states that the system returns to when it is perturbed. Such systems are well suited to modeling the regularity of human behavior in the presence of input coming from multiple modalities, with unpredictable and only loosely coordinated timing, because the attractors reduce the variability to a few interpretable patterns. In fact, most current formal models of VWP data are dynamical systems with attractors, typically implemented as groups of neuron-like (connectionist) elements with feedback connections. While some of these models focus on individual word perception (e.g., Allopenna et al., 1998; Spivey, 2007), others address the integration of information coming from multiple words in a sentence (e.g., Mayberrry et al., 2009; see also Roy & Mukherjee, 2005).

Here, we offer a general framework, called Impulse Processing, for understanding the coordination of verbal and visual input in support of action. We thus focus, for the action part, on the movement of the eyes across a visual scene. Our framework addresses incremental interpretation of multi-word sentences, and models the progression of eye movements as the sentence unfolds. This problem is a complex one, involving many aspects of language processing: speech recognition, syntactic, semantic, and pragmatic interpretation, coordination with motor systems, etc. We do not implement solutions to all of these problems here. Instead, we show that if we make well-motivated assumptions about the information structure in each of

these areas, then we can build a dynamical model that smoothly integrates the information, and makes several distinctive empirical predictions about the way structures form. Implementation and empirical testing of one part of this dynamical model are described in Section 3. We believe that the attractor-based approach provides a particularly effective foundation for the kind of large, interdisciplinary undertaking that is needed to make a full, functioning model of language understanding.

The central hypotheses of Impulse Processing are:

1. The system's dynamics involve noisy movement on a landscape of attractors.
2. The landscape's shape is continually adjusted by small impulses reflecting the current interpretation of verbal and visual input.
3. States of the system correspond to actions.

The landscape of attractors proposed under Impulse Processing is a continuous surface with ridges and valleys, as depicted in Fig. 1. The landscape may be thought of as characterizing an aspect of the nervous system of a human who is listening to language, while focusing on a visual scene, and looking, at each moment in time, at an object in the visual scene. Regions in the landscape are associated with objects in the visual scene: if the system is at a location in the region corresponding to a particular object in the visual scene, then, assuming that other features of the nervous system's state do not suppress the impulse, the system directs its gaze at that object. In other words, the location on the landscape specifies the action (cf. Jacobs & Michaels, 2007).

We focus here on the action of the eyes, which, in the context of the motionless scenes we consider here, consists of a series of fixations of various durations. In Fig. 1, the regions associated with some sample objects are indicated on the plane below the surface. The landscape

is a *potential surface* for the dynamics of the system. This means that the behavior of the system can be conceptualized as the action of a drop of water sliding down the landscape. The local minima of the landscape thus correspond to attractors of the system.[1] If such a landscape of smooth ridges and valleys were unchanging, then the system would eventually gravitate to an attractor and stay there, thus fixating indefinitely on whatever object's region contained the attractor. However, the system is not static—it changes due to the impulses, which can be caused by the speech signal, by other changes in the external environment, by changes in where the system is looking, which affect its relationship to the external environment, or by changes in other aspects of the nervous system (e.g., memories). Although we focus on verbal and visual input at present, the framework is sufficiently general to handle inputs from any number of sensory sources.

The forces which adjust the landscape can, in principle, be very complex. We assume that one aspect of the adjustment consists of minute, random variation in the position of the system on the landscape. Such *noise* may be thought of as corresponding to the small-scale, apparently random variation in the activity levels of neurons. The noise does not, in general, have a large influence on the shape of the system's trajectories, and the corresponding looking behaviors. However, when the system is near a saddle point of the surface (e.g., the point labeled S in Fig. 1), then the noise becomes potent: it determines which basin the system falls into and therefore, potentially, which object it looks at. Thus the system may be thought of as mainly deterministic, with stochasticity playing a role at certain juncture points. The determinicity corresponds to the broad structural properties of language perception and reference. For example, if there is a scene containing a cat, and the language says, "Look at the cat," then the landscape will be shaped so

[1] Note that the object regions do not necessarily line up with the *attractor basins,* or sets of points from which the system gravitates to a common attractor.

that the system directs its gaze at the cat. But if the scene contains two cats, one of which is on a chair, and the language says, “Look at the cat that’s on the chair,” then, when the system has heard only the partial input, “Look at the cat…,” it will have a saddle point between the basins corresponding to the two different cats. In this case, the stochasticity will cause the system to fixate on one cat or the other with equal probability, assuming that there are no additional biasing factors.

In an experimental trial involving a scene with multiple, clearly separated objects on a computer screen, attractive regions form for each object, dimpling the landscape with basins. As the language identifies referents of interest in the context, or gleans evidence that a particular region contains relevant information, the strength of the attractors is modulated. More relevant basins grow stronger, making them more likely to capture the current state of the system, and less relevant basins grow weaker. We assume that the signals arriving from sensory organs and other parts of the nervous system impinge only minutely (in small pulses) on the basin structure at each moment in time. Therefore, the attractors exhibit an inertia that serves to simplify and combine the information coming from a variety of modalities on a complex schedule. In this regard, our modeling approach contrasts with the symbol processing mechanisms of digital computer models of the mind, which face challenging problems of coordinating the timing of events related to signals arriving erratically from multiple sources.

Many current approaches to language processing focus on interpretation of linguistic entities, like words or phrases, as a critical intermediate goal of processing. In contrast, our approach makes a complete link between perceptions (of language) and the landscape that specifies actions (here, movements of the eyes). We assume that linguistic interpretations arise emergently in this process: that is, fully formed, coherent linguistic structures sometimes occur

as a consequence of the combined effect of several action-oriented impulses, but they are not required to form fully, and indeed there may be moments when they are only partly formed, and cases (e.g., in some garden path sentences) when they fail to fully form. An advantage of linking the perceptions to the actions without insisting on fully-formed linguistic structure in the middle is that it confers some robustness: the system acts, regardless of whether it understands or not, and this action is sometimes effective (e.g., upon hearing "Grasp the skirpet under the dial," it may be helpful to look under the dial, even if one has no idea what a *skirpet* is). The direct linking assumption also has the consequence that the system exhibits states of partial order that are not expected under standard assumptions about coherent structure formation (cf. Konieczny, Müller, Hachmann, Schwarzkopf, & Wolfer, 2009; Tabor, Galantucci, & Richardson, 2004). Here we argue that such errant structure formation predictions provide an empirical basis for distinguishing the current framework from other approaches.

1.3. Self-organization in dynamical systems

Self-organization refers to the formation of global structure among a group of independent but interacting elements, under restricted environmental conditions. This phenomenon manifests in a wide variety of settings: for example, in Rayleigh-Bénard convection, molecules in a fluid organize themselves into regular convection cells under particular conditions of temperature and viscosity (Koschmieder, 1993). Similarly, reagents in the Belousov-Zhabotinsky reaction form nonhomogeneous spatial patterns like stripes and spirals when mixed in particular concentrations (Zhabotinsky, 1991), pebbles in Greenland organize themselves into lenticular and ring-shaped patterns under particular conditions of surface water

flow and granularity (Kessler & Werner, 2003), and muscles of the human vocal apparatus coordinate to achieve articulatory goals under conditions in which a centralized controller is not plausibly the source of coordination (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). There is a developing formal theory of self-organizing phenomena, including studies which link self-organization to phenomena of grammar (Crutchfield & Young, 1990; Tabor, 2002) and systems of neural units (Haken, 2004). The current formal theory is remote from psycholinguistics partly because it has not been clear how this strongly bottom-up approach can address complex linguistic behavior. This paper extends Tabor and Hutchins' (2004) claim that self-organization explains some otherwise puzzling phenomena of parsing, suggesting that it can provide insight into the challenging question of how parsing is integrated with general perception and action. A helpful outcome of the current work is that it indicates a way that the relatively macroscopic observations we can make about sentence-level language processing are connected to the more microscopic insights about particle/fluid/neural interaction that prior work on self-organization has established.

In self-organizing systems, small, interacting elements form larger, global structures. Global organization develops as tensions among potentially incongruous local structures are resolved. Impulse Processing is a self-organizing process in the sense that each impulse leaves an impression on the potential surface which interacts with the impressions of other impulses to produce structured behavior. Thus, Impulse Processing makes the prediction that even grammatically inconsistent bottom-up structure can form during the global organization process. Tabor et al. (2004) and Konieczny et al. (2009) provide evidence for one particular kind of errant bottom up effect, referred to as "local coherence:" in a sentence like "The coach smiled at the player tossed the Frisbee," the phrase "the player tossed the Frisbee" is intended as a noun

phrase, with “tossed the Frisbee” as a reduced relative clause modifier on “player.” Tabor et al. argue that readers are temporarily distracted by the possible interpretation of the phrase as an active clause with “player” as its subject, even though the preceding grammatical context (“The coach smiled at”) seems to rule out such an interpretation. These are called “local coherence” effects because the locally (not the globally) coherent clause interpretation is the claimed source of the interference.

Evidence for the formation of inconsistent bottom-up structure (including local coherence as one type of case) has been reported in a number of language processing situations, in line with the self-organization hypothesis. In each case, global consistency of the relevant representations was not enforced by the cognitive system. Swinney (1979) and Tanenhaus, Leiman, and Seidenberg (1979), for example, demonstrated that even in biasing syntactic contexts (e.g., “spiders, roaches, and other…”), both senses of subsequent ambiguous words (e.g., espionage vs. insect “bugs”) were activated. Although these findings were initially taken as evidence for the partial independence of lexical and and syntactic modules, from the perspective of self-organization, they represent the formation of incongruous local *lexical* structure, despite global sentential context (Kawamoto, 1993, observed these phenomena in an attractor network similar to the one we employ here). More recently, Kukona, Fang, Aicher, Chen, and Magnuson (in press) demonstrated anticipatory looks in predictive contexts (e.g., “Toby arrested the…”) to both contextually appropriate patients of the verb (e.g., crook), and contextually inappropriate agents (e.g., policeman), supporting formation of incongruous local *thematic* structure, despite sentence context. Similarly, Tabor et al. (2004) and Konieczny et al. (2009) demonstrated that locally coherent but globally ungrammatical words strings (e.g., the case of “The coach smiled at the player tossed the frisbee”) interfered with processing, supporting the formation of

incongruous local *syntactic* structure, despite global sentential context. Moreover, Van Dyke (2007) provides evidence that globally ruled-out structures can also form bottom-up from elements that are not adjacent in the speech stream (non-local coherence effects). Relatedly, Allopenna et al. (1998) demonstrated the activation of rhyme competitors during spoken word recognition in the VWP (e.g., looks to a speaker on hearing "beaker," despite clear differences in onset), supporting the formation of incongruous local *sub-lexical* structure, despite global lexical context.

Tabor and Hutchins (2004) implemented a self-organizing model of sentence processing, called SOPARSE, which we adopt as a component of the Impulse Processing framework. SOPARSE assumes that words activate "treelets" (Fodor, 1998; Marcus, 2001) which interact to form global syntactic tree structures. Although we do not include an implementation of SOPARSE in the model presented in Section 3, we make assumptions about the timing of structure formation in the current model that are consistent with the behavior of SOPARSE.

Finally, the abstractness of language seems like a challenge to the claim that perceptions always specify actions. The framework of self-organization offers an answer to this challenge that is closely tied to the phenomenon of local coherence. We take up this issue in the General Discussion.

Next, we describe the visual world experiment we conducted.

1.4. A dynamical systems implementation: ambiguity in the visual and linguistic signals

To more carefully test the predictions of the Impulse Processing framework, we examined eye movements in VWP settings involving ambiguities of two types, which

specifically help to distinguish our approach: an ambiguity of reference (referential ambiguity) and an ambiguity of lexical interpretation and reference (lexical plus referential ambiguity).

We created VWP contexts containing seven items: four items in the four corners of a computer screen display, a reference item (a star) between the top two items, another reference item (a square) between the bottom two items, and a small fixation cross in the middle of the screen (see Fig. 2). The participant heard a spoken instruction over headphones of the form "Click on [Noun Phrase]," where [Noun Phrase] either had the form "the [Noun]" or "the [Noun] that's beside the [Noun]." The task for participants was to listen to each command and to click on the relevant location with the mouse. In simple, *unambiguous* cases, the participant would see an array like Fig. 2A or 2B, and hear either "Click on the snail" or "Click on the snail that's beside the star." Their task was to simply click on the snail, as instructed. In cases of *referential ambiguity*, the target indicated by the language was temporarily consistent with multiple images in the visual context: for example, "Click on the cat…" in the visual context of two cats (Fig. 2A). In cases of lexical plus referential ambiguity, the language contained a lexical ambiguity, and both interpretations of the ambiguity were present in the visual context: for example, "Click on the bat…" in the visual context of a baseball bat and a mammalian bat (Fig. 2B). For both ambiguity types, a relative clause ultimately disambiguated the target referent (e.g., "Click on the cat/bat *that's beside the star,*" such that only one cat/bat was beside a star in the visual context).

To introduce the conceptual organization of our model, we now review its processing of several types of examples. First, we consider a simple unambiguous sentence which includes some redundancy: "Click on the snail that's beside the star," uttered while the model views the scene in Fig. 2A. In this section, for ease of exposition, we focus on describing the pulses that drive the model's behavior, thus highlighting the model's predictions. Later, in Section 3, we

identify guiding principles underlying the pulse structure.

The model assigns attractive regions to items in the visual display. At the beginning of processing, the system has seven attractive regions, one corresponding to each of the items. These regions lie in the high-dimensional activation space of a neural network, which is described in detail in Section 3 and Appendix A. Here, to convey the principles of the model's operation graphically, we use a two-dimensional space to depict a subset of the attractive regions, corresponding to four of the items: the snail, the glove, the star (which is beside the snail) and the square (which is beside the glove). [2]

The strengths of the attractive regions are indicated approximately by the circles in Fig. 3. The relative sizes of the radii indicate how likely the model is to fixate each item in the display.

Initially, the regions are of roughly equal size (we assume there are no biasing factors, such as the visual saliency of items). Therefore, across repeated trials, the network spends approximately the same amount of time in each region (Fig. 3A). Since all items in the display are equally clickable, the words "Click on the" contain no information that favors one item over another, and thus the attractive regions remain the same size during the perception of these words (note that this would not be true of all contexts: the words "Pour the…," for example, would favor items which are pourable over those which are not; Chambers et al., 2004). The arrival of the first informative word, "snail," causes the region corresponding to the snail to gradually grow larger at the expense of the other regions (Fig. 3B). As a consequence, the system tends to spend more time in this region during and shortly after the utterance of "snail".

The arrival of the preposition "beside" causes the regions corresponding to the item *next*

[2] Fig. 3 shows these four attractive regions in locations that serve as reminders of the spatial relationships between the items associated with the regions. However, these depicted spatial relationships between the regions in two dimensions do not correspond directly to the spatial structure of the corresponding neural attractive regions because the geometry of the four-dimensional space does not map isomorphically onto the geometry of the two-dimensional space. For the purposes of the present illustration, however, the differences do not matter.

to the region currently being fixated to expand. Fig. 3C depicts the situation in which the system was nearest to the attractive region corresponding to the snail when the word "beside" was uttered; therefore, the attractive region for the star, which is beside the snail, grows. The arrival of the second noun, "star," causes the attractive region corresponding to the star to grow even larger, at the expense of the other regions (Fig. 3D). The net effect is that as the words "beside the star" are being perceived, the rate of looking at the star increases. The properties illustrated in the remaining frames of Figure 3 depend on syntactic interactions, which are based on the system of SOPARSE (Tabor & Hutchins, 2004). We discuss SOPARSE next and then return to Figure 3.

Impulse Processing assumes that syntactic structural interpretation happens by self-organization, as in SOPARSE (Tabor & Hutchins, 2004; Tabor, 2006). Pieces of linguistic information (coming from different words) bond together to form larger interpreted chunks on a schedule determined by activation dynamics in a system of "syntax neurons" which are a component of Impulse Processing. We have not incorporated the SOPARSE implementation into the Impulse Processing simulation we describe in Section 3. However, the formation of syntactic chunks in Impulse Processing is based on the predictions of the SOPARSE framework.[3] The bonding process for the current example is shown in Fig. 4. Note that chunks for "the snail", "beside", and "the star" form before the chunk corresponding to the unified phrase "the snail beside the star" forms. As each chunk comes online in SOPARSE, a neural node corresponding to that chunk's tree-diagram node undergoes a rapid ramp-up in activation. One may think of the model as "attending" to the semantics of the chunks that are undergoing rapid ramp-up at any

[3] *This importation of mechanisms from SOPARSE, along with the stipulation of various parameters in the Impulse Processing implementation that we describe in Section 3, implies that our results here are not strongly "emergent," in the sense that we do not derive a large variety of behaviors from just a few, low-level stipulations (e.g., a learning rule, and some basic assumptions about environment encoding). Instead, we include, by fiat, various specific mechanisms (e.g., features of SOPARSE, feature detectors at multiple scales, memory as scale-manipulation – see Section 4) which previous work suggests are plausibly present in self-organizing ("emergent structure") systems in order to understand how these mechanisms may work together. We see this as a step toward developing a more unitary treatment.*

moment in time. We assume that the succession of pulses corresponds to the successive attentional states of SOPARSE. Thus, the first three frames of Fig. 4 specify the pulse events we have so far enumerated: widening of the "snail" region during "snail" (Fig. 4A), widening of the star region during "beside" (Fig. 4B), and further widening of the "star" region during "star" (Fig. 4C). Fig. 4D specifies that after the arrival of the second noun "star," the local structures begin to assemble into a complex NP ("the snail beside the star"). This causes the attractive region corresponding to the snail, the head of the complex NP, to grow and engulf the lion's share of the action space, as indicated in Figs. 3E and 3F.

