Cognitive Science 33 (2009) 583–609

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01022.x

Incrementality and Prediction in Human Sentence Processing

Gerry T. M. Altmann, Jelena Mirković

Department of Psychology, University of York

Received 10 May 2008; received in revised form 29 August 2008; accepted 15 December 2008

Abstract

We identify a number of principles with respect to prediction that, we argue, underpin adult language comprehension: (a) comprehension consists in realizing a mapping between the unfolding sentence and the event representation corresponding to the real-world event being described; (b) the realization of this mapping manifests as the ability to predict both how the language will unfold, and how the real-world event would unfold if it were being experienced directly; (c) concurrent linguistic and nonlinguistic inputs, and the prior internal states of the system, each drive the predictive process; (d) the representation of prior internal states across a representational substrate common to the linguistic and nonlinguistic domains enables the predictive process to operate over variable time frames and variable levels of representational abstraction. We review empirical data exemplifying the operation of these principles and discuss the relationship between prediction, event structure, thematic role assignment, and incrementality.

Keywords: Sentence processing; Prediction; Simple recurrent network; Thematic roles; Incrementality

> *The one well-known comprehension model that does have prediction as a fundamental part of its architecture (Elman, 1990; see also Altmann, 1997), although frequently acknowledged as an interesting case of neural network modeling, has been equally lightly discarded as irrelevant to human language comprehension (e.g., see Jackendoff, 2002, p. 59, note 17).*
>
> Van Berkum, Brown, Zwitserlood, Kooijman, and Hagoort (2005, p. 444)

Language unfolds in time. And yet the words in this sentence unfold across space. This distinction, between the necessary unfolding in time of *spoken* language, and the necessary unfolding across space of *written* language, is often overlooked. Structure in time and

Correspondence should be sent to Gerry T. M. Altmann, Department of Psychology, University of York, Heslington YO10 5DD, UK. E-mail: g.altmann@psych.york.ac.uk

structure in space are not analogues of one another—after all, mechanisms able to move through space are all around us; mechanisms able to move through time are not. In this paper, we consider one particular mechanism for processing structure in time, a simple recurrent network (SRN) as proposed by Elman (1990), and consider the basic principles by which this mechanism accomplishes the task of both learning and representing linguistic structure. We consider also a simple extension of this mechanism, as proposed by Dienes, Altmann, and Gao (1999) and Altmann (2002) in which the learning of structure in one domain is constrained on the basis of structure previously learned in another. Our purpose in considering this mechanism is to identify whether and how principles manifest in its operation might also manifest in human sentence processing. In so doing, we shall attempt to reconcile distinct theoretical vocabularies that have become prevalent in recent years. Specifically, we shall focus on the relationships among theoretical constructs such as *thematic roles* and their assignment, *event structure*, *affordance*, *context*, and supporting all these, *prediction*. Throughout, our claim is not that the ''human sentence processing mechanism'' is an SRN (it is no more likely an SRN than it is a sausage machine or an augmented transition network—cf. Frazier & Fodor, 1978; Wanner, 1980), but rather, that the SRN embodies principles of representation and process (to the extent that they are separable) that capture, in part, the essence of what it means to ''understand'' human language.

Elman (1990) extended the original SRN (Jordan, 1986) in two ways that are relevant here: first, the input to the model was modified to become a combination of both the ''sensory'' input (i.e., an external signal) and the system's previous internal state (Fig. 1A). Second, the network learned time-varying structure by attempting to *predict*, at each moment in time, what the input would be at the *next* moment in time.[1] Thus, given a sequence of words presented one word at a time to the input units, the network's task was to predict (i.e., to output) at each time step the next word that would be input. The network did not learn to predict precisely what word would follow, but it did learn the range of words that would most likely follow (Elman, 1990, 1993). In effect, it learned which words were appropriate at each point in a sentence given the *prior context* (as encoded in the network's ''history'' of its prior internal states). The precise details of how the SRN learns only those statistical dependencies that are predictive of what will come next, and why it does not learn *all* the distributional facts about its input (including irrelevant facts such as a word like ''the'' occurring 14 words after a previous occurrence of ''the''), and how these lead to abstraction across experience, have been explained elsewhere (e.g., Altmann, 1997).

Elman's (1990 and 1993) models received, and attempted to predict, only linguistic input. Thus, and notwithstanding the significance of the SRN's ability to induce hierarchical structure on the basis of its linguistic experience, these models did not ''do anything'' with that structure. Unlike children, who learn linguistic structure through a process that ''grounds'' language in the external world (for reviews, see Gleitman, 1990; Gleitman & Gillette, 1995), Elman's SRN ''knew'' nothing more than the linguistic world. The fact that variation in its linguistic world was not grounded in variation in some external world meant that, during learning, variation in that nonlinguistic domain could not constrain learning in the linguistic domain. Dienes et al. (1999) and Altmann (2002) applied the SRN to precisely this problem—how sensitivity to variation in structure in one domain might influence the

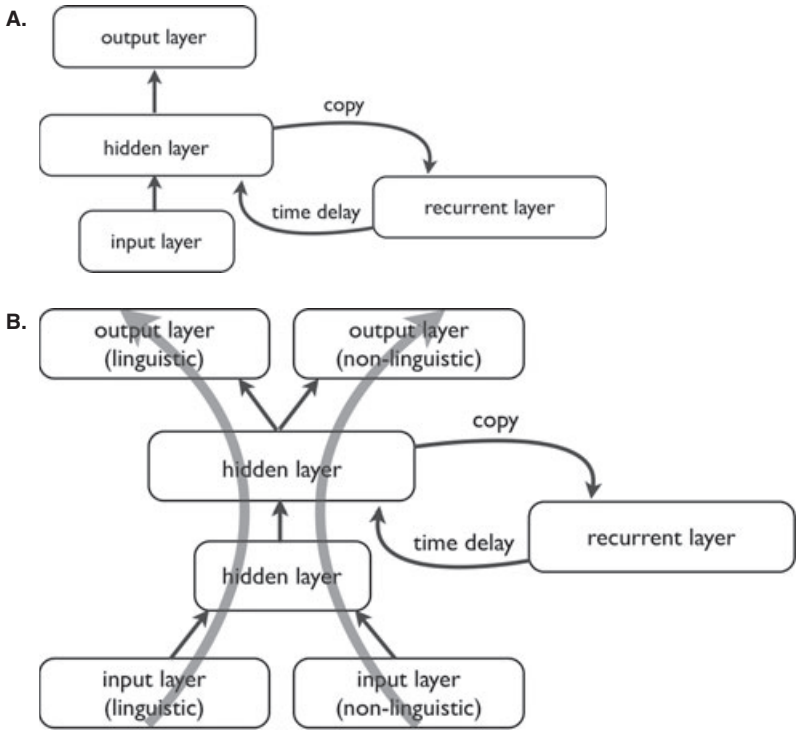

Fig. 1. (A) Elman's (1990) simple recurrent network and (B) the modified version from Dienes et al. (1999). The large arrows in (B) illustrate the two training ''routes'' described in the main text.

development of sensitivities to variation in structure in another (see also Altmann & Dienes, 1999). The architecture of their network is shown in Fig. 1B. The network's task was to predict successive inputs. When exposed to variation in one domain (e.g., when trained along the rightmost ''nonlinguistic'' route), the network encoded structure within that domain among the hidden layers. When the network was then exposed to structure in the other domain (e.g., when subsequently trained along the leftmost ''linguistic'' route), the network was biased, as it laid down in the hidden layers its encoding of structure in this new domain, by the configuration of weights (the encoding of structure) previously set by exposure to the first domain. The model demonstrated how the mapping of structure across domains could come about through prediction in time within each domain separately, and through a common substrate within which to represent structure in the two domains.

In the remainder of this paper, we explore how the following four principles, embodied in these SRN models, characterize language comprehension:

1. *Mapping across domains*: Structure in language has significance only insofar as it covaries with, and enables predictions of, structure in the external world (*event* structure). Sentence comprehension consists in realizing a mapping between sentence structures and event structures.

2. *Prediction*: ''Knowledge'' of the language can be operationalized as the ability to predict on the basis of the current and prior context (both linguistic and, if available, nonlinguistic) how the language may unfold subsequently, and what concomitant changes in real-world states are entailed by the event structures described by that unfolding language. Such predictions constitute the realization of the mapping between sentence structures and event structures.[2]
3. *Context*: Concurrent linguistic and nonlinguistic inputs, and the prior internal states of the system (together comprising the context), each ''drive'' the predictive process, and none is more privileged than the other except insofar as one may be more predictive than the other with respect to the subsequent unfolding of the input.
4. *Representation across time*: The representation of prior internal states enables the predictive process to operate across multiple time frames and multiple levels of representational abstraction. The ''grain size'' of prediction is thus variable, with respect to both its temporal resolution and the level of representational abstraction at which predictions are made.

We shall consider these principles in the context of the relationship between language, events, and attention to the external world. Although we shall use the terms ''prediction'' and ''anticipation'' interchangeably, we do distinguish theoretically between two different senses of these terms, reflecting a difference manifest in Elman's SRN: on the one hand, prediction is the *task* that the network is ''innately'' endowed with; activating across the output layer at time t what will be the input across the input layer at time $t + 1$. On the other hand, prediction/anticipation is what the network is able to do after a period of learning, and which is reflected in activation patterns across the hidden layers at time t, contingent on inputs at time $t - 1$ (and earlier), which enable at time $t + 1$ (and subsequently) activation patterns across the output. These activation patterns reflect, in turn, the previous experience of the contingencies between input and output patterns. Henceforth, unless explicitly referring to the network's task, we use the terms ''prediction'' and ''anticipation'' to refer to prediction as it reflects prior learning. Unsurprisingly, the general approach we take is one in which knowledge emerges through experience and is distributed within a dynamical system across a representational substrate supporting spreading activation (cf. Elman et al., 1996; Rogers & McClelland, 2004; Rumelhart & McClelland, 1986). Unlike more symbolic approaches, in which a variable (e.g., <*agent*>) can become instantiated at a particular moment in time with some value or other (e.g., <*Sam*>), there is no equivalent, within the theoretical framework we adopt, to this ''magic moment'' of symbolic instantiation—the challenge, then, is to explain the equivalent within this framework to theoretical constructs that sit more comfortably within symbolic approaches to linguistic representation, such as *thematic role assignment*. To the extent that the task for the comprehender is to determine *who did what to whom*, thematic role assignment is central to any theory of sentence comprehension. And to the extent that language, and events, unfold in time, the task for any theory of sentence comprehension is to explain how time enters into the theoretical equation.

The following sections begin with a review of the empirical evidence regarding language comprehension situated in a visual world, followed by further discussion of the theoretical implications of our approach.

1. Mapping language onto events

Broadly speaking, language either refers to states or to events (cf. Dowty, 1979). Events have internal complexity—causal structure—that is lacking in states; an event requires, as a minimum, an initial state and an end state, with one or more participants in the event undergoing some change between the initial and end states. The changes in state entailed by events are predictable, depending on the participants—dogs chase cats and postmen; balloons pop; and butter is spread most often, but not always, on bread. The predictability with which certain types of entity participate in certain types of event is, of course, mirrored in the language—in English, words referring to edible objects tend to follow words (verbs) referring to tasting or eating actions. Similarly, verbs referring to eating-like actions tend to be preceded by words referring to animates and tend to be followed by words referring to edible things. The influence of predictability on language comprehension is well established (for early studies, see Fischler & Bloom, 1979; Morton, 1964a, 1964b; Tulving & Gold, 1963). Elman (1990, 1993) demonstrated that the SRN is able to develop internal representations mirroring this predictable structure in the language. These representations *emerged* as a result of the prediction task in conjunction with the regularities in the input, such that the SRN could, given the current input (and access to its previous internal states), anticipate the range of words that could subsequently appear in the input. In the cross-domain version of the SRN (Fig. 1B; Altmann, 2002; Dienes et al., 1999), the model in fact was able to anticipate, given input in one domain (e.g., linguistic) what the corresponding input would be in the other domain (e.g., visual). This leads to the following empirical question: given a sequence such as ''the boy will eat…'' what, if anything, does the *human* sentence processing mechanism anticipate at ''eat''? Would it anticipate the upcoming language, or the unfolding conceptual correlates of the event which that language describes? Or both? In the following sections, we describe data that answer these, and related, questions. The bulk of these data come from studies employing a paradigm that allows the real-time evaluation of how language is mapped onto the visual world.