In sum, the model predicts that a person hearing "Click on the snail that's beside the star," while viewing Fig. 2A, should first look at the snail, as a consequence of the impulse for "snail" (Fig. 4A), then look at the star, as a consequence of the impulses for "beside" and "star" (Fig. 4B and C), and then look back at the snail, as a consequence of the formation of the complex NP "snail beside the star" (Fig. 4D).

More complex dynamics occur in cases involving ambiguities. Fig. 5 shows one possible progression of attractive region relationships upon hearing "Click on the cat that's beside the star" while viewing Fig. 2A. Several features of Fig. 5 are worth noting. The arrival of the word *cat*, in a referentially ambiguous context containing two cats, like Fig. 2A, causes the attractive region for both cats (i.e., cat_1, beside the star, *and* cat_2, beside the square), to grow (Fig. 5B). This is because the concept **cat** is simultaneously associated with both cats in the scene. In such a case, as indicated above, the low-level noise in the action space causes the system to make a random choice between looking at cat_1 and looking at cat_2, but both attractors remain present even while the system is fixating on one particular cat. Fig. 5 assumes that the system primarily fixates cat_2 during the utterance of "cat". Therefore, when "beside" arrives, the system fixates

items beside cat_2, causing the attractive region for the square to grow in magnitude (Fig. 5C). Subsequently, however, the word "star" arrives. Analogously to the previous example with the snail, SOPARSE then attends temporarily to "star" and the size of the star region grows (Fig. 5D). Finally, the complex NP corresponding to "cat that's beside the star" ramps up, as the local structures assemble into the more complex type of structure depicted in Fig. 4C. Crucially, the complex NP structure is unambiguously consistent with only a single item in the display, and thus the region for cat_1 becomes dominant over all other regions.

In sum, the model predicts that a person who garden paths on "cat" (where by "garden path," we mean that the person initially focuses on the irrelevant cat in the display), while hearing "Click on the cat that's beside the star" and viewing Fig. 2A, will first look at cat_2, and then will ramp up slightly on looks to the square, then ramp up slightly on looks to the star, and finally will settle on looking at cat_1 (on the basis of the complex NP structure).

Fig. 6 shows a possible progression of attractive region radii on hearing "Click on the bat that's beside the star" while viewing Fig. 2B. In this lexical plus referential ambiguity case, the development associated with the utterance of the ambiguous first noun ("bat") is different from the preceding referential ambiguity case. Unrelated senses of a word (e.g., baseball bat vs. mammalian bat) compete, so that after brief initial activation of multiple senses in a constraining context, only one sense has a large attractive region. In this regard, our model is consistent with results from Swinney (1979), Tanenhaus, Leiman and Seidenberg (1979), and Simpson and Kang (1994; see also Raczaszek-Leonardi, Shapiro, Tuller, and Kelso, 2008; Rodd, Gaskell, & Marslen-Wilson, 2002; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Simpson & Burgess, 1985) which indicate suppression of irrelevant senses of an ambiguous word soon after initial perception in a biasing context (here, the image first fixated in the visual display is the

biasing context). Our implementational assumptions (discussed in Section 3) in this regard are similar to the recurrent network implementation of Rodd, Gaskell, and Marslen-Wilson (2004), which treats alternative meanings of truly ambiguous words (as opposed to polysemous words) as mutually exclusive attractor basins. The choice of which bat is fixated is made randomly, driven by the noise (see Section 3). In Fig. 6, we have assumed that the region for the mammalian bat (bat_2) is the one that grows. As a consequence, the system tends to be looking at the mammalian bat when the word "beside" arrives. In line with the previous examples, this causes the attractive region for the square to grow. Then, when the word "star" arrives, the region for star briefly grows. Finally, the region corresponding to "the bat that's beside the star" (i.e., bat_1) becomes the dominant region, on the basis of the formation of structure at the level of the complex NP (as in Fig. 4D), and captures the large majority of looks.

In sum, the model predicts that a person who garden paths on "bat" while hearing "Click on the bat that's beside the star" and viewing Fig. 2B, will first look at bat_2, and then will ramp up slightly on looks to the square, then ramp up slightly on looks to the star, and will finally settle on looking at bat_1. This progression is very similar to the referential ambiguity case illustrated in Fig. 5. However, there are two notable differences: First, the attractive region for cat_1, ultimately the correct cat, is already large at the onset of disambiguation. Therefore, the transition from primarily looking at cat_2 (Fig. 5D) to primarily looking at cat_1 (Fig. 5E and F) happens relatively quickly. By contrast, the transition from primarily looking at bat_2 (Fig. 6D) to primarily looking at bat_1 (Fig. 6E and F) happens comparatively slowly, because the attractive region for bat_1 has to be built up from a small size in order to capture the bulk of the looks. Second, because the attractive region for bat_1 takes a relatively long time to become dominant, there is a large likelihood (compared to the referential ambiguity case) that the regions

corresponding to the reference items, star and square, will capture looks during this phase. Thus, for garden paths, where participants initially focus on the irrelevant cat/bat, the framework predicts that the rate of fixation on reference items during disambiguation of a lexical ambiguity will be greater than the rate during disambiguation of a referential ambiguity.

1.5. Distinctive claims of Impulse Processing

Several features of the behavioral sequences just reviewed reveal distinctive claims of the Impulse Processing framework.

First, the prediction (which we will call "Prediction 1: *Local coherences in unambiguous contexts*") that the system should switch during unambiguous processing (e.g., hearing "Click on the snail that's beside the star" in the context of Fig. 2A) from looking at the first noun (e.g., snail), which we call the *target* (i.e., to be clicked on), to looking at the *reference item beside the target* (e.g., star beside the snail) is unexpected on a symbolic, information-driven view of parsing (Levy, 2008; Hale, 2001).

Information-driven approaches compute only structure that is consistent with global constraints, not structure that is merely consistent with bottom-up constraints. If we assume that information-driven models "rationally" direct eye movements toward referents which are consistent with the global structure implied by the compositional semantics of a sentence, then for a sentence like "Click on the snail that's beside the star," where there is a clear, unambiguous, globally preferred referent (i.e., snail), these models most naturally predict eye movements to that referent. Eye movements to lower-scale structures, like the reference item beside the target (e.g., star), provide a challenge to such information-driven models because local

semantic structures are not predicted to form on these accounts (especially given that the global structure is *already* preferred, making reinterpretation unnecessary. See Levy, Bicknell, Slattery, & Rayner, 2009). Moreover, the star and the square are in the same positions in the display on every trial, including the practice trials. Thus, a person viewing the scene will not glean useful new information by looking at the star beside the snail (i.e., reference item beside the target).[4] Impulse Processing claims that the looking happens nevertheless because of the self-organizing nature of the process: each interpretation must be formed anew from bottom-up input. Long years of practice fixating on stars when stars are mentioned imply that the system has a strong bottom-up tendency to perform this action. Note, however, that self-organization straddles a critical line between local and global structure. Word-level models of the VWP (e.g., Allopenna et al., 1998; Spivey, 2007) also predict the transition in looks from the target (snail) to the reference item (star), but they do so because they have *no* global sentence-level structure. What is critical is that, in addition to such low level effects (Prediction 1), impulse processing *also* makes predictions about global sentence-level structure (e.g., see Prediction 3) where word-level models fail.

One might suppose, though, that because of uncertainty about the behavior of the world, an information-driven system needs to double-check on information. Such an assumption could predict elevated looks to the reference item even in unambiguous context as a way of confirming the identity of the reference item. However, this assumption provides no reason for expecting more checking in the case of lexical plus referential ambiguity than referential ambiguity. Thus

[4] It is true that structure-based prediction models like Levy (2008) predict future linguistic structures that previous information anticipates. For example, they could plausibly predict "star" after hearing "Click on the snail that's beside the…" in a context where there is only one snail and it is beside the star. However, the model's prediction of the word does not imply that it specifies looks to the referent of the word. Under our assumption that the least assumption-laden extension of these models to looking behaviors is that they look at objects which are likely to provide useful information, given the current state of the interpretation at each point in time, these theories have little reason to predict looks to the reference items (though see discussion of a possible alternative view in the next paragraph below).

the prediction that looks to the reference items should be greater during transition out of a garden path (i.e., where participants look initially to the irrelevant, non-target cat/bat in the display) in lexical plus referential ambiguity than referential ambiguity further distinguishes Impulse Processing ("Prediction 2: *Local coherences in garden path contexts*"). This claim mirrors a general property of self-organizing systems: structure is built up recursively with autonomy of all the subunits. Therefore, the loss of functionality at one structural scale does not usually crash the system; instead, it causes the system to fall back onto a lower scale of coherence. This property may be one of the reasons why the self-organizing mechanisms found typically in living systems are more robust than corresponding digital computer models which are based on brittle information structures (Smolensky, 1988).

Third, on a classical information processing approach, ambiguity resolution is driven by symbolic computation. The information that drives the switch from one cat to the other cat in the resolution of the referential garden path is the same as the information that drives the switch from one bat to the other bat in the lexical plus referential garden path, namely, the current hypothesized focus object lacks the required property (e.g., being beside the star). Therefore, the classical approach does not predict a difference in the rate of transition under disambiguation between lexical and referential processing. We refer to the fact that Impulse Processing predicts a faster rate of transition in referential disambiguation as "Prediction 3: *Differential difficulty of recovery from a garden path*." This prediction of Impulse Processing stems from the assumption that, in referential ambiguity, the attractor basin for the alternative meaning is already present, but in lexical plus referential ambiguity, only the attractor basin for the competitor object is present at the point of disambiguation.

We turn now to investigating these predictions empirically.

2. Experiment 1: Ambiguity in the visual world paradigm

To test the predictions of Impulse Processing, we conducted a visual world experiment involving the referential and lexical plus referential ambiguities described in Section 1.4.

2.1. Methods

2.1.1. Participants

Twenty-five students from the University of Connecticut participated for course credit. All participants were native speakers of English with normal vision.

2.1.2. Materials

Lexical ambiguity (ambiguous vs. unambiguous target) was crossed with referential ambiguity (single- vs. two-target display) in a 2 x 2 design. Thirty-two lexically ambiguous homographic homophones (e.g., “bat”), containing at least two roughly equibiased noun interpretations, were selected (Gorfein, Viviani, & Leddo, 1982; Nelson, Mcevoy, Walling, & Wheeler, 1980; Wollen, Cox, Coahran, Shea, & Kirby, 1980). Thirty-two lexically unambiguous words (cat) were also chosen. Our unambiguous words were strongly biased toward a single meaning (given that few words are *truly* unambiguous), which we confirmed by sampling 20 occurrences of each word from the Corpus of Contemporary American English (COCA). At minimum, 70% of the occurrences of an unambiguous word were consistent with the meaning we employed ($M = 0.93$, $SD = 0.09$), given a fairly conservative metric in which we did not

count figurative uses (e.g., “I commend Oprah Winfrey for reaching out with her golden sword…” was not counted as an occurrence of the canonical weapon sense of “sword”). The Kučera and Francis (1967) frequency per million of unambiguous words (M = 52, SD = 94) and ambiguous words across all semantic senses (M = 50, SD = 65) were balanced, $t(62) = 0.09$, p = .93, as were the number of syllables in unambiguous words (M = 1.19, SD = 0.40) and ambiguous words (M = 1.16, SD = 0.37) words, $t(62) = 0.33$, p = .75.

In two-target displays, two images corresponding to the target noun were present on the screen, in addition to two unrelated distractors. For lexically unambiguous words, two identical pictures of the word were present (e.g., “cat” in Fig. 2a). For lexically ambiguous words, images corresponding to two different semantic senses of the word were present (e.g., “bat” in Fig. 2b). In single-target displays, the display contained just one image of the target noun (e.g., just one sense of a lexical ambiguity), and three unrelated distractors. Additionally, all visual displays contained two reference items (star and square). In two-target displays, the potential targets appeared beside different reference items.

Eight lists were assembled to rotate the lexically ambiguous and unambiguous targets through the one- and two-target displays, and to counterbalance the target interpretation for lexical ambiguities, the position of the target relative to the reference items (i.e., beside the star or square), and the relative orientation of the target and competitor in the two-target displays (vertical or diagonal). While items rotated between referential ambiguity conditions (e.g., one cat vs. two cat display; one bat vs. baseball and mammal bat display), items did not rotate between lexical ambiguity conditions (e.g., there was not a two baseball bat display). Participants were presented with 16 trials in each of the four conditions, and 48 filler trials. In 32 filler trials, a lexical or referential ambiguity was present in the display, but these images were not the target of

the trial. In 16 filler trials, an image of just one sense of a lexically ambiguous word was present, but it was not the target of the trial. Thus, the presence of a visual ambiguity (lexical or referential), or an image corresponding to a lexically ambiguous word, was not a cue to the target of a given trial. The experiment was divided into 16 blocks of seven trials each, with the four critical conditions, and 3 filler types, represented in each block. The order of trials within each block, and the order of targets across the experiment, was randomized. Participants saw each critical word only once, in a single condition.

For each of the 64 target nouns, two sentence frames were recorded (incorporating both the star and square as the reference item): “Click on the bat/cat that’s beside the square/star.” Thus, in the two-target displays the target remained ambiguous until the reference item was identified. Sentences were recorded by a native female speaker of American English using Praat software. Our visual stimuli were color photographs with white backgrounds (see Fig. 2).

2.1.3. Procedure

Participants listened to the recorded sentences over headphones while they viewed the visual displays on a computer monitor. Participants were instructed to use the mouse to click on the target image that was identified in each sentence. Participants’ eye movements were tracked with an R6 remote optics eye tracker with a head-tracking device (Applied Scientific Laboratories, MA, USA). A 500 ms preview of the display preceded the presentation of sentences, and the experiment began with 10 practice trials with feedback. The full experiment lasted approximately 45 minutes.

2.2. Results

2.2.1. Mixed logit modeling and growth curve analysis

Our analyses focused on the distinctive claims described in Section 1.5: Impulse Processing predicts local coherences in unambiguous contexts (Prediction 1), local coherences in garden path contexts (Prediction 2), and differences in the difficulty of recovery from a garden path (Prediction 3). We used both mixed logit modeling (Jaeger, 2008) and growth curve analysis (e.g., Mirman, Dixon, & Magnuson, 2008; Singer & Willett, 2003), where appropriate, to quantify differences in the trajectories of looks over time between relevant items and/or conditions. Because the proportions of fixations over time to the various items in the display have a complex form, we did not analyze looks across the entire sentence. Rather, we focused our analyses on smaller temporal windows, which were time locked to relevant events in the speech, where we predicted differences to occur. Given the typical 200 ms lag observed between eye movements and information in the language (e.g., Allopenna et al., 1998), our windows were also shifted forward in time by 200 ms from the relevant speech events.

For Prediction 1, in which we compare (non-independent) looks to items in the same display, we used mixed logit modeling, and a categorical looking measure. Our models included fixed effects of condition, and random effects of participant and item.

For Predictions 2 and 3, in which we compare looks to items in separate displays, we used growth curve analysis. Our analyses used orthogonal power polynomials to capture linear, quadratic, and in some cases, cubic and quartic, effects of time on the proportions of fixations to items in the visual display at the condition by subject level. Effects of condition were introduced onto the intercept, linear, and quadratic terms, as well as the cubic and quartic terms where appropriate. Models also included fixed effects of subject on each term. We quantified

differences between conditions by examining the effect of condition on each term in the growth curve models. With orthogonal power polynomials (where time is re-centered), condition impacts the curves as follows (see Mirman et al., 2008): an effect of condition on the intercept reflects a difference in the average height of the curves; an effect of condition on the linear term reflects a difference in the overall slope of the curves; and an effect of condition on the quadratic term reflects a difference in the rise and fall of the curves around the center. For analyses of target fixations (Prediction 3), where fixation curves grew monotonically to a peak, our growth curve models included intercept, linear, and quadratic terms. For analyses involving non-targets (Prediction 2), where fixation curves grew to a peak, then sank back downward, models also included cubic and quartic terms (thus capturing the three inflection points in the curves); an effect of condition on these terms reflects differences in the steepness of the curves near the inflection points.