1.1. Mapping language onto the visual world

Roger Cooper first observed that as participants listen to a sentence referring to objects in a concurrently presented visual scene, the eyes move seemingly automatically to the objects in the scene as expressions referring to those objects are heard (Cooper, 1974). Tanenhaus et al. (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Salverda, Dahan, & McQueen, 2003) demonstrated that these eye movements are extremely finely time-locked to the unfolding acoustic signal, reflecting graded effects on lexical access of phonetic variation in the input. The essential finding is that as a word

unfolds in the acoustic input, so the eyes move toward whatever in the visual scene that unfolding word *could* refer to. This paradigm is not without limitations. For example, it is often the case that the visual ''world'' in such studies is restricted in terms of the number of objects that are depicted. However, this is not unlike real discourse; typically, only a very small number of entities are referred to across successive sentences of a discourse. However, there is no reason, in principle, why the visual scenes in this paradigm should not reflect the real-world complexities of the actual visual world to which we are more often exposed (different studies do vary in terms of the complexities of the scenes they employ—see Henderson & Ferreira, 2004, for discussion). Notwithstanding such caveats, the paradigm enables us to explore exactly the issue that is the focus of the current paper: the mapping of language structure onto real-world entities, as depicted in the concurrent visual world, and the events in which those entities might participate.

To address whether information at one point in a sentence could be used to anticipate information at a subsequent point, Altmann and Kamide (1999) showed participants scenes depicting, for example, a boy, a cake, and a number of other objects, all of which were inedible. Eye movements were recorded as participants heard each sentence. We found that participants looked more toward the cake at the verb when it was ''eat'' than when it was changed to ''move''—in this latter case, the verb did not select for one object more than any other, and so eye movements were split between the different objects that were moveable in the scene. Critically, this increase in eye movements toward the cake in the ''eat'' case occurred *before* the onset of the postverbal phrase ''the cake.'' As such, these were *anticipatory* eye movements. We accounted for this pattern of eye movements by proposing that, at the verb, participants anticipated which entity or entities in the visual context could take part in the eating event by virtue of being eaten. At a more linguistic level of description, this corresponds to the entity or entities that could fill the thematic role associated with the *theme* of the verb. Interestingly, there was no reason why it *should* have been one of the objects in the visual context that would be mentioned next—indeed, on half the trials, the sentences would continue as in ''the boy will eat the ice cream'' (and there was no ice cream in the scene). And yet participants acted as if they assumed that the postverbal noun phrase *would* refer to something in the scene. We return below to why we believe that it is inevitable that the system should act in this way.

A study that was equivalent to this in many respects, but which did not employ the visual world paradigm, was reported in Altmann (1999). There, participants read short passages such as the following:

> A car was driving downhill when it suddenly veered out of control. In its path were some dustbins and a row of bicycles. It injured/missed…

Participants had to judge, for each word in the word-by-word presentation, whether the sentence continued to make sense. In this case, there were more ''stops making sense'' judgments and longer reading times at the verb ''injured'' than at the verb ''missed.'' There were no such differences between the verbs in the following case:

A car was driving downhill when it suddenly veered out of control. In its path were some tourists and a row of bicycles. It injured/missed…

We accounted for this pattern of data in much the same way as we did in the Altmann and Kamide's (1999) study; participants anticipated at the verb the likely theme, drawing on the entities introduced in the context (in this case, a linguistic context, not a visual context). Where there was no plausible entity for the theme role associated with the verb ''injured'' (i.e., the first case), participants deemed the sentence to stop making sense at the verb (or if they did not, took longer to read the verb). An alternative account, still based on the notion of anticipation, is that participants anticipated the class of event that could occur given the context (the car veering out of control, the bicycles, and the dustbins or tourists); that is, in the final sentence at ''It,'' only certain kinds of event might be anticipated given a context containing an out-of-control car, with different kinds of event depending on whether the context contained only inanimate or also animate entities (Altmann, 2002; see also McRae, Hare, Elman, & Ferretti, 2005; who showed that typical participants in events prime the verbs that denote the event action). In fact, we do not view as different these two accounts (anticipating at the verb a likely theme, or anticipating before the verb is encountered what kind of verb might follow)—we argue below that they are part-and-parcel of a mechanism that learns to anticipate, on the basis of its current and preceding input, what input may follow. Either way, and as in the visual world case of Altmann and Kamide (1999), there was no reason for participants to assume that the entities introduced in the prior context would participate in the event unfolding in the final sentence; the first case (with dustbins and bicycles) could as easily have continued ''it injured some tourists who were standing nearby.'' Again, we return below to why there is a ''preference'' to make such assumptions, and in effect, to assign thematic roles to whatever entities are available in the context (as distinct to some as yet unintroduced entity).

These data, from two different experimental paradigms, suggest that *thematic fit* (how likely an entity is to have taken on some role in an event) can be determined in advance of the linguistic material that would unambiguously signal which entity should receive that role (i.e., which entity is actually going to be referred to in the position within the sentence that would mark that entity as having played that particular role in the event). Moreover, this determination of thematic fit occurs even in the absence of obligatory syntactic dependencies within the sentence that would signal which entity should receive which role; in the case of wh-constructions, for example, the wh-phrase signals an obligatory structural dependency between the wh-filler and the position in the sentence that unambiguously signals the role it should receive—thus, in ''which cake did the boy eat?'' the wh-phrase ''which cake'' signals that the cake participated in an event, and at the verb ''eat'' the appropriate thematic role can be assigned to the cake on the basis of this dependency (cf. Tanenhaus, Carlson, & Trueswell, 1989).

On the face of it, the ''eat the cake'' data are a straightforward manifestation of prediction during sentence comprehension, and of the mapping between the unfolding language and event structures; as the sentence unfolds, the system anticipates the kinds of participant that could participate in the eating event. On hearing ''eat,'' an abstract representation—that

is, reflecting abstraction across previous experience, both of events in the world and of structure in the language—becomes activated that reflects those things that are edible; in effect, an abstract set of potential entities that could be eaten. The existence of a potential member of that set in the concurrent scene attracts the eyes toward it.