2.2.2. Prediction 1: Local coherences in unambiguous contexts

Impulse Processing predicts elevated looking to the reference item beside the cat (or bat; e.g., star) when listeners hear "Click on the cat/bat that's beside the star," even when there is only one cat (or bat) in the display. To test this prediction, we compared looks to the reference item beside the target with looks to the distractors in visually unambiguous contexts (e.g., one cat, or one baseball bat and no mammalian bat, or visa versa). Average proportions of fixations in accurate trials are plotted for lexically unambiguous words (e.g., cat) in Fig. 7A, and for lexically ambiguous words (e.g., bat) in Fig. 7B. The plot extends from the mean onset of the target noun to the mean offset of the reference item. Trials were aligned at the onset of the reference item. Distractor fixations reflect mean looks to the non-target and non-competitor

items, excluding the reference items.

We used a temporal window that spanned between the mean onset and offset of the reference item (plus a 200 ms lag), and we coded trials (using a categorical outcome measure) as having a "look" to an item if participants looked to it at any point during the window (including "looks" which were launched prior to the onset of the window). We submitted the categorical looking measure to a mixed logit model with a fixed effect of item (reference item beside the target vs. distractor). For unambiguous words, the statistical model revealed a reliable fixed effect of item, coefficient (*reference item beside the target*) = 1.36, $SE = 0.18$, $p < .001$, with more looks to the reference item beside the target ($M = 0.40$, $SD = 0.49$) as compared to the distractors ($M = 0.16$, $SD = 0.36$). For ambiguous words, the statistical model also revealed a reliable fixed effect of item, coefficient (*reference item beside the target*) = 0.91, $SE = 0.15$, $p < .001$, with more looks to the reference item beside the target ($M = 0.54$, $SD = 0.50$) as compared to the distractors ($M = 0.33$, $SD = 0.47$). Both results are consistent with Prediction 1 of Impulse Processing.

2.2.3. Garden path trials

The remaining distinctive claims of Impulse Processing are concerned with *garden paths*. We defined *garden paths* as trials in which listeners made looks to the competitor but not the target within 500 ms of disambiguation, when the reference item was named. This *garden path* window spanned roughly the second half of the window between target offset and reference item onset (duration $M = 856$ ms), at which point we expected listeners to have fully processed the target word. Given the typical 200 ms lag observed between eye movements and information in the language (e.g., Allopenna et al., 1998), the *garden path* window extended from 300 ms

before reference noun onset to 200 ms after this point. Looks to the target, competitor, and reference item beside the target in *garden path* trials are plotted from target noun onset in Fig. 8 (aligned at reference item onset), with the *garden path* window shaded gray. During the 500 ms *garden path* window, among accurate trials in the visually ambiguous condition, listeners made at least one fixation to the competitor and no fixations to the target on 26% of trials (i.e., *garden paths*); listeners made at least one fixation to the target and no fixations to the competitor on 32% of trials; listeners made at least one fixation to both the target and competitor on 36% of trials; and listeners made fixations to neither the target nor the competitor on 6% of trials. There were no reliable differences in the average number of *garden paths* with referential ambiguities ($M = 3.88$, $SD = 2.31$) and lexical plus referential ambiguities ($M = 4.52$, $SD = 1.53$), $t(24) = 1.44$, $p = .16$. [5]

2.2.4. Prediction 2: Local coherences in garden path contexts

At disambiguation following a *garden path*, Impulse Processing predicts more looks to the reference item beside the target with lexical plus referential ambiguities as compared to referential ambiguities, given the greater influence of coherence at the simple noun phrase scale, and hence of the reference item, in the lexical plus referential ambiguity case. Looks to the reference item beside the target in accurate *garden path* trials, following the onset of the reference item, are plotted by ambiguity type in Fig. 9A. The plot extends from 200 ms following the onset of the reference item to 200 ms following the offset of the reference item. Trials were aligned at the onset of the reference item.

[5] Identical analyses for Predictions 2 and 3 were performed without isolating garden path trials. These analyses revealed a similar pattern of results with referential and lexical plus referential ambiguities. However, looks to the target were greater across conditions, and looks to the reference item beside the target were reduced across conditions, when trials in which participants looked to the target during the *garden path* window were included. Here, we only report results for the *garden path* analyses.

Growth curve fits, with effects of ambiguity type (referential or lexical plus referential) on the intercept, linear, quadratic, cubic, and quartic terms, are plotted as curves in Fig. 9A. Fixations to the reference item beside the target with referential and lexical plus referential ambiguities differed reliably in quadratic (*Estimate* = -0.20, *SE* = 0.03, $p < .0001$), and quartic terms (*Estimate* = 0.10, *SE* = 0.03, $p < .01$), but not in intercept (*Estimate* = -0.04, *SE* = 0.02, $p = .11$), linear (*Estimate* = -0.15, *SE* = 0.11, $p = .17$), and cubic terms (*Estimate* = -0.03, *SE* = 0.03, $p < .32$). Critically, the reliable effect on the quadratic captures the reliably steeper rise in the lexical plus referential curve past the temporal midpoint relative to referential ambiguities, consistent with Prediction 2 of Impulse Processing.

2.2.5 Prediction 3: Differential difficulty of recovery from a garden path

At disambiguation following a *garden path*, Impulse Processing predicts a faster transition from the competitor to the target with referential ambiguities as compared to lexical plus referential ambiguities, given the larger structural transition required in the lexical plus referential case (note the relative sizes of the target attractor basins for cat_1 and bat_1 in Figs. 5D and 6D). Looks to the target in accurate *garden path* trials are plotted in Fig. 9B, across the same window described above for Prediction 2. Additionally, growth curve fits, with effects of ambiguity type (referential or lexical plus referential) on the intercept, linear, and quadratic terms, are plotted as lines in Fig. 9B. Fixations to the target with referential and lexical plus referential ambiguities differed reliably in intercept (*Estimate* = 0.17, *SE* = 0.02, $p < .0001$) and linear terms (*Estimate* = 0.75, *SE* = 0.11, $p < .0001$), and there was a marginal effect on the quadratic term (*Estimate* = 0.06, *SE* = 0.03, $p = .09$). For Prediction 3, the critical difference is not simply on the intercept, which reflects a greater average rate of looking at the reference item

beside the target in referential ambiguities as compared to lexical plus referential ambiguities. Rather, the effect on the linear term reflects the greater overall steepness of the referential curve supports Prediction 3, indicating a quicker recovery from the garden path in purely referential ambiguities.

2.3. Summary

The mixed logit and growth curve analyses of eye movements in Experiment 1 confirmed the distinctive claims of Impulse Processing described in Section 1.5. In visually unambiguous contexts, listeners looked reliably more to the reference item beside the target as they heard "the cat/bat that's beside the…" as compared to looks to unrelated distractors. The shift in looks from the unambiguous target to the reference item beside the target is consistent with a locally coherent interpretation of the language (Prediction 1). In visually ambiguous contexts, during the transition from competitor to target after a *garden path*, listeners also looked more to the reference item beside the target with lexical plus referential ambiguities as compared to referential ambiguities, consistent with a locally coherent interpretation of the language in a garden path context (Prediction 2). Finally, fixation curves to the target following disambiguation were steeper with referential ambiguities as compared to lexical plus referential ambiguities, consistent with the larger structural transition required with lexical ambiguities (Prediction 3).

2.4. Discussion

The experimental findings were consistent with the predictions of Impulse Processing and thus provide some initial support for the framework. The experiments point to several additional empirical questions. We identify these here to indicate ways that alternative accounts could be ruled out, and ways that the current account could be falsified, thus suggesting some avenues for future empirical work in this area.

2.4.1. Visual similarity of the referentially ambiguous stimuli

The current design used identical images for the two referentially ambiguous items (e.g. cats). One might wonder if it is the visual similarity of these images that makes it easy to switch from one cat to the other cat. As a first check on this idea, we analyzed the looking behavior during the 500 ms preview interval (at the beginning of each trial, before any language was spoken) to see if there was any evidence for more frequent transitions between identical objects than between nonidentical objects. For this analysis, we constructed a 7x7 Markov transition probability matrix (7 equals the number of objects in the display, counting the fixation cross) for each subject and compared the averages of these matrices in the referential conditions to the average in the lexical plus referential conditions. This analysis yielded no evidence of greater tendency to switch between referential competitors (e.g. cat to cat) than between lexical competitors (e.g. bat to bat). However, this is a weak test because it is a short time window, and participants' biases in the presence of language may be different from their biases in its absence. A stronger test would be to use different images for the two referentially ambiguous items in the display and to employ a visual similarity metric (e.g., offline similarity ratings) to assure that there is no greater similarity in the referential than the lexical plus referential trials.[6] If the contrasts of Predictions 2 and 3 disappear when visual similarity is controlled for, the hypothesis

[6] We thank an anonymous reviewer for suggesting the possibility of this kind of control experiment.

that lexical representations are mediating the garden path recovery transitions will be ruled out. However, there will still be a need to explain the results of the current experiment; in this case, the self-organization account will still be a contender, but with the adjustment that the primary representational revision is occurring in a visual encoding space rather than a lexical one.

2.4.2. Differences between lexically ambiguous and lexically unambiguous words

We are referring to the nouns used in the straight referential ambiguity conditions as "unambiguous." In fact, very few words, including the "unambiguous" words in our design, are truly unambiguous. Our stimuli were chosen so that "unambiguous words" showed a strong asymmetry: the meanings depicted in our displays were highly preferred (these meanings were employed 93% of the time on average for unambiguous words in the COCA samples, and, as noted above, never less than 70% of the time). As noted in Materials, we also controlled the frequency of our items so that the overall frequency of ambiguous and unambiguous forms was equal. We chose this frequency control because we hypothesized that it would put the phonological processing of the forms being compared on an equal footing. However, one might argue, in parallel to the points raised about visual similarity above, that the lower frequencies of the ambiguous word semantics (each being roughly half the frequency of the unambiguous word semantics) may have made the transition easier in the referential ambiguity case. One way to rule this possibility out would be to use equibiased ambiguous words (rather than unambiguous words) as controls (e.g., while a lexical plus referential ambiguity condition would have a baseball bat and a mammalian bat present, a referential ambiguity condition would simply have two baseball bats, or two mammalian bats, present). We did not build this design because it was difficult to find enough reasonably frequent equibiased ambiguous words to run a within-

participants design under these conditions. One might consider a between participants design or an artificial lexicon approach (e.g. Magnuson, Tanenhaus, Aslin, & Dahan, 2003).

Finally, one could use lower-frequency words for the referential ambiguity case, thus equating semantic frequency rather than phonological frequency across the two conditions. Again, if the garden path recovery differences disappear in these cases, but the tendency to look at the reference items during recovery remains, then self-organization will still be a viable hypothesis, but the data will be suggesting that the locus of transition difficulty is in the frequency-properties rather than the semantic properties of lexical items.

2.4.3. A symbolic (non-self-organizing) account

It's possible that participants look at the reference items during garden path recovery because (1) they are (symbolically) reanalyzing their parse and (2) they do not know what the new interpretation is going to be, so they position their eyes at an intermediate location on the screen, thus minimizing the expected length of the next saccade. Such an account predicts that participants should return to the fixation cross during garden path recovery (the data do not support this claim), but it could be assumed that the fixation cross is not a very interesting object so they look at the star and the square which are nearby and more interesting. This symbolic account can be distinguished from the self-organization account via a design in which the positions of the objects, including the star and the square are randomized after every trial. The self-organization account still predicts looks to the mentioned reference item in unambiguous trials and the same garden path recovery difference, but the symbolic account predicts looks to objects near the center of the screen in this case.

We feel that these further empirical projects will be especially worth undertaking if the

self-organization approach can be shown to be formally coherent. To that end, we next describe a formal model which shows that the predictions plausibly follow from the assumptions.

3. Attractor network simulation

In this section, we describe an implementation of the Impulse Processing framework described in Section 1 using a connectionist attractor network (see Fig. 10). The details of the model's processing are described in Appendix A. The values of its free parameters are shown in Table 2. A Matlab implementation can be downloaded from http://solab.uconn.edu/People/Kukona/papers.html. Here we provide an overview, indicating how the main assumptions of the theory are implemented.

3.1. Architecture

The model consists of four layers of nodes: a phonological layer, a lexical semantics layer, a cross-word layer (i.e. a rudimentary kind of syntax layer), and an action-space layer (Fig. 10). All of these layers employ localist representations (one unit on per concept). We do not think localist representations will capture many subtle aspects of meaning and behavior, but they are a useful, simple case to consider first in developing the dynamical picture. There are feed-forward connections from each layer to the next and recurrent connections within the lexical semantic layer, the cross-word layer, and the action-space layer. The activations in the phonology layer correspond to (spoken) words. The activations in the lexical semantic layer correspond to word concepts. Activations in the cross-word layer allow the meanings of previous

words to affect the interpretations of subsequent words, useful for keeping track of syntactic dependencies. In the present study, for simplicity, we implement the syntactic constraints by fiat, specifying that after hearing a reference noun (e.g. “star” or “square”), the weights from the head noun’s concept in the cross word layer to the action layer should focus on just the relevant object, and the lexical semantic activations gravitate toward the corresponding concept. This means that our model does not implement a general theory of the encoding of syntactic dependencies in neural machinery (as do models like Elman, 1990, 1991; Tabor, 2000, 2003). Nevertheless, the cross-word layer does implement an account of the carry-over of information between lexical items that produces contrasts between garden path effects. Finally, the activation patterns in the action-space layer map onto fixation choices. Although, as indicated above, there were seven objects in each display, we were able to illustrate the important features of the dynamics by studying the interactions of units corresponding to just four of those objects. Additional units can be added to study interactions among more than four objects. Unlike some connectionist models in which particular units are statically anchored to properties in the world (e.g., “green at pixel 475”), we posited that nodes in the action-space layer correspond to objects in the visual display. In this sense, the assignment of interpretations to nodes in the action-space layer bears a resemblance to variable binding in the classical computational theory of mind (e.g., Marcus, 2001). We note that this assumption leaves a large explanatory gap with regard to the question of how abstract concepts are related to architecturally static neural tissue (as they seem to be in adult organisms). Since our purpose is not to try to solve this long-standing “symbol grounding problem” here, we make the simplifying assumption that concept “nodes” with appropriate connections are rapidly created each time a person encounters a new scene.

3.2. Processing

The activations of the nodes range from 0 to 1. The model is interpreted as fixating object *i* when its action-space vector, $a = [a_1, a_2, a_3, a_4]$, is closer to the *i*'th indexical bit vector (i.e. the vector with a 1 in the *i*'th position and 0's elsewhere) than to any other indexical bit vector in R^4.

The phonological layer has a node for each word in the model's vocabulary. Perception of a word is implemented as setting this node's activation to 1 and setting all other vocabulary activations to 0 for a number of timesteps corresponding to the duration of perception of the word.[7] We assume that, after detecting the sense of the last word in a phrase, the model detects the sense of the whole phrase as described in section 1.4 above (on SOPARSE dynamics).

Each phonological unit sends a "pulse" to the lexical semantics layer at every time step that it is activated. This pulse influences the shape of the attractor landscape in lexical semantic space. A pulse from a particular word strengthens an attractor corresponding to each meaning of the word. In this implementation we assume that "cat" has one meaning (and hence one lexical semantic attractor) and "bat" has two meanings (hence two lexical semantic attractors), and that "beside" gently encourages the listener to look at the reference item, and "star" and "square" encourage the listener to look at the star and square, respectively, but more weakly than do the experimentally varying items ("bat", "cat", etc) because "star" and "square" occur on every trial (See Table 1).[8]

The lexical semantic layer activations travel on the attractor landscape created by the

[7] A more realistic model could employ phonetic features that ramp-up gradually in activation, as in TRACE (McClelland & Elman, 1986).

[8] We tested a version of the model in which both "cat" and "bat" were lexically ambiguous, but only one sense of "cat" ever appeared in the displays. Except for the fact that the model sometimes activated the non-present sense of "cat" in the lexical semantic layer when it heard the word "cat", its behavior was similar to the current simulation and the same effects occurred..

phonological pulses. Noise is injected into their activations at every time step, creating small-scale random changes in the trajectory. In cases of lexically ambiguous words (e.g. "bat" in the present simulation), the noise pushes the state into an attractor corresponding to one or the other of the word's meanings.[9] The effect is that the lexical semantic units tend to hover briefly at low activation levels when a phonological stimulus is first presented, then one of the concepts corresponding to that phonology rockets to its maximum level and stays there for the duration of the word.

The lexical semantic units, in turn, send pulses that shape the attractor landscape of the cross-word layer. Unlike the lexical-semantic layer, which erases old structure when new structure comes in, the cross word layer captures the accumulation of information coming from sequenced lexical items. For example, if the words "cat beside (the) star" are presented, the cross-word units head, at first, toward an attractor corresponding to "cat" in general, then toward attractors corresponding to the reference items, then toward an attractor corresponding to "star," but from the direction of the "cat" attractor so the memory of the word "cat" is still influencing the state, and finally toward an attractor corresponding to the particular cat that's beside the star (and away from other meanings).

The cross-word units, in turn, shape the attractor landscape in the action-space layer. If an object corresponding to an activated cross-word concept is present in the display, then an attractor develops in the action-space corresponding to that concept. When the cross-word activations head away from concepts related to objects in the display, then the action-space attractors corresponding to those objects diminish in strength.

The action-space activations travel on the attractor landscape created by the cross-word

[9] A more sophisticated version of the model could employ feedback from visual perception, so that if the system were fixating on one kind of "bat", then the visual bias, instead of random noise, would push the lexical semantic representation in the direction of the matching interpretation.

pulses. As with the lexical semantic activations, noise is injected into the action-space activations at every time step. The noise plays an exploratory role for the visual system, causing the system to sometimes look at objects other than the one it is currently fixating on. The system's tendency to switch objects is a function, simultaneously, of the fixed noise magnitude and the attractor strengths, which vary in response to the incoming words.

We ran the simulation 20 times to model the collection of data from 20 different "listeners." Each simulation consisted of 16 trials with referential ambiguities, 16 trials with lexical plus referential ambiguities, and 16 trials with visually unambiguous contexts. We defined *garden paths* for referential and lexical plus referential ambiguities as trials in which the model was not fixating the target item between the midway point for the pulse for "beside" and the onset of the critical noun (cat/bat).

3.3. Results

Our analyses of the performance of the model closely parallel the analyses of the behavioral data in Experiment 1. As in Experiment 1, we used mixed logit modeling and growth curve analysis to quantify differences in the trajectories of simulated looks over time between relevant items and/or conditions. Our aim is to show that, with respect to the predictions of our theory, the model exhibits the same significant contrasts as the human data do.

3.3.1. Prediction 1: Local coherences in unambiguous contexts

Impulse Processing predicts greater looks to the reference item beside the target (e.g., star) during "beside the star," even in an unambiguous context, as compared to the unrelated

distractors. Simulated average proportions of fixations to the reference item beside the target and the unrelated distractor in the unambiguous context are plotted in Fig. 11, beginning at the preview window which preceded sentence onset (with looks to the target also plotted). For the analysis, we used a temporal window that spanned between the onset and offset of the pulse for the complex NP (using a categorical outcome measure) as having a "look" to an item if there was a look to it at any point during the window. We submitted the categorical looking measure to a mixed logit model with a fixed effect of item (reference item beside the target vs. distractor). The statistical model revealed reliably more looks (*Estimate* = 0.80, SE = 0.31, $p < .01$) to the reference item beside the target (M = 0.94, SD = 0.23) as compared to the distractor (M = 0.88, SD = 0.32).

3.3.2. Prediction 2: Local coherences in garden path contexts

At disambiguation following a *garden path*, Impulse Processing predicts greater looks to the reference item beside the target with lexical plus referential ambiguities as compared to referential ambiguities, given the greater reliance on local structure following a *garden path* in the lexical case. Simulated average proportions of fixations to the target, competitor, and reference item beside the target in *garden path* trials are plotted for referential and lexical plus referential ambiguities in Fig. 12, beginning at the preview window which preceded sentence onset. Simulated average proportions of fixations to the reference item beside the target are plotted by ambiguity type as symbols in Fig. 13A. The plot extends between the onset and offset of the pulse for the complex NP. Growth curve fits, with effects of ambiguity type (referential or lexical plus referential) on the intercept, linear, quadratic, cubic, and quartic terms, are plotted as curves in Fig. 13A. Fixations to the reference item beside the target with referential and lexical

plus referential ambiguities differed reliably in intercept (*Estimate* = -0.12, SE = 0.02, p < .0001), linear (*Estimate* = -0.78, SE = 0.16, p < .0001), quadratic (*Estimate* = -0.10, SE = 0.03, p < .01), cubic (*Estimate* = 0.15, SE = 0.03, p < .0001), and quartic (*Estimate* = 0.09, SE = 0.03, p < .01) terms. The reliable intercept difference reflects the greater mean height of the lexical plus referential curve as compared to the referential curve, consistent with Prediction 2, and the pattern in Experiment 1 (although not the precise curvature; see below).

3.3.3 Prediction 3: Differential difficulty of recovery from a garden path

At disambiguation following a *garden path*, Impulse Processing predicts a more rapid transition to the target with referential ambiguities as compared to lexical plus referential ambiguities, given the greater structural change required following a *garden path* in the lexical case. Simulated average proportions of fixations to the reference item beside the target are plotted by ambiguity type in Fig. 13B, across the window described for Prediction 2. Growth curve fits, with effects of ambiguity type (referential or lexical plus referential) on the intercept, linear, and quadratic terms, are plotted as lines in Fig. 13B. Fixations to the target with referential and lexical plus referential ambiguities showed a reliable effect in the linear term (*Estimate* = 0.26, SE = 0.04, p < .0001), a marginal effect in the intercept (*Estimate* = 0.04, SE = 0.02, p = .07), and a non-reliable effect in the quadratic term (*Estimate* = -0.04, SE = 0.04, p = .29). The marginal difference in the intercept term suggests a higher average rate of looking at the target in the referential case, and the reliable difference on the linear term indicates a faster recovery from the garden path in referential rather than lexical plus referential trials. Both of these results are consistent with Experiment 1, except that the model curves rapidly coincide after the syntax takes over (see Figure 12), while the human data show a much more gradual

convergence.

3.4. Discussion

The simulations show how the assumptions of the Impulse Processing framework can be formalized and how the three touted predictions follow from the assumptions under this formalization. In this section, we outline the main insights provided by the simulation, and we identify shortcomings of the model, suggesting avenues for future development.

The simulation shows how the assumptions outlined in intuitive terms at the beginning of the paper can give rise to data patterns very similar to the ones we observed in Experiment 1. In particular, the attractors at the conceptual and action-space levels can be implemented with positive on-diagonal elements and negative off-diagonal elements in the weight matrices for these layers. Although we did not implement learning of the encodings here, this kind of encoding arises via well-known Hebbian or Delta Rule learning processes in recurrent networks where the elements are mutually exclusive. Thus the simulation suggests the viability of attempting to derive the weights of these layers via learning (however it will be important to test this claim with an implemented learning model). In the cross-word layer, we employed a form of normalized recurrence (Spivey, 2007) because this kind of feedback among units provided an appropriate degree of spreading of the influence of information across time. Although we implemented syntactic constraints by fiat in the weights issuing from this layer in the current simulation, it is again possible that these weights could be learned, thus providing insight into how to predict real-time processing of words in a syntactic context.[10] In fact, we tried for a long

[10] The currently most successful attractor based model of syntactic processing, the Simple Recurrent Network (Elman 1990, 1991) can predict word sequencing data, but it provides no explicit model of within-word temporal

time to construct the model with only the phonological, lexical semantic, and action-space layers present. It turned out to be hard for one layer (the lexical semantic layer) to both quickly resolve the lexical ambiguity (bat vs. bat) and yet keep the memory of "cat" around while it was processing the disambiguating word ("star" or "square"), so that the lexical representation of "cat" could influence garden path recovery.[11] Although the success of this model does not prove that a syntax layer is necessary, it nevertheless suggests an effective formal way to integrate lexical and syntactic ambiguity resolution in a network attractor-model context.

It seems, at first, that there is a paradox in the juxtaposition of our empirical findings with the well-known results on multiple-access of ambiguous meanings in the literature, which we cited above (e.g., Swinney, 1979; Tanenhaus et al., 1979). Our effects depend on claiming that one meaning of an ambiguous word can be suppressed in such a way that there is a cost associated with resurrecting it later. The well-known results from the literature indicate that both meanings of an ambiguous word are initially enhanced. According to the model, however, there is no conflict between these findings. The difference is in the timing. Our experiment with human participants indicates that the suppression of the irrelevant meaning needs to have occurred by several words downstream from the onset of ambiguity (e.g., at "star" in "the bat that's beside the star"). On the other hand, the well-known findings on multiple access indicate that activation of the contextually irrelevant meaning is near baseline by three syllables (Swinney, 1979) or 600 ms (Tanenhaus et al., 1979) following the offset of ambiguity.[12] In fact,

---

processing in syntactic contexts, as the current model does.

[11] A more sophisticated version of the model would allow the reference word to impose the syntactic constraints in a context-sensitive way. In this case, the recovery from the lexical garden path will require recurrence (memory) in the phonological layer, so that the alternative lexical sense can be discovered.

[12] There is an important difference between those studies and ours in that in ours, the context did not resolve the ambiguity until the reference word ("star" or "square") arrived, but in theirs, the preceding context resolved the ambiguity, Nevertheless, our model claims, the ambiguity gets resolved quickly: small biases pushing the system one way or the other end up reinforcing each other through the recurrent connections causing the system to rapidly select an interpretation. This, then, is a prediction of the model: the referential/lexical ambiguity differences

a test of the model's lexical semantic activations showed elevated activations of the to-be-rejected sense relative to baseline for about 24 time steps, or about one-third of a word duration, starting from the onset of the ambiguous word.[13] In this sense, the model predicts both kinds of effects. The early multiple activation occurs because the (unbiased) phonological information positions the system on a ridge ("separatrix") between two attractor basins and drives the system along this ridge toward a saddle point which is associated with both senses (See Figure 1). Only when the noise pushes the system off the ridge does one meaning get the upper hand and shut the first meaning down (recall that we are treating noise as a stand-in for the biasing aspects of specific contexts.)

The current implementation has several shortcomings. For simplicity of design and analysis, we used localist encodings of phonology, lexical semantics, cross-word representations, and action-space representations. Distributed encodings would allow the model to capture representational similarity effects (see further discussion in Section 4.2 below). The model is sensitive to the values of its free parameters: the strengths of the phonological pulses, the rates of convergence of the recurrent subnetworks, the initial activations of recurrent units, the magnitude of the noise, the word durations (See Tables 1 and 2). If these values are too small or too large, the attractors fail to develop and the activations either stay at baseline levels or the model gets stuck in the first attractor basin it falls into. We find it encouraging that the model only shows a few qualitatively different kinds of behavior. Nevertheless, it would be desirable to set the parameters in a principled way. For the phonological pulse magnitudes, this may be achieved by

---

observed here should not occur or should be reduced in magnitude if the syntactic disambiguating information arrives while both senses of lexically ambiguous words are still activated.

[13] The findings with humans in the literature generally detect the multiple activations starting from the offset of ambiguous words. We suspect that this difference from the model stems from the fact that natural language spoken words only become uniquely recognizable after several phonemes have been spoken and because the process of activating a complex natural language representation takes more time than activating the very simple localist representations employed here. Further probing of a more realistic version of the model (e.g. with TRACE-like inputs) is a natural next step in this regard.

learning. Finally, it is desirable to implement a principled encoding of syntactic structure, including its recursive aspects.

4. General Discussion

4.1. Summary

We have outlined a solution to the problem of information integration in support of action, called Impulse Processing, which models eye movements in the VWP according to principles of self-organization. Central to our proposal is the claim that structure at one scale is built upon structure at lower scales: thus, Impulse Processing predicts local coherence phenomenon in the VWP, which we have confirmed in an empirical study and an implemented model.

4.2. Relation to prior work

As we noted in the introduction, a number of dynamical systems proposals have been advanced which model the integration of spoken and visual information in the VWP. These proposals form an essential foundation for our project, and we incorporate many of their insights into the present work: for example, continuous activation values, feedback connections, and melding of different sources of constraint in an interactive activation framework. Now, through careful comparison of our account with these projects, we clarify some of the ways that the current work contrasts with and/or extends these approaches.

Allopenna et al. (1998) examined the fine-grained temporal dynamics of spoken word recognition in the VWP. In a behavioral study, they instructed people to pick up a target item in a visual array. They found that the amount of looking to each item in the array (i.e., to the target itself, and to items sharing a phonological onset or rhyme) was predicted by lexical activations in TRACE (McClelland & Elman, 1986), an attractor model of spoken-word recognition that takes spoken, but *not* visual information, as its input. Allopenna et al. (1998) also observed a kind of local coherence in eye-movement behaviors: looks to competitor items that shared a rhyme, but not an onset (e.g., *beaker – speaker*). Despite the clear difference in onsets, listeners were more likely to look to rhymes than to unrelated distractors that shared no phonological overlap with the target (e.g., *carriage*). This result is closely related to the reference item looks we observed, suggesting the formation of localized structure, as predicted by self-organization.

However, Allopenna et al. (1998)'s model is only partially constrained by the visual information. For purposes of simulating looks (using the Luce choice rule, given activation across the entire lexicon), they restricted their analysis to only those nodes in TRACE that corresponded to objects in the visual display, thus emphasizing the dynamics of the linguistic portion of the signal, and not the visual portion (see also Tanenhaus et al., 2000). Spivey (2007) modified this TRACE-based approach to allow feedback from the visual component to influence the dynamics: he simulated Allopenna et al.'s behavioral findings with a recurrent network with three layers: a lexical layer, with word nodes fed raw activation levels from TRACE, a visual layer, with object nodes activated when an object was present in the display, and an integration layer, which connected the lexical and visual layers. Like Spivey (2007), we have allowed both visual and verbal information to modulate the dynamics. However, neither Spivey nor Allopenna et al.'s approaches handle structure above the lexical level: note that if TRACE were fed the

words in a phrase like “The cat that’s beside the star” in succession, at the last word in the sentence, it would continue looking to the star, although behaviorally listeners return to the cat implied by the larger phrasal structure. The model we have implemented, takes a step toward clarifying how dynamical models like these might approach the challenge of integrating syntax-level information across words in sentences.

The dynamical systems approach we have proposed, however, is not the first to address sentence processing in the visual world. Mayberry et al. (2009), for example, simulated *anticipatory* looking behaviors (e.g., Altmann & Kamide, 1999, 2007, 2009; Knoeferle & Crocker, 2006, 2007) in the visual world using an augmented simple recurrent network (SRN; Elman, 1990) that processed sentence-level input. Behaviorally, listeners robustly use information from the language and the visual context to anticipate upcoming linguistic referents. For example, Altmann and Kamide (1999) showed that listeners hearing “The boy will eat the…” were more likely to look at edible objects like a cake, as compared to inedible objects like a ball or truck, as predicted by the verb in the sentence. Accordingly, Mayberry et al. (2009) used a multi-layer network, with recurrence in the hidden layer, to predict role assignment of arguments in a scene, given the visual and linguistic contexts. They presented their model (C_{IA}-N_{ET}) with both word-by-word (German-based) sentences, as well as visual contexts depicting scenes. The task of their model was to activate a filler-role representation of the event within the scene that the language referred to (each scene was assumed to contain two possible events). Consistent with the behavioral data, the model can use information from the linguistic signal up through the verb (e.g., “The princess is painting the...”), and information from the visual signal (e.g., a pirate who is washing a princess who is painting a fencer), to anticipate the direct object predicted by the union of the language and the scene (e.g., fencer). Also consistent with the

behavioral data, the model favors visual information over stereotyped linguistic information when they conflict: given "The pilot was spied on by the…," for example, in a visual context depicting a wizard who is spying on a pilot who is being fed by a detective, the model anticipated the wizard, consistent with the visual context, and not the detective, a stereotypical and predictable agent of the verb spy based on the language. Thus, the model is highly sensitive to the relationship between sentence-level structure in the language, and interactions among different items in a visual context.

A very appealing property of the Mayberry et al. (2009) model is that, as an SRN, it learns to relate visual and linguistic information, and to use this information to focus looks appropriately. Although the network we have described is not a learning model, it is nevertheless compatible with such an approach. We assume, for example, that the linguistic pulses that modulate the network's action landscape reflect learned associations between stereotyped (eye-movement) behaviors, linguistic contexts, and visual contexts. Consistent with robust behavioral findings, Mayberry et al.'s model also acts anticipatorily. In this regard, we also found evidence for a kind of anticipation: listeners tended to fixate the reference items as they heard "beside," before "star" or "square" was named in the sentence, a behavior exhibited by our model. Our model demonstrated this anticipatory behavior because of the semantics assigned to each pulse. The effect of the pulse for "beside," for example, is to deepen the attractor basins for the reference items, which are beside the item usually being fixated at this point in the trial. This definition of the effect of beside is a pure stipulation in our model, unlike in Mayberry et al.'s. We think it plausible that experience with the word "beside" induces a context-independent tendency to look at objects beside the object currently being looked at.[14] This assumption about

[14] We also implemented a version of the model in which the response to beside was contingent upon which object the model was currently looking at. This version produced the same pattern of results as those reported here.

adult behavior is consistent with a learning paradigm that drives an organism toward helpfully exploratory behavior (e.g., Oudeyer, Kaplan, & Hafner, 2007; Sutton & Barto, 1998).