One interpretive problem with these data is that they could be due just to the relationship between the verb and the context—it is a property of the word ''eat'' that likely words to follow will refer to edible things. Indeed, this same interpretive problem applies to Elman's (1990) demonstration of anticipatory activation of edible objects after a verb such as ''eat.'' Does anticipation at the verb reflect information about the verb alone (and an *association* between the verb and the words that tend to follow it), or does anticipation reflect information about the verb and whatever preceded it (in effect, an association between the verb and what may follow it conditional on what preceded it)? Elman (1993) demonstrated that the SRN *could* learn to predict the upcoming input on the basis of the current word conditional on its preceding context. Kamide, Altmann, and Haywood (2003) asked the equivalent question of their visual world data: Did anticipation of the cake at the verb ''eat'' reflect what could plausibly be eaten (reflecting only the lexical semantics of the verb), or what could plausibly be eaten *by the boy*? That is, was it the unfolding *sentence* (i.e., event with particular participants), or the unfolding *word* (i.e., action), which led to these predictive behaviors?

Kamide et al. (2003) contrasted sentences such as ''the man will taste the beer'' and ''the girl will taste the candy'' in the context of scenes depicting, for instance, a grown man, a girl, a beer, and some candy. The stimuli were designed so that, on the basis of world knowledge and the actually depicted individuals (i.e., the girl was a toddler, not an adult), the man would be more likely to taste the beer than the candy, and the girl more likely to taste the candy than the beer. At issue was where the anticipatory eye movements would be directed on hearing ''taste.'' If the prior demonstrations of anticipatory processing had been driven by the verb alone, we should see anticipatory looks toward both the candy and the beer, as both can be tasted. But if they had been driven by the integration of the verb with its linguistic subject (i.e., ''the man'' or ''the girl''), and, analogously, the action with the agent, we should see more looks to the beer than to the candy in the case of ''the man will taste…'' than in the case of ''the girl will taste…,'' and conversely for looks to the candy. This is in fact what we found. Thus, it would appear that anticipatory processing in sentence comprehension, at least as evidenced by this study, is the result of the integration of each unfolding word with the prior linguistic context, the concurrent visual scene, and general world knowledge—it is the result of mapping the unfolding sentence onto the event structures afforded by, in these cases, the linguistic and visual contexts. Kamide et al. (2003) also demonstrated, in a study using Japanese sentences, that anticipatory processing of this kind is not restricted to verbs, but is driven by whatever cues in the sentence signal the kinds of role that can be assumed by the participants in the event being described; a sequence of two noun phrases, marked for nominative and dative case, respectively, caused anticipatory looks toward a plausible object (subsequently referred to in the accusative case—i.e., as a *theme*) that could be transferred from the entity marked in the nominative case to the entity marked in the dative case—in effect, anticipating the kind of event (an event involving transfer) that would be denoted by the subsequent sentence-final verb. The Kamide et al. study thus

demonstrated how verbs (and their subjects) can be used to anticipate upcoming arguments, and how nouns (and their accompanying morphosyntax) can be used to anticipate aspects of upcoming verbs (cf. earlier discussion of Altmann, 1999).

These data leave unresolved, however, the nature of the representations that were constructed, or activated, at the verb (or before, in the Japanese study). Did anticipatory eye movements toward the cake or the beer/candy reflect anticipation of what *word* would follow, or of what *entity* would most likely be referred to next? Or could it have been a combination of both? The distinction here is between, on the one hand, predicting subsequent linguistic input and, on the other, predicting the conceptual and/or real-world correlates of that input. For example, with respect to eye movements during reading, McDonald and Shillcock (2003a, 2003b) demonstrated that eye movements can reflect the likelihood of what *word* could follow given the preceding word (i.e., sensitivity to statistical distributions of lexical bigrams). Other studies have demonstrated that anticipation can reflect properties of a word such as its likely *phonological form*. DeLong, Urbach, and Kutas (2005) presented participants with written sentences one word at a time such as ''The day was breezy so the boy went outside to fly a kite/an airplane.'' The continuation ''an airplane'' was less plausible, but possible. Critically, the more plausible continuation required the article ''a,'' whereas the implausible continuation used ''an.'' DeLong et al. found that the N400 to the postverbal article correlated inversely with the cloze probability of the article (itself derived from the cloze probability of the noun). Thus, the less likely ''airplane'' was as a continuation, the less likely would be the article ''an,'' and the greater would be the N400 to this article (see also Van Berkum et al., 2005, for an auditory equivalent that led to the same conclusions). These data demonstrate that readers and listeners can use the available context to rapidly predict the specific words that are likely to come next. In one sense, these data therefore indicate that anticipation/prediction can consist of the projection of *linguistic structure*. In which case, are the predictions we make as the language unfolds *only* about the upcoming language per se, or might they *also* include information about the upcoming conceptual correlates of the events described by that language?

1.2. The role of object representations in language-mediation of visual attention

Altmann and Kamide (2007) reported a study that more directly manipulated the interaction between the interpretation of the visual scene and the unfolding interpretation of the concurrent language (see also Altmann & Kamide, 2004, for related findings). They showed participants scenes depicting, among other things, a cat, a few mice huddled together, and a pile of feathers. Participants heard ''The cat will kill all of the mice.'' More looks were directed toward the mice than toward the feathers before the onset of the phrase ''the mice.'' This is in accordance with the prior results. More interesting was what happened in the case of the sentence ''The cat has killed all of the birds.'' Here, we were interested in where the eyes would be directed prior to hearing ''the birds.'' In this case, we found more looks now toward the pile of feathers than toward the huddled mice. The feathers could not have been looked at because they would be referred to next—they violated the ''selectional restrictions'' of the verb (that is, they were at odds with what would normally be predicted, on the

basis of linguistic experience, to be referred to as the object of killing). Instead, they were looked at because the comprehender presumably anticipated after ''killed'' that whatever would be referred to next would be something that (a) was animate (thus satisfying the selectional restrictions of the verb), (b) was something that cats are likely to kill, and (c) had been, but was no longer, alive (thus satisfying the tense morphology of the verb). At this point, knowledge of real-world contingencies would have identified that whatever was killed would have been associated with some probability with the pile of feathers—based on the real-world contingency between events involving cats killing and resultant states in which feathers are scattered on the ground. On hearing ''has killed'' the eyes moved to the feathers because of our experience of what such piles of feathers can indicate—they indicate the past existence, and likely current nonexistence, of a bird.

A related finding was reported by Knoeferle and Crocker (2007). In their experiment, participants were briefly presented with a sequence of three pictures. For example, the first depicted a scene containing, among other things, a waiter, a chandelier, and a set of crystal glasses; the second showed the waiter polishing the chandelier (i.e., now performing an action); and the third was the same as the first (i.e., no action being performed). This sequence of pictures simulated the temporal properties of an event in the real world, indicating that the polishing action happened in the past, that is, that one of the properties of the chandelier has changed from the first to the last picture (i.e., it being polished—although the final picture did not depict the change in state that occurred between as a result of the polishing event). While the last picture was on the screen, participants heard ''The waiter polished recently the chandeliers.'' (The sentences were in German.) And even though there were two objects that could equally plausibly be polished (chandeliers and crystal glasses), participants looked more during ''polished'' at the chandeliers than at the crystal glasses. As in the ''cat has killed'' example, the representation of the event of polishing as having taken place in the past (given the final picture) caused the eyes to anticipate at the verb the entity on which the action denoted by the verb had been performed.

One interpretation of these past tense results is that the eye movements triggered by the past tense verb reflected the fit between what would likely be referred to next (what do cats typically kill? What objects are typically polished?) and the affordances of the objects in the scene (see Chambers, Tanenhaus, & Magnuson, 2004; for an initial demonstration that language-mediated eye movements are modulated by the affordances of the objects in the scene). On this interpretation, there is an element of prediction: Both mice and birds, or glasses and chandeliers (depending on the study), would plausibly be predicted to be referred to next, but only the birds would, in conjunction with the depicted feathers, satisfy the tense morphology. Similarly for the polished chandeliers—hence more looks toward the feathers/chandeliers than toward the mice/glasses. The notion of ''fit'' with the affordances of the objects within a scene is central to the consideration of why the eyes move toward anything at all in the visual world paradigm. The details of why they do move are fleshed out in Altmann and Kamide (2007). We summarize those details, and the mechanism they entail, below. As will become apparent, it is this mechanism that gives rise to the ''preference'' to anticipate that objects and entities in the context will play a role in the event described by the unfolding language.

Viewing a scene results in the activation of object representations that precede the arrival of the language (this knowledge includes the object's affordances—knowledge based on our experience of how that object interacts with other objects; that is, knowledge of the dynamically changing contexts in which that object may be experienced[3]); when a sequence of words is subsequently heard, the representations that they engender may overlap to some degree with the preexisting representations already activated through the prior interpretation of the concurrent scene (in Altmann & Kamide, 2007; we described these representations as featural and multimodal, which is common with other approaches to semantic cognition based on distributed representations, e.g., Farah & McClelland, 1991; McRae, de Sa, & Seidenberg, 1997; Rogers & McClelland, 2004; Tyler & Moss, 2001). To the extent that there is overlap among the representations, those representations increase in activation (the overlapping components increase in activation because they receive dual support, and this increase spreads to the remainder of the representation). These changes in activation necessarily increase the activation of that part of the object representation that encodes the object's location, and it is this change in activity that causes the eyes to shift, with some increased likelihood, toward that location. We view the change in activation of an object's representation as a change in the attentional state of the cognitive system, with this change either *constituting*, or *causing*, a shift in covert attention (see Altmann & Kamide, 2007; for further detail). This account does not require an external mechanism (e.g., working memory; Knoeferle & Crocker, 2007), because attention is instantiated within the same representational substrate as linguistic and nonlinguistic information (see also Cohen, Aston-Jones, & Gilzenrat, 2004); in other words, different states of this representational substrate represent the attentional modulation that drives eye movements. In contrast, other models of the visual world data (e.g., Knoeferle & Crocker, 2006, 2007) postulate independent representations of the utterance meaning, scene information, and linguistic expectations, and these are related through processes of coindexation and subsequent reconciliation of the utterance meaning and scene information. The principles we adduce here, however, are part-and-parcel of a system in which utterance meaning, scene information, and linguistic expectation are representationally indistinguishable and exist within a *unitary* system that learns, represents, and processes language and the world.

This mechanistic account of language-mediation of visual attention relies on a representational substrate that is shared across different cognitive domains. Within the context of the SRN model of cross-domain ''structure sharing'' (Dienes et al., 1999), activation of the hidden layers (the common representational substrate) can reflect either input from one domain, the other domain, or both (for the sake of exposition, we restrict discussion of this model to just two domains). The activation state of the hidden layer thus has consequences for what will be output across each domain's output units *regardless* of the source of the signal that caused that activation state. In effect, the activation of the hidden layer reflects a mapping, acquired through prior experience, of structural variation within one domain onto structural variation within the other. Altmann (2002) reported hierarchical clustering analyses, which suggested that the network encoded the equivalence between sequences in one domain and sequences in the other. We assume that the sharing of representational substrate between language processing and event encoding gives rise to an equivalence between linguistic

representations and corresponding nonlinguistic representations that drives not only language-mediation of visual attention but also the mapping of language onto event structures. This equivalence arises, even though, as we discuss below, sequential variation in language has different temporal properties to the changes in time that accompany the unfolding of real-world event structures.

To summarize our account of the earlier ''cat has killed'' example: We assume that, on seeing the feathers, a representation is activated that encodes the likelihood of the prior existence of a bird (and its currently deceased state) and that on hearing ''the cat has killed'' a conceptual representation is activated that encodes the likely participation in the killing event of a bird, and, given the tense morphology, the requirement that whatever will be referred to next must no longer be alive. The representational overlap between the two sets of activated representations causes the eyes to move toward the feathers. This account also explains why, after the fragment ''the boy will eat...,'' the eyes anticipate whatever is edible in the visual context (even though the sentence could go on to refer to something new): the conceptual representations activated at the verb overlap with those preactivated by the scene, causing anticipatory eye movements toward the edible object (the cake, in the original example). We can generalize from this case, where the visual context determines the prospective participants in the event, to the case where there has been *no* visual correlate to the objects being described by the language, as in the case of the purely linguistic narratives employed in Altmann's (1999) study: in an example such as ''Jeff took a Hershey bar out of his cargo pants. He ate…,'' a mental representation corresponding to the chocolate bar is activated by the first sentence and overlaps with the conceptual representation subsequently activated on hearing ''He ate.'' The overlap between the two causes a boost in the activation of the representation corresponding to the chocolate. In effect, ''attention'' shifts within the mental representation of the situation toward the chocolate. On the account where, in an example such as this, the verb ''eat'' is in fact anticipated by the prior occurrence of the chocolate (cf. Altmann, 2002; McRae et al., 2005), the chocolate in the first sentence activates conceptual representations corresponding to the affordances of a chocolate bar—chocolate affords eating—and when the subsequent verb ''ate'' is encountered, the representations corresponding to this particular ''feature'' of chocolate increase in activation due to the dual support they receive, and this increase spreads back to the remainder of the chocolate representation, causing, again, a shift in the attentional state of the system.