The Mayberry et al. (2009) model, however, is limited in a number of ways. First, the output of the model consists of looks to holistic scenes: for example, the model might activate a scene involving a princess who is painting a fencer, rather than a pirate who is washing a princess. However, the model does not generate looks to individual items within each scene (e.g., princess versus fencer). At a finer level of analysis, listeners do look to individual items in the display. Additionally, Mayberry et al.'s model makes the assumption that listeners have a rich mental representation about the relationships between all items in a display: for example, their model assumes that listeners do not simply know that a pirate, princess, and fencer are present; they also know precisely what each item is doing to all of the other items. However, there is evidence to suggest that listeners often store only a minimal amount of information about items in a visual context (Ballard, Hayhoe, Pook, & Rao, 1997), according to the task at hand. Additionally, there is the problem of understanding how listeners could grapple with very rich visual contexts, in which items are interacting in an infinite number of ways, like we might encounter in the real (visual) world. By employing a combinatorial generation mechanism – the looking behavior in Impulse Processing arises from the combination of pulses created by the word sequence and the context – our model is situated to exhibit an appropriately open-ended variety of behaviors.

Roy and Mukherjee (2005) have also addressed the integration of sentence-level and visual information in VWP-like settings. Their model (Fuse) is a probabilistic rule model which is trained to interpret referential expressions about items in a visual context, and to find the items identified by the language. Fuse processes sentences incrementally, by generating a probability

distribution across the items in a visual context, based on their fit with the language. As each new word is processed, Fuse modulates the distribution of probabilities across the visual display. As Fuse processes an utterance like "The large green block in the far right beneath the yellow block and the red block," for example, it first allocates higher probabilities to large blocks in the display ("The large…"), then to large green blocks ("…green…"), then to large green blocks on the right ("…block in the far right…"), and so forth.

Like Mayberry et al. (2009)'s $C_{IA}N_{ET}$, Roy and Mukherjee (2005)'s Fuse has the desirable trait that it learns to perform its task: it is trained on corpora of real language spoken by real people about real visual contexts. Like our model, Fuse also "interprets" complex referring expressions. Unlike our model, however, and the others we have discussed, Fuse is not explicitly a model of eye movements: Roy and Mukherjee interpret the model's probability distributions as distributions over attentional foci. If one assumes that attentional foci correspond to fixation locations, then the model can be interpreted as a model of eye movements. Under this assumption, the model makes incorrect predictions about the fixation patterns associated with complex noun phrases: to arrive at an interpretation, the model divides the complex noun phrase into sub-phrases, such that one phrase identifies the target (e.g., "The large green block…"), and the other phrase identifies landmark items (akin to our reference items) that serve to disambiguate the target (e.g., "beneath the yellow block and the red block"). As the model begins to process the noun phrase identifying the landmark item, its attention shifts to the landmark item in the visual context, rather than remaining on the target item (e.g., higher probabilities are allocated toward yellow and red blocks which are above a large green block, rather than toward a large green block which is below a yellow and red block). Thus, in processing the last word in "The cat beside the star…," Fuse would allocate a higher probability to the star in the display,

although behaviorally listeners return to the target.

Interesetingly, although Roy & Mukherjee do not address the issue in their discussion, Fuse appears to exhibit local coherence behaviors. Roy & Mukherjee (2005) plot the distribution of probabilities to visual objects during the processing of "The large green block…" in a visual context containing large green blocks and small green blocks. The probability bars accompanying the figure suggest that while their model allocated the highest probabilities to large green blocks, it also allocated elevated probabilities to *small* green blocks, which were at least consistent in color with the language, as compared to small red blocks, for example. Like looks to rhymes (e.g., Allopenna et al., 1998), and the looks to reference items that we observed, this suggests the formation of local structure despite incongruence with the global context. This behavior of the model is likely a consequence of the way individual words impinge on the system's probability distributions. Each word (e.g., "green") is mapped to consistent items in the visual context, and these context-independent probabilities are multiplied together as a sentence is processed. While this independence assumption about the effect of words on the system recapitulates the notion of bottom-up priority, it is nevertheless limited, in so much as it cannot naturally handle more complex expressions (e.g., the model does not shift attention from the landmark to the target after hearing the modifying clause of a complex noun phrase; instead, the reference of the whole phrase is computed independently of the locus of attention).

Our dynamical systems approach is generally consistent with the feature-based approach of Altmann and Kamide (2007, 2009). These authors are especially concerned with results from the visual world that indicate that listeners do not simply look to items in a visual context as they are named, but that listeners also look to items which are related in any number of ways with the language. Their theoretical approach provides a rich account of a large number of results from

the VWP: for example, looks to a rope on hearing "snake" as a consequence of physical featural overlap (Dahan & Tanenhaus, 2005), and looks to a trumpet on hearing "piano" as a consequence of categorical featural overlap (Huettig & Altmann, 2005; see also Yee & Sedivy, 2006). Their proposal assumes that entities in the language (e.g., words) and in the visual world (e.g., objects or images) activate corresponding representations in our *mental world*: mental representations which are featural in composition, and which include information about the form, function, associations, and so forth, of the words and visual objects being processed. Their proposal assumes that a visual representation receives a *boost* in activation when it shares features with a linguistic representation, increasing the likelihood of a saccade to that object in the display. Such effects can be predicted by recurrent networks like the one we describe here that use distributed codes: the featural encodings of distinct entities overlap. Consequently, if one object gets activated, it will partially activate other objects that share features with it. This kind of behavior is well-documented in recurrent networks, closely related to ours, with feature overlap (e.g. Kawamoto, 1993; McClelland & Kawamoto, 1986; McRae, de Sa, & Seidenberg, 1997; Harm & Seidenberg, 2001, 2004). For simplicity, and because our focus is not on feature overlap effects, we have used localist encodings in the current model, and thus do not consider fine-grained semantic and physical feature overlap, although this is not a necessary restriction. An important question for future modeling research in this area is whether the same sentence processing dynamics can be sustained when more complex distributed codes are used in the recurrent layers.

4.3. The abstractness of language

As we noted in the introduction, one may well wonder if a theory that directly connects language and action can handle abstract uses of language. What about situations where an action specified by language is not immediately carried out (e.g., hearing on the telephone, "Could you pick up a quart of milk on the way home?") or where the language evinces a mental change that does not require any physical response (e.g., "You see, the morning star and the evening star are one and the same object!")?

An in-depth discussion of these issues is not within the scope of this paper. Nevertheless, we note that the approach we have outlined has a reasonable answer to this question: In Impulse Processing, perceptions modify a landscape that specifies actions. But the shape of the landscape at any point in time, and the location of the system on the landscape, is determined by the perceiver's cumulative interaction with the environment. So the obvious action specified by a particular piece of language (e.g. "pick up a quart of milk") may not dominate the behavior at the moment the utterance occurs. It is useful to think about this issue in terms of contrasting structural scales. In the model discussed above, we considered examples in which the referent of a modifying noun (e.g. "star") attracted some looks during the utterance of a complex noun phrase (e.g., "the cat that's beside the star") during the time when the modifying noun was being spoken. However, this tendency was modulated by the strength of the attractor of the head noun, as indicated by the comparison of referential (i.e., cat) and lexical plus referential (i.e., bat) ambiguities. Since the scale of the head noun attractor was relatively large compared to the scale of the modifying noun, the influence of the modifying noun on the looking behavior was minimal (e.g., given the large attractor basin for cat_1 in the *garden path* case; see Fig. 5).

Relatedly, we hypothesize that, when someone hears a statement that refers to events associated with a remote time, as in the milk example, then, although there *is* an effect on the

action landscape of the hearer at the time of perception of the request, this effect is a relatively small deformation. It will cause some minimal activation of motor-related neural pathways associated with the process of purchasing milk, but it will not cause the person to leap up and begin milk-purchasing activities at the moment. This is because the current situation constraints cause the magnitude of the deformation of the prospective action to be minimized in relation to the magnitude of deformations related to the task at hand (in this case, talking on the phone). In the milk purchasing case, we assume that the deformation caused by the request, though small, sticks around in a portion of the mental space of the person connected with her plans for traveling home, and, at the appropriate point in the journey, the small deformation becomes enlarged to the point where it causes appropriate action (e.g., driving into the parking lot of a store that sells milk, etc.) Similarly, in the case of an abstract mental revision, like learning the common identity of the morning star and the evening star, Impulse Processing claims that the comprehender's landscape for action is revised at a small scale when the utterance occurs, and this deformation becomes enlarged later at points where it becomes relevant (e.g., acts of drawing a diagram of the solar system).

In the restricted domain of syntactic comprehension on a seconds-long timescale, where words that occur at one point in time constrain the possibilities for words at certain future points, Tabor (2000, 2003, 2009) discuss a neural activation framework, called *fractal grammars*, that works according to the scale manipulation principle just outlined. Although the attractors in this framework do not modulate behavior at the millisecond timescale appropriate for modeling eye-tracking data, the framework nevertheless suggests that neural memory manipulation could be a matter of scale manipulation.

This scale-manipulation view is generally consistent with accounts which ground abstract

conceptual knowledge in perceptual and motor systems (e.g., Barsalou, 1999; Barsalou, Simmons, Barbey, & Wilson, 2003), and is supported by data on neural responses to language which indicate neural activation in regions relevant to the action associated with the language in the absence of overt muscular responses (e.g., Moody & Gennari, 2009; see also Pulvermüller. 2005). We hypothesize that these neural responses are weak versions of the activation dynamics that would take place if the person actually engaged in the action described by the language. The view is also consistent with the finding that when people are asked to imagine a described scene while viewing a blank wall, the scan paths of their eyes resemble those they would produce if they were actually viewing the scene (e.g., Spivey & Geng, 2001). In this case, the blank wall provides such a weak global context that the eye movements that are naturally associated with the words are not suppressed and can be detected. This view also helps explain how the strongly input-driven self-organization approach is compatible with the finding that different task constraints produce very distinct scan-path characteristics for the same image (e.g., Yarbus, 1967) – the task constraints amplify different attractor basins.

We suggest, then, that Impulse Processing is a sufficiently flexible framework to make headway on the problem of integrating multiple, loosely coordinated information sources, and that the framework makes distinctive, empirically justified predictions, and that it has a plausible take on the well-known challenges of handling concrete and abstract language in a common framework.

Appendix A

(An implementation can be downloaded from http://solab.uconn.edu/People/Kukona/papers.html)

The values of the free parameters in the model are shown in Tables 1 and 2.

The activation dynamics in the concept layer are implemented by the following equations:

$$wll = (1 + linhib) \cdot ppulse_t - linhib \quad [1]$$

$$netl_i = \sum_j wll_{ij} \cdot l_j \quad [2]$$

$$\Delta l_i = dtl \cdot netl_i \cdot l_i \cdot (1 - l_i) + lexnoise \quad [3]$$

where $ppulse_t$ is the lexical semantic weight matrix specified by the current phonological input, *wll* is a scaled and recentered version of that matrix which produces attractors of appropriate strength, *linhib* is the strength of inhibition among mutually inconsistent lexical semantic units, l_i is the activation of the ith lexical semantic unit, and *dtl* is the base rate of change of the lexical semantic units. The lexical noise always distorts the activation toward the middle of lexical semantic space:

$$lexnoise = \mu mag \cdot -sign(l_i - 0.5) \cdot \mu \quad [4]$$

where μ is a uniform noise distribution on [0, 1], and μmag scales the noise. Equations [1]-[4], have the effect of creating attractive regions in the concept space corresponding to the meaning(s) of the phonology being activated.[15]

[15] Equation [3] is a single equation that combines the excitatory and inhibitory equations used in other language processing models like Interactive Activation (McClelland & Rumelhart, 1981), and TRACE (McClelland & Elman, 1986), but without the decay terms used in those models. It

The crossword activations derive from the lexical semantic activations as follows:

$$netc_i = \sum_j wcl_{ij} \cdot l_j \quad [5]$$

$$\Delta c_i = dtc \cdot netc_i \quad [6]$$

$$c_i(t + dtc) = \frac{c_i(t)}{\sum_j c_j(t)} \quad [7]$$

where *wcl* is the matrix of weights from the lexical semantic layer to the cross-word layer, c_i is the activation of the *i'*th cross-word unit, and *dtc* specifies the base growth rate of cross-word activations. Equation [7] implements a form of normalized recurrence similar to the normalized recurrence of Spivey (1996, 2007). The normalization causes the influence of preceding information to fade slowly when new information comes in, thus producing cross-word interactions.

The cross-word activations specify the self-weights of the action-space dynamics:

$$waa_{ii} = \sum_j wac_{ij} c_j \quad [8]$$

where *wac* is the weight matrix from cross-word space to action-space and *waa* specifies the action space recurrent connections. waa_{ij} = *-ainhib* for $i \neq j$, where *ainhib* is positive. The action-space dynamics are similar in form to the lexical-semantic dynamics:

$$neta_i = \sum_j waa_{ij} \cdot a_j \quad [9]$$

$$\Delta a_i = dta \cdot neta_i \cdot a_i \cdot (1 - a_i) + fixnoise \quad [10]$$

where a_i is the activation of the *i*'th action-space unit and *dta* is the base growth rate of action-space activations. The fixation noise is given by:

---

creates a sigmoidal response in the lexical semantic units as a function of net input.

$$fixnoise = \eta mag \cdot -sign(a_i - 0.5) \cdot \eta \quad [11]$$

where η is a uniform noise distribution on [0, 1], and ηmag scales the noise.

The activation update implied by equations [3], [6], and [10] takes place via an update computation of the form:

$$x(t + dt) = x(t) + \Delta x \quad [12]$$

The activations of the lexical semantic units, the cross-word units, and the action-space units are all initialized to the same small value, *actinit*, in the range (0, 1). In the limit as the growth rates, *dtl*, *dtc*, and *dta* go to 0, the activation change equations [3], [6]/[7], and [10] approximate differential equations which are invariant on the unit hypercube in the respective spaces. Since we intend the discrete equations as approximations of the continuous equations, and we assume that the noise models a physical process which is subject to the same constraint, we restrict the activations to this range, moving the value of any given dimension back inside the space if it ever strays out.

References

Abraham, R.H., & Shaw, C.D. (1992). *Dynamics: The geometry of behavior*. Redwood City: Addison-Wesley.

Allopenna, P., 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., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: eye-movements and mental representation. *Cognition, 111*, 55-71.

Altmann, G.T.M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*, 502-518.

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

Ballard, D.H., Hayhoe, M.M., Pook, P.K., & Rao, R.P.N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 723–742.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences, 22*, 577-660.

Barsalou, L.W., Simmons, W.K., Barbey, A., & Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.

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.

Crutchfield, J.P., & Young, K. (1990). Computation at the Onset of Chaos. In W.H. Zurek (Ed.),

Complexity, Entropy, and the Physics of Information (pp. 223-270). Redwood City: Addison-Wesley.

Dahan, D., & Tanenhaus, M.K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12*, 453-459.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-224.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry, 29*, 1-36.

Gorfein, D.S., Viviani, J.M., Leddo, J. (1982). Norms as a tool for the study of homography. *Memory & Cognition, 10*, 503-509.

Haken, H. (2004). *Synergetic Computers and Cognition, 2nd Enlarged Edition*. Berlin: Springer.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL, Volume 2* (pp. 159-166).

Harm, M. W., & Seidenberg, M. S. (2001). Are There Orthographic Impairments In Phonological Dyslexia? *Cognitive Neuropsychology, 18*, 71-92.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review, 111*, 662-720.

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*, 23-32.

Jacobs, D., & Michaels, C.F. (2007). Direct learning. *Ecological Psychology, 19*, 321-349.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not)

and towards Logit Mixed Models. *Journal of Memory and Language. 59*, 434–446.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32*, 474-516.

Kelso, J.A.S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C.A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 812-832.

Kessler, M.A., & Werner, B.T. (2003). Self-Organization of Sorted Patterned Ground. *Science, 299*, 380-383.

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.

Konieczny, L., Müller, D., Hachmann, W., Schwarzkopf, S., & Wolfer, S. (2009). Local syntactic coherence interpretation. Evidence from a visual world study. Paper presented at the 31st Annual Conference of the Cognitive Science Society.

Koschmider, E.L. (1993). *Bénard Cells and Taylor Vortices*. Cambridge: Cambridge University Press.

Kučera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Kukona, A., Fang, S., Aicher, K., Chen, H., & Magnuson, J. S. (in press). The time course of anticipatory constraint integration. *Cognition*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126-1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner. K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086-21090.

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

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Mayberry, M.R., Crocker, M.W., & Knoeferle, P. (2009). Learning to Attend: A Connectionist Model of Situated Language Comprehension. *Cognitive Science, 33*, 794-838.

McClelland, J. L. & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology, 18*, 1-86.

McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings. *Psychological Review, 88*, 375-407.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*, 99-130.