We now consider a final set of studies suggesting that the eye movement data we have described above generalize beyond concurrent visual input, and which provides more direct evidence that the conceptual representations activated as the language unfolds anticipate the structure of the *event* which that language describes.

1.3. What you see is not what you get: Eye movements in a mental world

Altmann and Kamide (2004, 2009) report a study in which participants heard a short narrative describing a change to the world as depicted in a concurrent scene. At issue was whether the accompanying eye movements as the language unfolded would reflect the depicted world, or the *changed* world. Participants were shown a scene depicting a room

within which was a woman, a table and, on the floor, a bottle and an empty wineglass. Participants heard one of the following two passages:

> The woman will move the glass onto the table. Then, she will pick up the bottle and pour the wine carefully into the glass.
>
> The woman is too lazy to move the glass onto the table. Instead, she will pick up the bottle and pour the wine carefully into the glass.

Prior studies had shown that, after the verb ''pour,'' the eyes would anticipate the *goal* of the pouring (i.e., where the wine would be poured) during the postverbal region and prior to hearing ''the glass.'' We predicted that the eye movements would reflect the *situation* as modulated by the language, and that anticipatory eye movements would reflect the glass as being on the table in the first passage above, but as being on the floor, still, in the second. In one version of this study (reported in Altmann & Kamide, 2009), the ''blank screen paradigm'' was used (Altmann, 2004); participants saw the scene for 5 s and then the screen went blank, and 1 s later the two-sentence auditory stimulus began with the screen remaining blank throughout. By the time participants heard, ''the wine'' in the second sentence, the screen had been blank for over 6 s. We found the predicted effect on anticipatory eye movements; more saccades launched to where the table-top *had been* during ''the wine carefully into'' in the ''moved'' case than in the ''unmoved'' case. Critically, during ''the glass'' itself, more saccades were launched to where the table-top had been in the moved than in the unmoved condition, and indeed, there were no more looks in the moved condition to where the glass had actually been than toward where a distractor object (a bookcase) had been (just as, in the unmoved condition, there were no more looks during ''the glass'' toward where the table-top had been than toward where the distractor object had been). The data suggest that the fact that the glass had actually been seen in a specific location did not have any residual ''pull'' on the eye movements—the eyes were directed solely toward the ''mental location'' of the glass (as determined by the interaction between the narrative and the visual memory for the objects, and their locations, that would take part in the events described in that narrative). To all intents and purposes, the location of where the glass had been seen was no more relevant than the location of where the irrelevant distractor had been seen; all that mattered, and all that drove the eyes, was where the objects were located in the dynamically updated mental representation of the external world as depicted by the previously seen visual scene.

These last data confirm that anticipatory eye movements (and even eye movements toward, for example, the glass *during* ''the glass'') do not reflect only the upcoming, or concurrent, language per se (cf. the studies by Van Berkum et al., 2005; DeLong et al., 2005; and McDonald & Shillcock, 2003a, 2003b), they reflect also the unfolding conceptual correlates of the event which the unfolding language describes; they reflect the likely involvement of one entity or another as represented within a dynamically changing representation of the *situation* within which the event being described unfolds. From the perspective of eye movement research, these data are interesting because they are a further demonstration of

goal-directed eye movements in the absence of anything in the visual field to move the eyes toward (cf. also demonstrations in Altmann, 2004; Brandt & Stark, 1997; Hoover & Richardson, 2008; Knoeferle & Crocker, 2007; Laeng & Teodorecu, 2002; Richardson & Spivey, 2000; Spivey & Geng, 2001). But more importantly, they demonstrate that it is not only the visual memory of where something was that drives the eye movements in these cases—rather it is some dynamically modifiable representation of the object's location that appears to be updated through an interaction between the sensory percept (where the glass had actually been) and the unfolding, but internalized, event representation (the repositioning of the glass); this interaction results in object representations that reflect the ever-changing state of the world that the unfolding language describes. Moreover, these internalized representations are no less privileged with respect to their influence on eye movement control than the memory of the actual sensory input—recall that the memory of where the glass had actually been in the ''moved'' condition had no bearing on where the eyes moved; those movements were determined solely by the internalized representation of where the glass would end up, given the events described in the language, and the anticipation that the glass would be the object into which the pouring of the wine would take place.

A series of studies by Knoeferle and Crocker (2006, 2007) further illustrates the dynamic activation of internalized representations of events and how these can be modulated by events depicted in the concurrent visual world. Knoeferle and Crocker contrasted the knowledge about events abstracted from experience (e.g., actions typically afforded by particular agents, for instance, detectives spying) with the knowledge accompanying experience of events depicted in the concurrent world. In their study, these two sources of knowledge were in conflict; for example, a wizard was looking at a pilot through a telescope while a detective was serving the pilot some food. The concurrent language described one particular event (''the detective/wizard will soon spy on the pilot,'' although in the German sentences, the pilot was mentioned first in the sentence, and was marked in the accusative case—''the pilot ACCUSATIVE spies on soon the detective/wizard NOMINATIVE),'' and at issue was whether it would be the event knowledge abstracted across experience (that detectives prototypically spy) that would drive eye movements at the verb toward the agent of the spying, or whether it would be the event knowledge associated with the concurrent situation that would drive these eye movements. If the former, the eyes would be driven to the detective, but if the latter, they would be driven toward the wizard given that it was the wizard who was using an instrument prototypical of spying. Knoeferle and Crocker found that, on hearing ''spies on,'' participants were more likely to look at the wizard than at the detective (or, in a blank screen version, to where the wizard had been than toward where the detective had been). Thus, the event representations that drove the eye movements in these studies appeared to be modulated primarily by the depicted situation; the information from the concurrent visual world appears to have taken precedence over information abstracted across prior experience. However, when the concurrent visual world did not provide information that conflicted with this abstracted knowledge (the scene did not show the telescope), participants did use their experiential knowledge to drive their eye movements (toward the detective). The fact that the visual world took precedence in these studies over experiential knowledge is not surprising, of course, given that the most reliable cue to *who is doing what*

to whom is whoever one *sees* doing it, not whoever one *thinks* is doing it. This is reflected in the principle discussed above which states that no input is more privileged than another except insofar as one may be more predictive than the other in a given situation.

To account for their data, Mayberry et al. (2009) used a recurrent network (a basic recurrent sigma–pi network) that mapped an input representation containing information both about the current word and about the participants and actions in the scene onto an output meaning representation comprising the action corresponding to the verb in the sentence, the entities referred to in the sentence, and the relationship of these entities to each other as agent or patient. The network's task was, in effect, to update the representation of the meaning of the utterance given the scene as the utterance unfolded. This implementation shares some properties with an Elman net: the meaning of an utterance was updated as the linguistic input unfolded, and thus the input representation developed over time. In addition, the meaning of the utterance was influenced by the visual context in which it occurred. However, the models of Mayberry et al. involve a mapping of the two domains (visual and linguistic) onto a common meaning output. The models are therefore explicitly trained on the mapping from linguistic and nonlinguistic input onto meaning and, moreover, onto a meaning representation with fixed structure corresponding to that associated with the meaning of SVO and OVS sentences. In our account of the principles underlying language comprehension situated in a visual world, this mapping, between language and ''meaning,'' is an *emergent* property of a training regime that maps input at time t onto input at time $t + 1$; no ''oracle'' is required that knows, in advance of training, what the correct mapping should be, nor what the correct meaning representation should be; the only information given to the network is the current input and the input at the next time step. We return in the sections below to discussion of how such mappings might emerge.

To summarize the data thus far: the behaviors we see using the visual world paradigm do not reflect the mapping of language onto the visual world per se, but rather onto a mental world whose structure is only partly determined by the visual world as previously or concurrently experienced. This mapping manifests as an ability to predict the upcoming structure both within the linguistic (e.g., predicting upcoming phonological information) and visual domains (e.g., predicting which visual objects will be referred to next, or which visual cues afford whatever will be referred to next), and within the conceptual domain (e.g., predicting which object-representations within the mental model of the situation will be engaged by subsequent input). In the next section, we further elaborate on the implications of prediction and emergent structures for language comprehension beyond the visual world.

2. Prediction, thematic roles, and incrementality

Conceptually, Elman's (1990) SRN is simply a device that learns, for each word it is input, the range of contexts in which that word can occur. Of course, at one level, the meaning of a word is just that—knowledge of the contexts in which the word can be appropriately uttered (cf. Wittgenstein, 1953). Learning these contextual dependencies is not trivial, for as many others have observed, words can occur in numerous contexts, and the trick is to learn

just those contexts that *matter*. Given the prediction task, the SRN learns just those contextual dependencies that increase the likelihood of making a correct prediction (Altmann, 1997). In principle, once the SRN has been exposed to sufficient example sentences, it will predict nouns given verbs or adjectives, verbs given nouns, and so on. In the literature on word-to-word priming, equivalent effects have been demonstrated by McRae et al.: Ferretti, McRae, and Hatherell (2001) showed that verbs primed typical (i.e., predictable) agents, patients, and instruments. McRae et al. (2005) showed conversely that nouns can prime verbs with which they typically occur—that is, they activate the event representations in which their referents typically participate (cf. earlier discussion above). With respect to the SRN, and as argued above, these predictions from nouns to verbs and from verbs to nouns are possible because the network encodes, and abstracts across, the contextual dependencies it is exposed to during training. McRae et al. (McRae, Ferretti, & Amyote, 1997; McRae et al., 2005) have suggested that thematic roles as concepts emerge in a similar way, through generalizations across multiple experiences of events and the specific objects that participate in those events (either as experienced through direct observation of the event, or as experienced through linguistic descriptions of events). Such generalization causes us, in effect, to encode events in terms of which kinds of objects can participate in them, and how, as the events unfold, the objects participating in them undergo change through time. Conversely, objects are encoded in terms of the events they participate in (cf. *affordances*). From the perspective of language, this means that the meaning of a verb is composed, in part, from the meanings of the nouns it co-occurs with, and the meaning of a noun is composed, in part, from the meanings of the verbs it co-occurs with. In some respects, such generalizations are akin to those manifest in Elman's SRN—exposed only to word sequences, representations emerged reflecting not just the structural properties of the sentences which the SRN was exposed to (distinguishing between nouns, transitive verbs, intransitive verbs) but also what might be considered more ''semantic'' properties—distinguishing between animates and inanimates, humans, food, and breakables. Of course, the emergence of such representations from serial structure in the language is one thing, but could the same principles that enable the emergence of these representations from the language apply to the emergence of equivalent representations through observing a world that does not obey the same rigid sequential structure?

Zacks, Speer, Swallow, Braver, and Reynolds (2007) described a model of event perception in which the perceptual system continuously makes predictions about the upcoming perceptual input (the Event Segmentation Theory). A partial implementation of their Event Segmentation Theory, by Reynolds, Zacks, and Braver (2007), exposed a modified version of an SRN (a gated recurrent network; Hochreiter & Schmidhuber, 1997) to three-dimensional motion in a visual world—an animated actor represented by 18 points on the body's joints. The actor on which the animation was modeled performed a range of motions, including, for example, opening a door, sitting down, or bowing. The network was exposed to random sequences of these actions (each action was represented as a sequence of positions of the 18 points on the body in successive frames from the animation). Its task was to predict the position of the 18 points in the next time frame. The network was able to take advantage of the increased prediction error at event boundaries (i.e., its ability to better

predict successive frames *within* an event than across an event boundary) to learn to segment events at the appropriate boundary points. The network thus demonstrated how events might be segmented solely on the basis of perceptual input, without any explicit labeling. This is analogous to Elman's SRN being more accurate at predicting the upcoming word within a sentence than the first word of the next sentence. However, learning to *segment* the perceptual stream into events is not the same as learning the *internal structure* of those events (the *who did what to whom*) and abstracting across these structures to develop concepts corresponding to the individual roles that participants can play in each event (the events given to the Reynolds et al., 2007, network did not have any internal structure in terms of different participants taking on different roles within the event). In part, the challenge for any model of the emergence of event structure is to develop representations that capture the spatiotemporal contingencies that give rise, in humans, to the perception of causality; the notion of ''role'' within an event is closely tied to notions of causality and the spatiotemporal contingencies between successive states of the world. Thus, sensitivity to the temporal contingency between one state at one time and another state at a future time is a prerequisite for the emergence of event structure, and the fact that the ability to predict later states given earlier states is a manifestation of such sensitivity suggests, following Zacks et al. (2007), that prediction across time is key to the emergence of event structure. The Reynolds et al. (2007) implementation of Event Segmentation Theory is the first step to meeting the challenge of emerging event structures within an SRN-like architecture. In Elman's SRN, segmentation at sentence boundaries, and the emergence of sentence-internal structure (e.g., nouns vs. verbs, animates vs. inanimates, and so on) went hand in hand. Whether the Reynolds et al.'s (2007) model, if exposed to more complex event types involving multiple interacting objects, would develop event-internal structure, and whether it would be sensitive to the similarity across events of that internal structure (a prerequisite to emerging the conceptual equivalent of thematic roles), is an empirical (albeit computational) question. Similarly, whether an architecture that supported dynamic linguistic *and* visual input/output, with common hidden layers (cf. Dienes et al., 1999), would learn to map linguistic input onto event structure is also an empirical question.