Mirman, D., Dixon, J.A., and Magnuson, J.S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*, 475-494.

Moody, C.L., & Gennari, S.P. (2010). Effects of implied physical effort in sensory-motor and pre-frontal cortex during language comprehension. *NeuroImage, 49*, 782-793.

Nelson, D.L., Mcevoy, C.L., Walling, J.R., & Wheeler, J. W. (1980). The University of South Florida homograph norms. *Behavior Research Methods & Instrumentation, 12*, 16-37.

Oudeyer, P., Kaplan, F., & Hafner, V. (2007) Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation, 11*, 265-286.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*, 576-582

Raczaszek-Leonardi, J., Shapiro, L.P., Tuller, B. & Kelso, J.A.S. (2008). Activating Basic Category Examplars in Sentence Contexts: A Dynamical Account. *Journal of Psycholinguistic Research, 37,* 87–113.

Rodd, J.M., Gaskell, M.G., & Marslen-Wilson, W.D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*, 245–266.

Rodd, J.M., Gaskell, M.G., & Marslen-Wilson, W.D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*, 89-104.

Roy, D., & Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language, 19*, 227-248.

Seidenberg, M.S., Tanenhaus, M.K., Leiman, J.M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based

processing. *Cognitive Psychology, 14*, 489-537.

Simpson, G.B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 28-39.

Simpson, G.B., & Kang, H. (1994). Inhibitory processes in the recognition of homograph meanings. In D. Dagenbach & T.H. Carr (Eds), *Inhibitory processes in attention, memory and language* (pp. 359-381). San Diego: Academic Press.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Spivey, M.J. (2007). *The continuity of mind*. New York: Oxford University Press.

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

Strogatz, S.H. (1994). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Cambridge: Perseus Books.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning*. Cambridge: MIT Press.

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

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks, 17*, 41-56.

Tabor, W. (2002). The Value of Symbolic Computation. *Ecological Psychology, 14*, 21-51.

Tabor, W. (2003). Learning Exponential State-Growth Languages by Hill Climbing. *IEEE Transactions on Neural Networks, 14*, 444-446.

Tabor, W. (2006). A unifed, self-organizing model of garden path phenomena, center-embedding phenomena, and interference effects. Paper presented at the 19th Annual CUNY Conference on Human Sentence Processing, New York, NY.

Tabor, W. (2009). Dynamical Insight into Structure in Connectionist Models. In J.P. Spencer, M.S.C. Thomas, & J.L. McClelland (Eds.), *Toward a Unified Theory of Development. Connectionism and Dynamic Systems Theory Re-considered*, (pp. 165-181). Oxford: Oxford University Press.

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.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*, 355-370.

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

Van Dyke, J.A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 407-430.

Waugh, N.C., & Norman, D.A. (1965). Primary memory. *Psychological Review, 72*, 89-104.

Wollen, K.A., Cox, S.D., Coahran, M.M., Shea, D.S., & Kirby, R.F. (1980). Frequency of occurrence and concreteness ratings of homograph meanings. *Behavior Research Methods & Instrumentation, 12*, 8-15.

Yarbus, A.L. (1967). *Eye Movements and Vision*. New York: Plenum.

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

Zhabotinsky, A.M. (1991). A history of chemical oscillations and waves. *Chaos, 1*, 379-386.

Acknowledgements

Preliminary results from Experiment 1 and Simulation 1 were presented at the 21st Annual CUNY Conference on Human Sentence Processing in Chapel Hill, North Carolina (March, 2008). We thank Jim Magnuson for allowing us the use of his lab, and for his helpful comments on an earlier version of this manuscript. We also thank four reviewers for their very helpful feedback. We gratefully acknowledge support from NICHD grant HD40353 to Haskins Laboratories and NICHD Predoctoral NRSA HD060414 to AK.

Figure captions

Figure 1. A three-dimensional rendering of a dynamical potential surface with attractor basins, and a saddle point at point S. An example trajectory of the system is depicted. Four regions corresponding to objects are depicted on the two-dimensional topographic portrait on the floor of the plot. When the system is above the region corresponding to a particular object, it fixates on that object.

Figure 2. Visual world displays depicting (A) a referential ambiguity with two cats, and (B) a lexical plus referential ambiguity with a baseball bat and a mammalian bat. Reference items (the star and square) were centered between the top and bottom items for purposes of disambiguating the target. The star and square appeared in the same positions on every trial. Listeners were instructed to "Click on the cat/bat that's beside the star/square."

Figure 3. Diagram illustrating changes in the attraction strength of each object during the processing of "Click on the snail that's beside the star" in the visual context of one snail and unrelated distractors (e.g., glove).

Figure 4. Processing of the complex noun phrase "the snail beside the star" by SOPARSE. Solid lines indicate strong links between nodes, and dashed lines indicated weak links.

Figure 5. Diagram illustrating changes in the attraction strength of each object during the processing of "Click on the cat that's beside the star" in the visual context of one cat beside a star (cat_1: target), and another cat beside a square (cat_2: competitor).

Figure 6. Diagram illustrating changes in the attraction strength of each object during the processing of "Click on the bat that's beside the star" in the visual context of a baseball bat beside a star (bat_1: target), and a mammalian bat beside a square (bat_2: competitor). The figures depict a *garden path* trial, in which the system fixates the competitor until the offset of the first noun.

Figure 7. Average proportions of fixations to the target, reference item beside the target, and (mean) distractors with lexically unambiguous (A; e.g., *cat*) and ambiguous (B: e.g., *bat*) words in visually unambiguous contexts (e.g., one cat, or one baseball bat and no mammalian bat, or visa versa) in Experiment 1. Standard error bars are shown, along with mean onsets and offsets for the target noun and reference item (dashed lines).

Figure 8. Average proportions of fixations to the target (△), competitor (▽), and reference item beside the target (□), with referential (A) and lexical plus referential (B) ambiguities in *garden path* trials in Experiment 1, where listeners looked to the competitor but not the target within 500 ms of reference onset (+ 200 ms). Standard error bars are shown, along with mean onsets and offsets for the target noun and reference item (dashed lines).

Figure 9. Average proportions of fixations to the reference item beside the target (A) and target (B) in *garden path* trials in Experiment 1. Black symbols indicate referential ambiguities (e.g., two cats), and white symbols indicate lexical plus referential ambiguities (e.g., baseball bat and

mammalian bat). Standard error bars are shown, along with the mean reference item offset (dashed line). Curves indicate model fits from the growth curve analyses for Predictions 2 (A) and 3 (B).

Figure 10. Architecture diagram for the artificial neural network used in Simulation 1.

Figure 11. Simulated average proportions of fixations to the target, reference item beside the target, and distractors with lexically unambiguous words (e.g., *cat*) in visually unambiguous contexts (e.g., one cat) in Simulation 1. Standard error bars are shown, along with onsets and offsets for the linguistic pulses as follows: pause | cat_1 | beside | star | cat_1 (complex NP).

Figure 12. Simulated average proportions of fixations to the target (△), competitor (▽), and reference item beside the target (□) with referential (A) and lexical plus referential (B) ambiguities in *garden path* trials in Simulation 1. Standard error bars are shown, along with onsets and offsets for the linguistic pulses as follows: pause | cat/bat_2 | beside | star | cat_1/bat_1 (complex NP).

Figure 13. Simulated average proportions of fixations to the reference item beside the target (A) and target (B) in *garden path* trials in Simulation 1. Black symbols indicate referential ambiguities (e.g., two cats), and white symbols indicate lexical plus referential ambiguities (e.g., baseball bat and mammalian bat). Standard error bars are shown, along with the mean onset and offset of the reference item. Curves indicate model fits from the growth curve analyses for

Predictions 2 (A) and 3 (B).

Tables

Table 1. Phonological pulses: Each phonological pulse is an nsem x nsem matrix where nsem is the number of nodes in the lexical semantic layer. (In this implementation nsem = 5. The five are *bat-mammal-sense, bat-baseball-sense, cat, star, square*). When the word corresponding to the pulse is being spoken, this matrix specifies the dynamics in the lexical semantic layer via Equation [1] in the Appendix. The matrices are hand-wired to reflect the semantics/pragmatics of each word. (*bm = mammal bat, bb = baseball bat, ca = cat, st = star, sq = square*).

"bat" pulse

	bm	*bb*	*ca*	*st*	*sq*
bm	1	0	0	0	0
bb	0	1	0	0	0
ca	0	0	0	0	0
st	0	0	0	0	0
sq	0	0	0	0	0

"cat" pulse

	bm	*bb*	*ca*	*st*	*sq*
bm	0	0	0	0	0
bb	0	0	0	0	0
ca	0	0	1	0	0
st	0	0	0	0	0
sq	0	0	0	0	0

"beside" pulse

	bm	*bb*	*ca*	*st*	*sq*
bm	0	0	0	0	0
bb	0	0	0	0	0
ca	0	0	0	0	0
st	0	0	0	0.2	0.2
sq	0	0	0	0.2	0.2

“star” pulse

	bm	*bb*	*ca*	*st*	*sq*
bm	0	0	0	0	0
bb	0	0	0	0	0
ca	0	0	0	0	0
st	0	0	0	0.4	0
sq	0	0	0	0	0

“square” pulse

	bm	*bb*	*ca*	*st*	*sq*
bm	0	0	0	0	0
bb	0	0	0	0	0
ca	0	0	0	0	0
st	0	0	0	0	0
sq	0	0	0	0	0.4

Table 2. Values of free parameters in the Simulation.

Parameter Description	Parameter Symbol	Value
Initial activation of all recurrent units (lexical sem., cross-word, action space)	*actinit*	0.01
Time constant for lexical semantics dynamics	*dtl*	2
Time constant for cross-word dynamics	*dtc*	0.05
Time constant for action space dynamics	*dta*	1
Inhibition between lexical semantics units	*linhib*	0.5
Inhibition among action space units	*ainhib*	1
Noise magnitude, lexical semantics units	*μmag*	0.05
Noise magnitude, action space units	*ηmag*	0.15
Word length (timesteps) (also = the duration of the phrasal pulse)	*wtime*	80

Figures

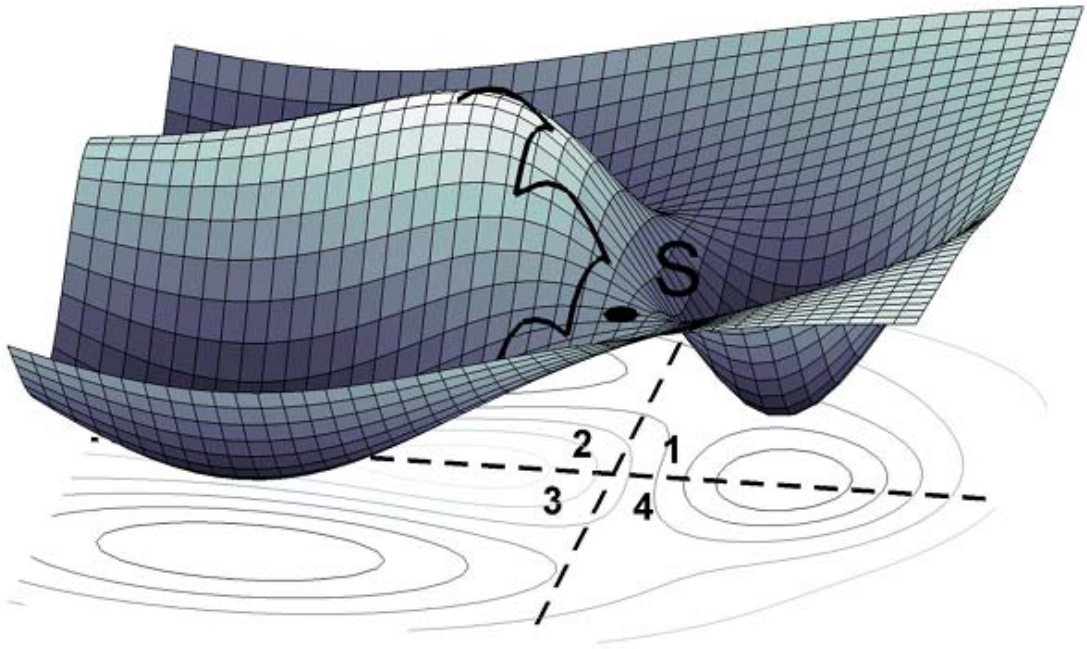

Figure 1.

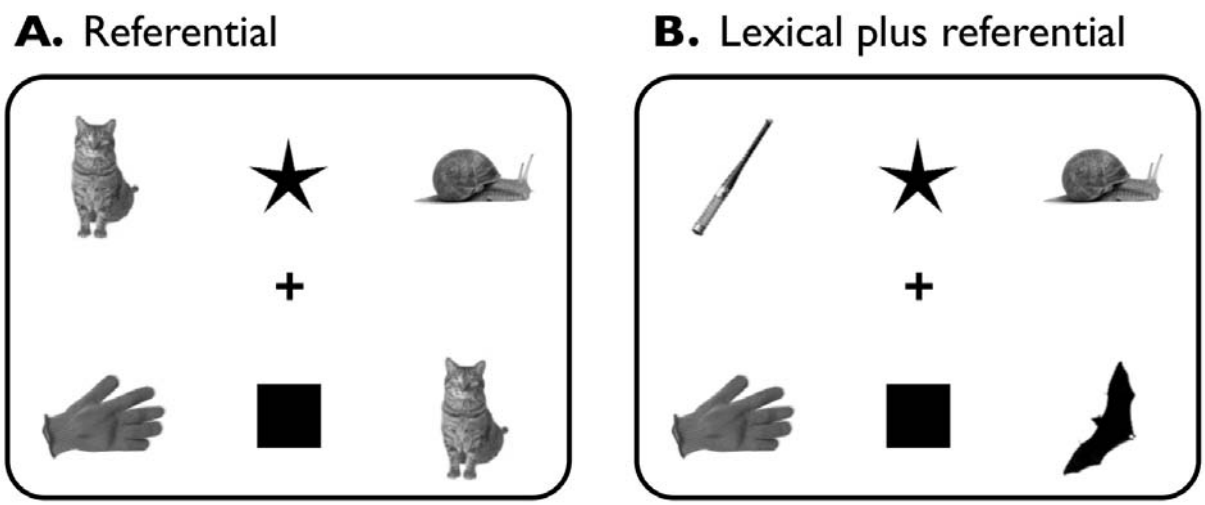

Figure 2.

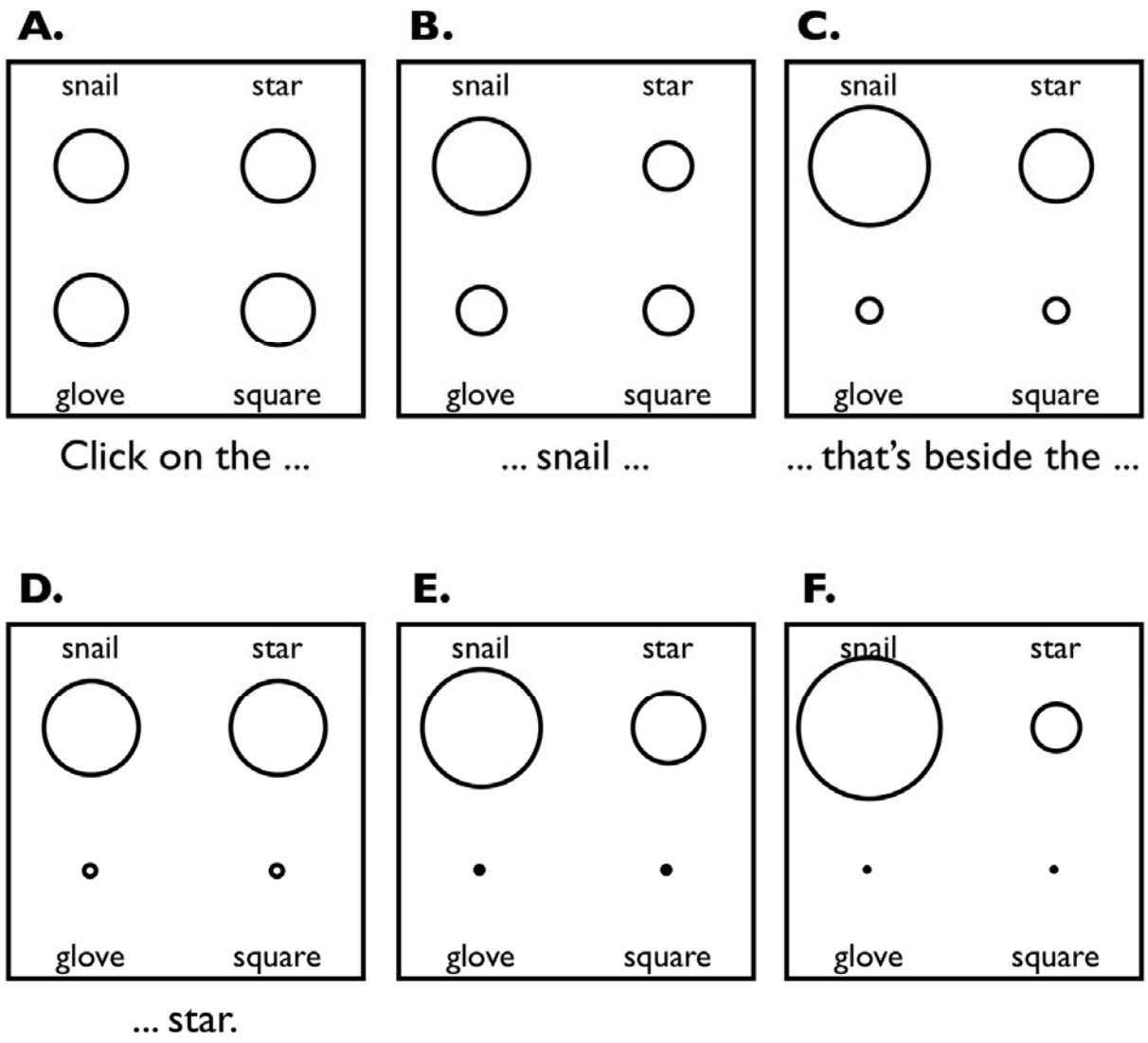
A.
snail
star
glove
square
Click on the ...
B.
snail
star
glove
square
... snail ...
C.
snail
star
glove
square
... that's beside the ...
D.
snail
star
glove
square
... star.
E.
snail
star
glove
square
F.
snail
star
glove
square

Figure 3.

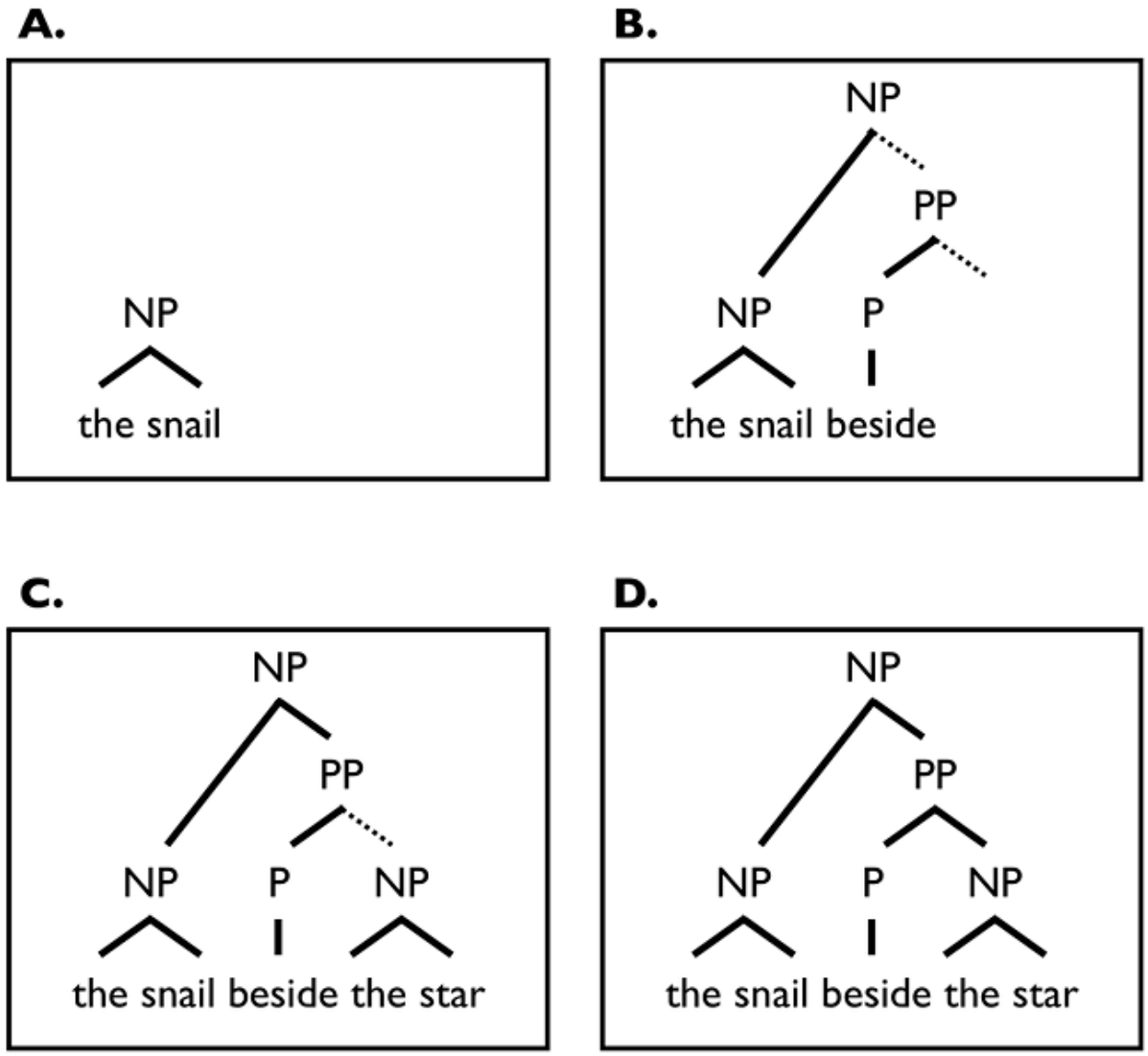
A.
NP
the snail
B.
NP
PP
NP
P
the snail beside
C.
NP
PP
NP
P
NP
the snail beside the star
D.
NP
PP
NP
P
NP
the snail beside the star

Figure 4.

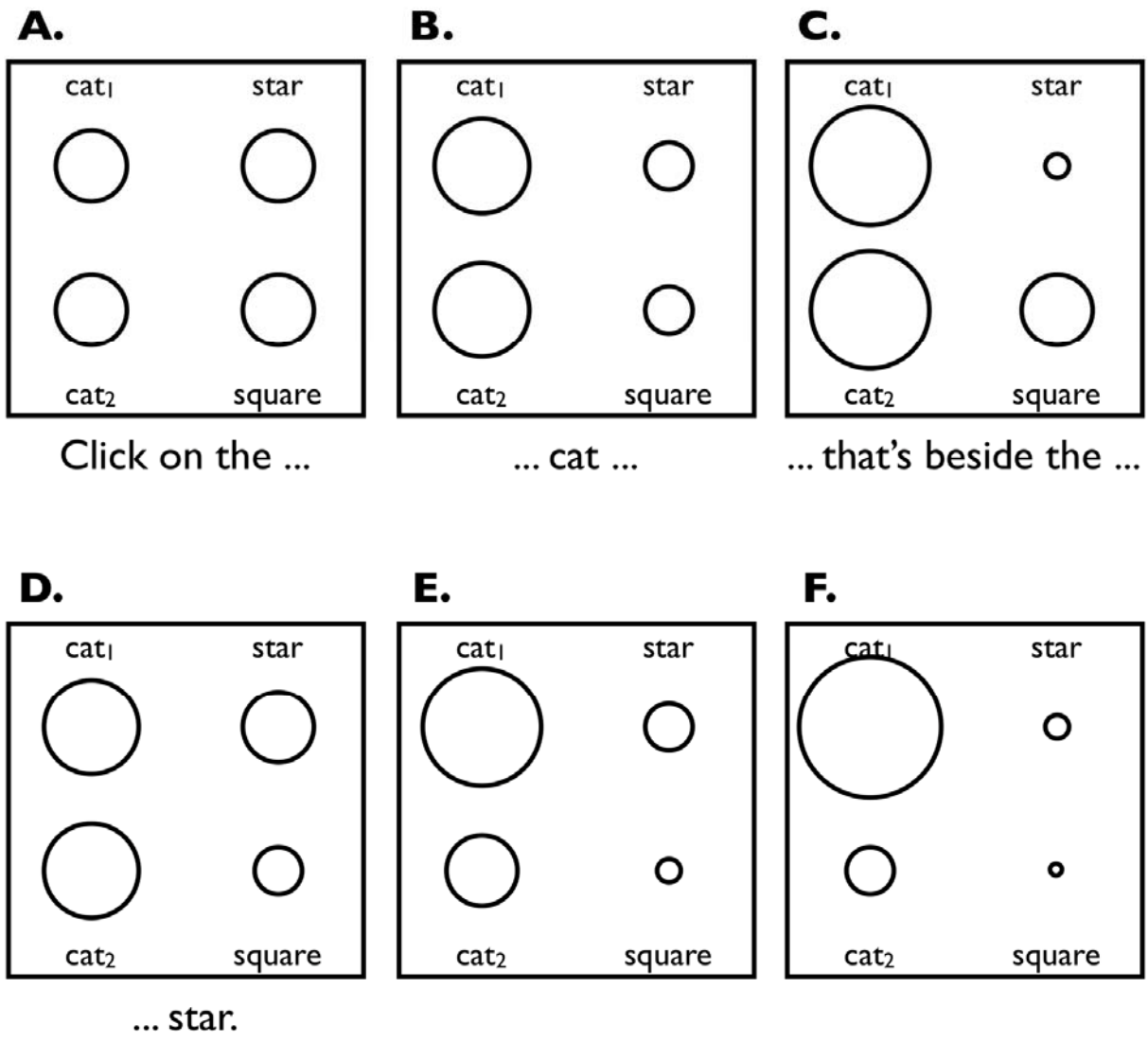
A.
cat1
star
cat2
square
Click on the ...
B.
cat1
star
cat2
square
... cat ...
C.
cat1
star
cat2
square
... that's beside the ...
D.
cat1
star
cat2
square
... star.
E.
cat1
star
cat2
square
F.
cat1
star
cat2
square

Figure 5.

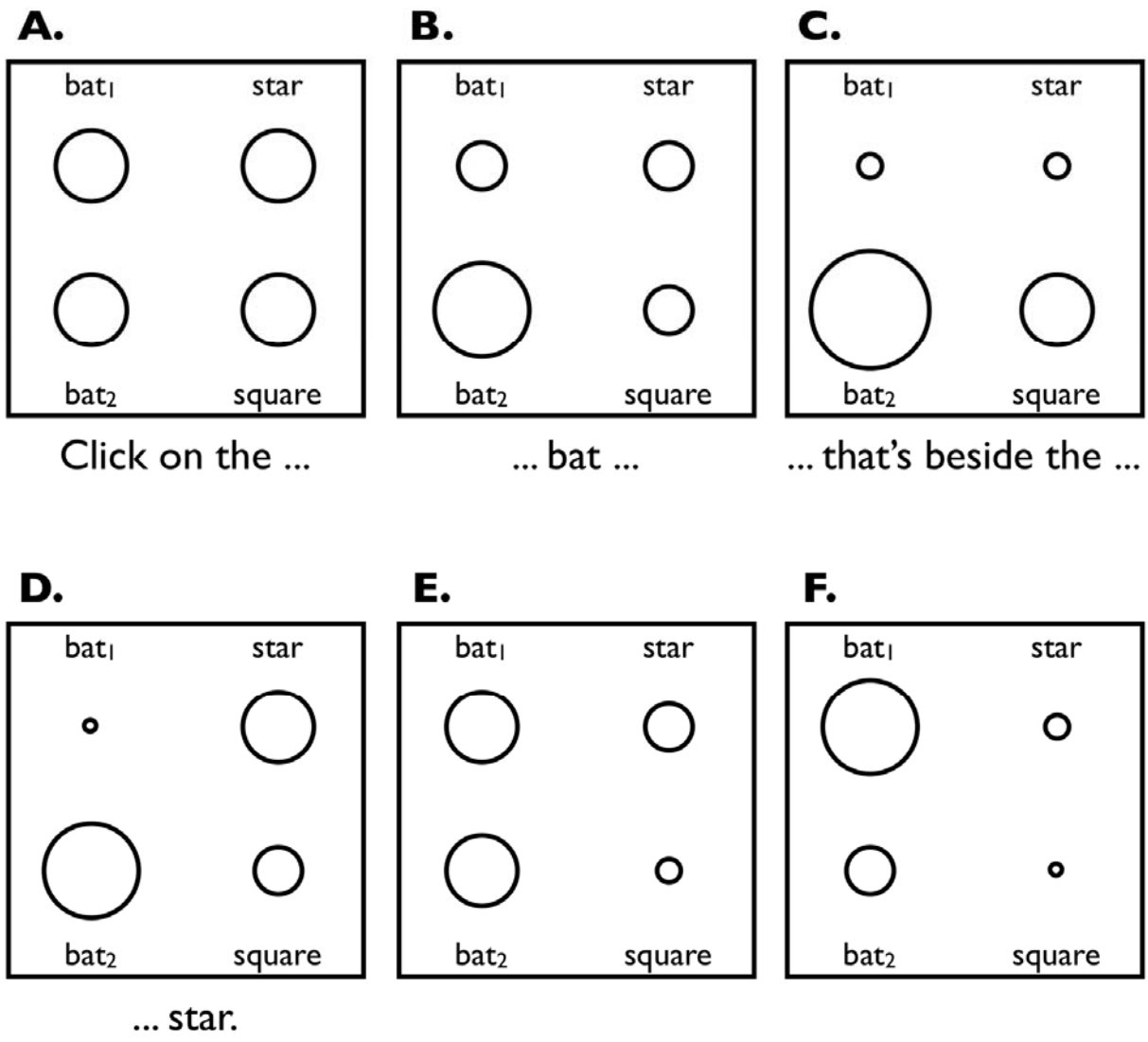
A.
bat1
star
bat2
square
Click on the ...
B.
bat1
star
bat2
square
... bat ...
C.
bat1
star
bat2
square
... that's beside the ...
D.
bat1
star
bat2
square
... star.
E.
bat1
star
bat2
square
F.
bat1
star
bat2
square

Figure 6.

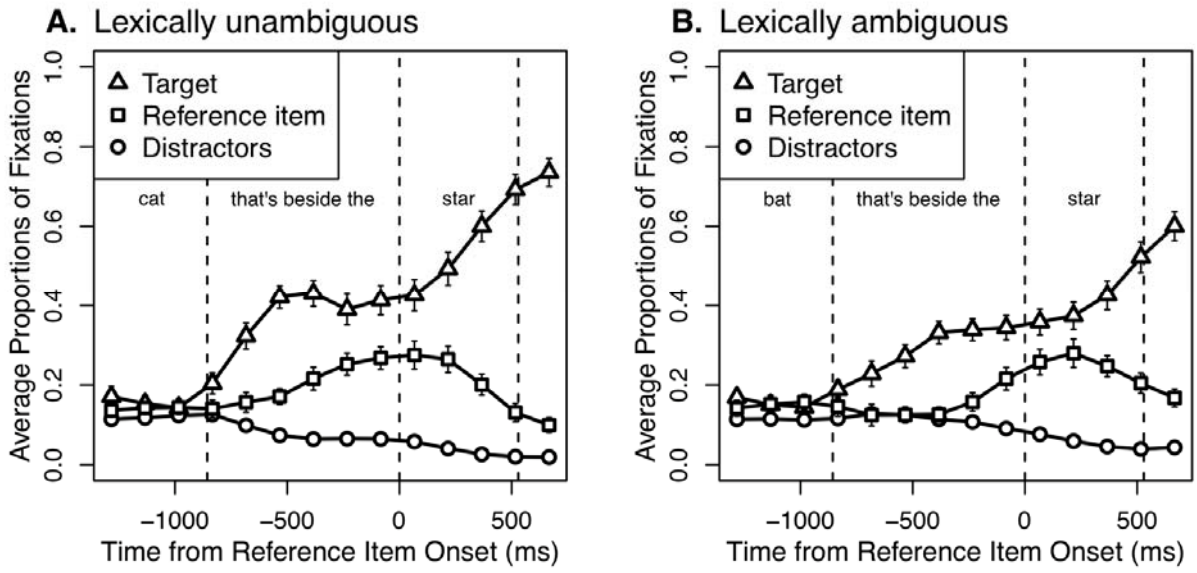
A. Lexically unambiguous
Target
Reference item
Distractors
cat
that's beside the
star
Average Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
−1000
−500
0
500
Time from Reference Item Onset (ms)
B. Lexically ambiguous
Target
Reference item
Distractors
bat
that's beside the
star
Average Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
−1000
−500
0
500
Time from Reference Item Onset (ms)

Figure 7.

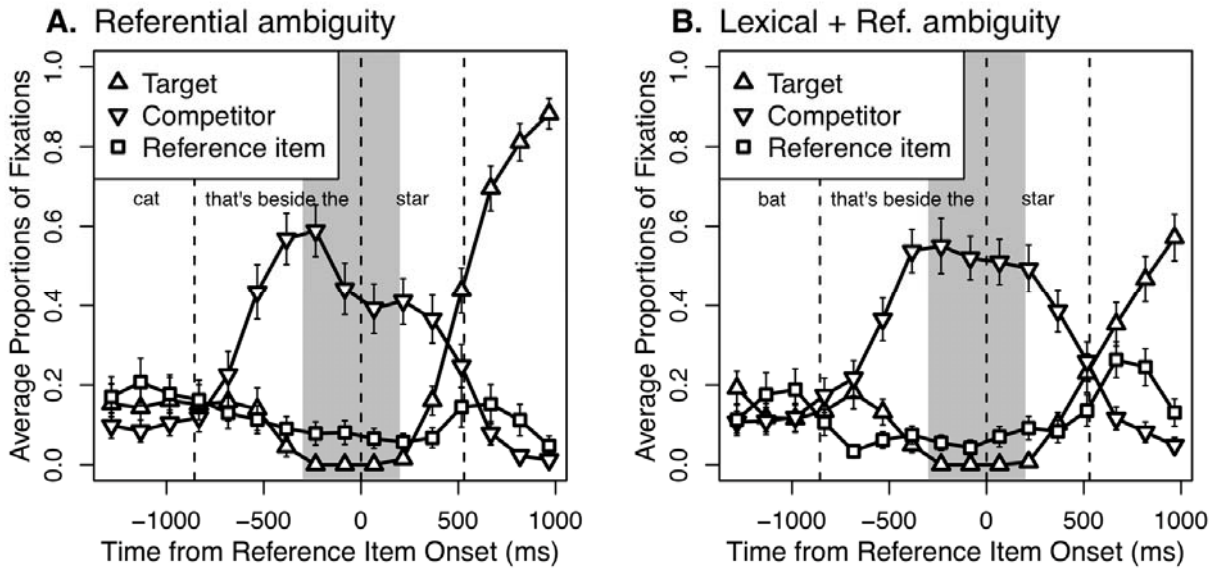
A. Referential ambiguity
B. Lexical + Ref. ambiguity
Target
Competitor
Reference item
cat
bat
that's beside the
star
Average Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
−1000
−500
0
500
1000
Time from Reference Item Onset (ms)

Figure 8.

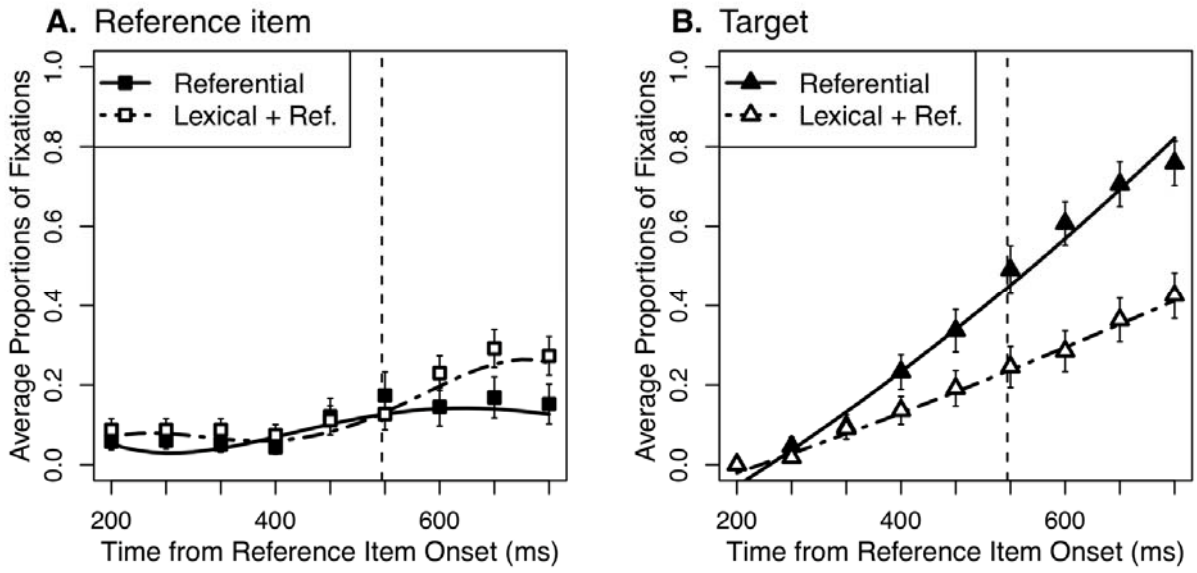
A. Reference item
Referential
Lexical + Ref.
Average Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
200
400
600
Time from Reference Item Onset (ms)
B. Target
Referential
Lexical + Ref.
Average Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
200
400
600
Time from Reference Item Onset (ms)

Figure 9.

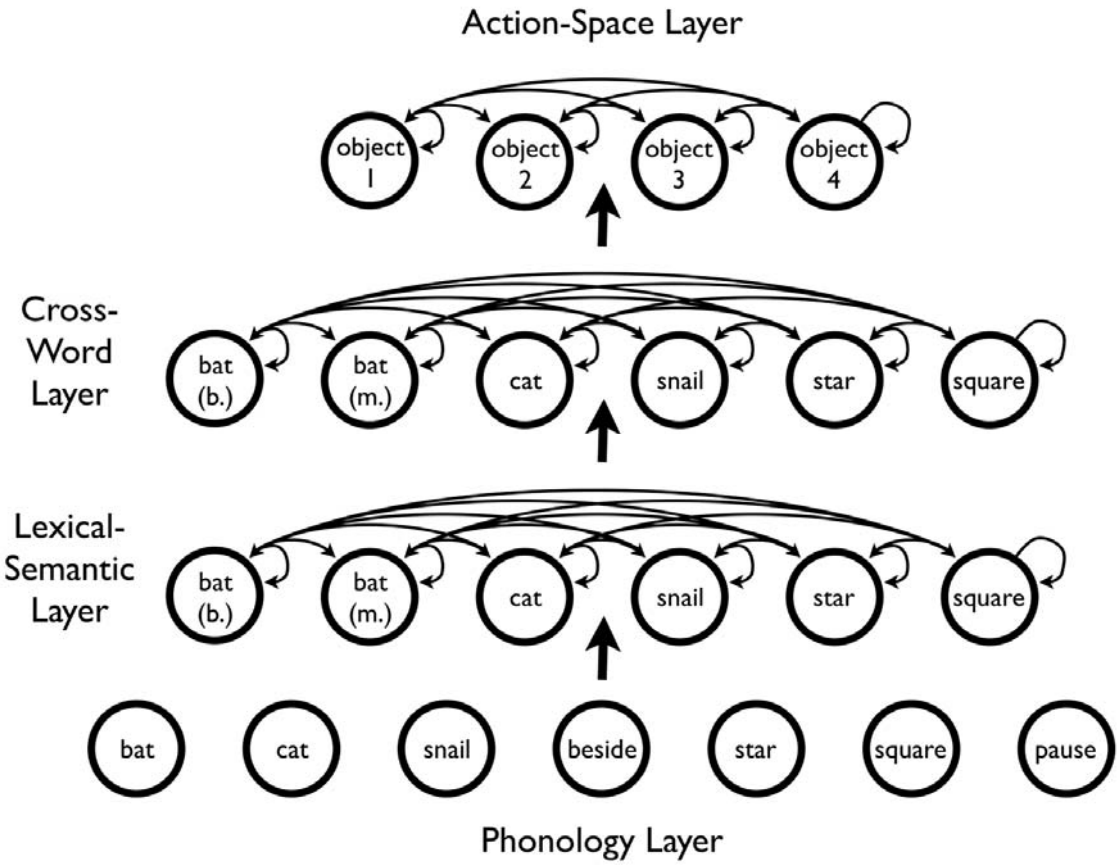
Action-Space Layer
object 1
object 2
object 3
object 4
Cross-Word Layer
bat (b.)
bat (m.)
cat
snail
star
square
Lexical-Semantic Layer
bat (b.)
bat (m.)
cat
snail
star
square
bat
cat
snail
beside
star
square
pause
Phonology Layer

Figure 10.

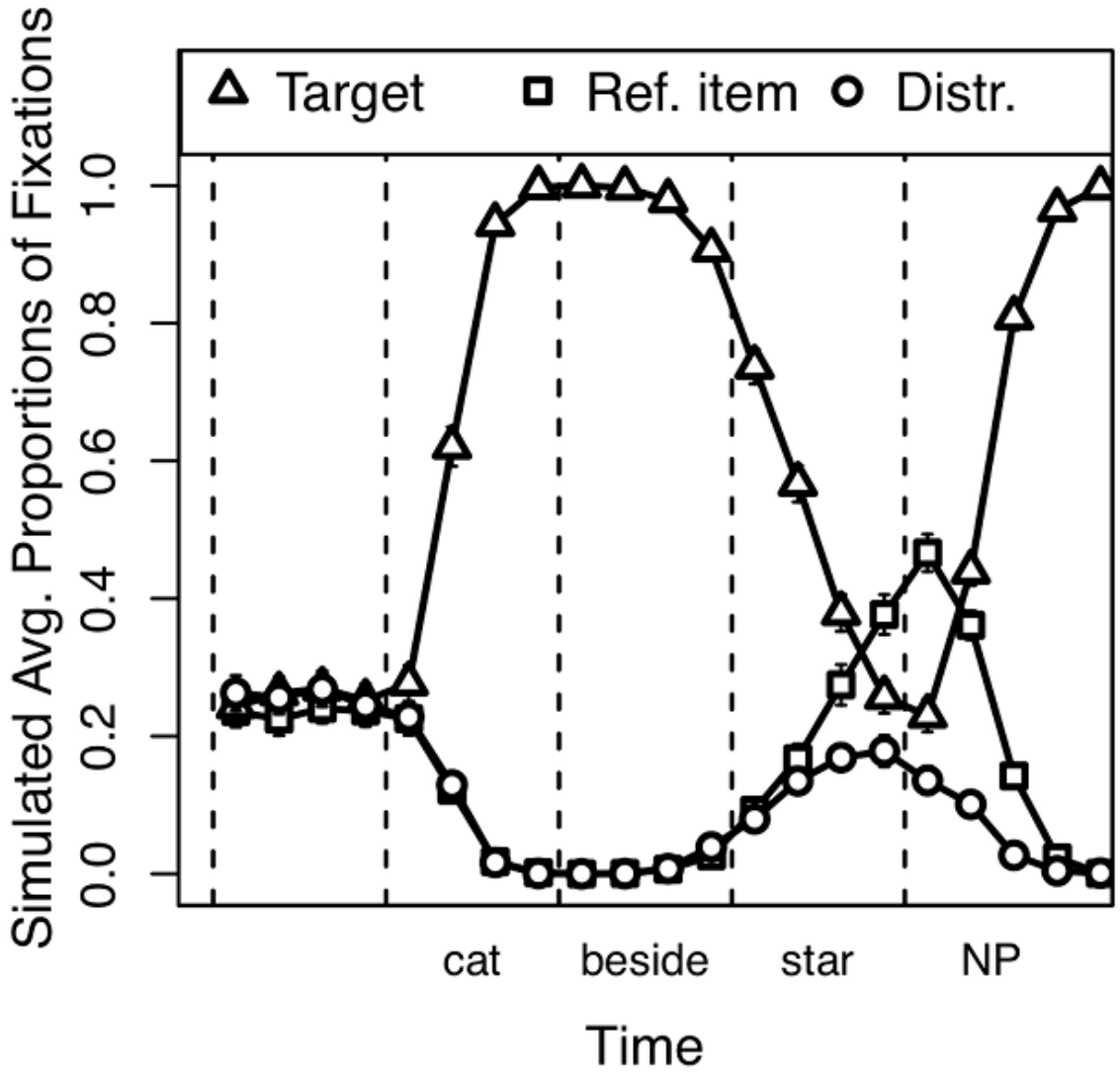
Target
Ref. item
Distr.
Simulated Avg. Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
cat
beside
star
NP
Time

Figure 11.

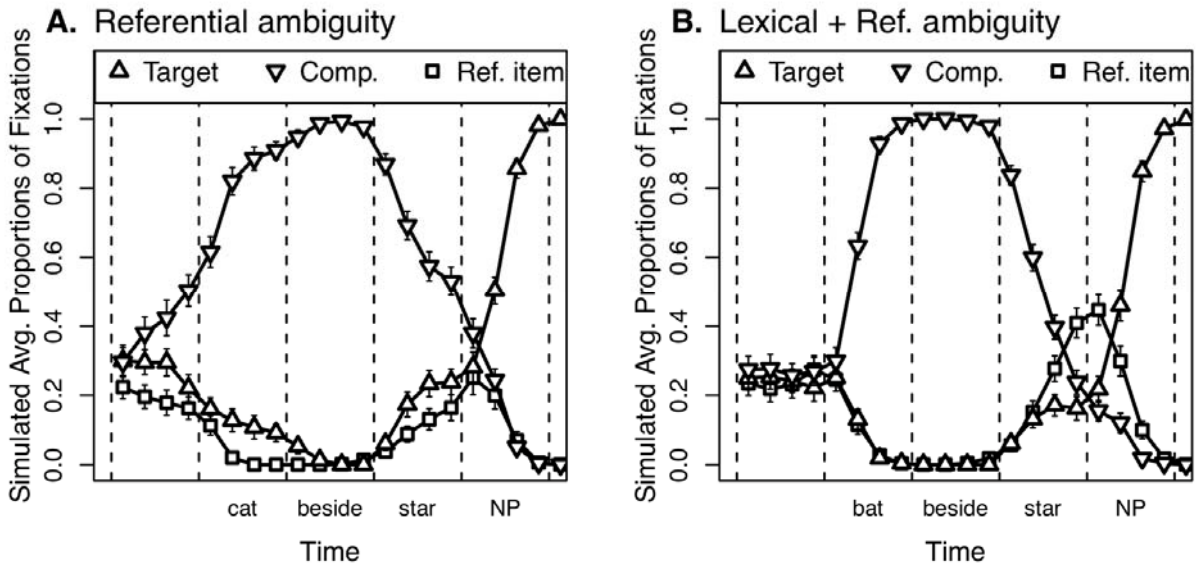
A. Referential ambiguity
Target
Comp.
Ref. item
Simulated Avg. Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
cat
beside
star
NP
Time
B. Lexical + Ref. ambiguity
Target
Comp.
Ref. item
Simulated Avg. Proportions of Fixations
0.0
0.2
0.4
0.6
0.8
1.0
bat
beside
star
NP
Time

Figure 12.

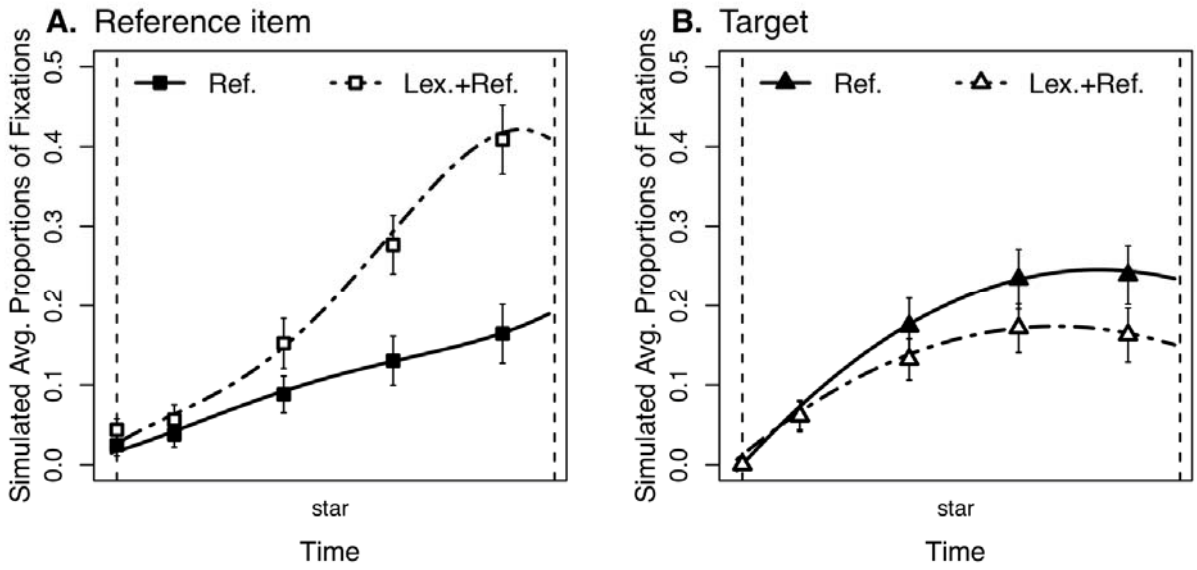
A. Reference item
Ref.
Lex.+Ref.
Simulated Avg. Proportions of Fixations
0.0
0.1
0.2
0.3
0.4
0.5
star
Time
B. Target
Ref.
Lex.+Ref.
Simulated Avg. Proportions of Fixations
0.0
0.1
0.2
0.3
0.4
0.5
star
Time

Figure 13.