2.1. Thematic role assignment

Notwithstanding these last computational uncertainties, the question remains of how, as a sentence unfolds, the internal state of the comprehender dynamically changes to encode not just the structure of the event being described by that unfolding language, but specifically the roles played within that structure by the specific participants in the event that are referred to by the language (i.e., how it encodes thematic role assignments). We would argue that, once again, the notion of prediction is the key to understanding this latter process. In fact, we would argue that there is no such process as thematic role assignment per se. In the earlier case of ''Jeff took a Hershey bar out of his cargo pants. He ate…,'' the activation at ''ate'' of representations that overlap with those activated earlier on hearing about the chocolate bar causes those earlier representations to change their activation state (because of the featural overlap between the two sets of conceptual representation). This

change, we argued, constitutes (or causes) a shift in the attentional state of the system, but it also encodes the relationship between eating and chocolate bars—it encodes the possible role that the Hershey bar can play in an eating event. Thus, it is this change in state that constitutes the thematic role assignment. As a sentence unfolds, conceptual entities ''receive'' event-specific roles only to the extent that they are anticipated to take part in the event that the unfolding language describes. If the example had been instead ''Jeff took a Hershey bar and a sandwich out of his cargo pants. He ate…,'' the representations corresponding to both the Hershey bar *and* the sandwich would receive a boost on hearing ''ate,'' encoding the possibility that either, or both, will take part in the eating event. And whereas in the first case, only a single conceptual entity is anticipated to take part in the eating event (and thus only the one entity ''receives'' the corresponding *theme* role), in the second case, two entities are each anticipated to take part in the eating event (in proportion to the plausibility of each being eaten), and as such, *both*, in effect, receive the associated thematic role, until such time as the unfolding language, and specifically the postverbal noun phrase, further constrains the intended event structure. This characterization of thematic role assignment is quite different from that associated with traditional linguistic formulations, in which referring expressions at different positions within a sentence would receive different thematic roles as a function of their position within the hierarchical structural configuration of the sentence (see, e.g., Pritchett, 1992). Within such formulations, an individual role is assigned to an individual entity—not to a set of entities as permitted in the account we propose here. However, this different characterization reflects the psycholinguistic data—if at ''eat'' in ''the boy will eat…'' the comprehender anticipates that a piece of cake in a concurrent visual scene is going to be referred to next (cf. Altmann & Kamide, 1999), the *theme* role must be assigned at ''eat'' (see also Boland, Tanenhaus, Garnsey, & Carlson, 1995; for a syntactically driven account of thematic role assignment at verbs in advance of their arguments). But if there are *two* edible objects in the scene, or prior context, we must assume that both objects will be entertained as plausible recipients of the *theme* role, at least momentarily before the postverbal noun phrase is encountered (a range of different data, including some of the earliest visual world studies—e.g., Eberhard, Spivey-Knowlton, Sedivy, and Tanenhaus, 1995—attest to such an assumption; cf. also Gennari & MacDonald, 2008).

This account of thematic role assignment leaves unanswered at least one major question. Huettig and Altmann (2005) showed that on hearing a word such as ''piano,'' participants would look more toward a trumpet than toward other unrelated distractors (for a similar effect induced by a discourse context, see Federmeier & Kutas, 1999). In Altmann and Kamide (2007), we argued that the same mechanism responsible for anticipatory eye movements, summarized above, could explain this latter case also—the conceptual overlap between the representation activated on hearing ''piano'' and the representation previously activated on seeing the trumpet would boost the activation of the representation corresponding to the trumpet with a consequent increase in the likelihood of launching an eye movement toward it. So why does the boost in activation of the chocolate on hearing ''eat'' in ''Jeff took out some chocolate…. He ate…'' constitute thematic role assignment when the boost in activation of the trumpet on hearing ''piano'' presumably does not? The answer,

we maintain, is that in fact there is no difference between these. The change in activation of the trumpet (that is, of the conceptual representation corresponding to the trumpet) reflects the relationship that pianos have with trumpets through category membership. Such membership in this case is an emergent property across objects that can take on similar roles in events of similar types (as well as possibly sharing other properties). The change in activation of the chocolate reflects the relationship that chocolate has with eating, and that relationship emerges across events in which objects participate in similar ways. We just happen to call one relationship ''thematic,'' and the other ''semantic.'' But within the Elman SRN, the manner in which nouns are abstracted across to form emergent representations is no different to the manner in which verbs are abstracted across. The similarity structure among different verbs comes about through the same process as that among different nouns. And the relationship between one noun and another, or between one verb and another, or between a noun and a verb, or between an agent and a patient, is again emergent through the same underlying process and manifest across the same representational substrate. The difference in relationship between trumpets and pianos, and between chocolates and eating, is a difference in nomenclature alone.

2.2. *Incrementality and prediction*

The data discussed thus far are consistent with a framework in which predictions are made simultaneously at multiple levels of representational abstraction (or ''unit'')—from fine-grained phonetic structure (e.g., McMurray, Tanenhaus, & Aslin, 2002; Salverda et al., 2003), through phonology (cf. the DeLong et al.'s [2005] study), to event representation (cf. the feathers' role in attracting anticipatory eye movements in Altmann & Kamide's [2007] study). Earlier, we distinguished between prediction as a learning task and prediction/anticipation as an ability consequent on that learning task. In an SRN trained with prediction, the task has, in effect, a grain size of a single time slice; the task, on a trial-by-trial basis, is to compare the activation pattern across the output units at time t with the activation pattern across the input units at time $t + 1$, and the temporal resolution that distinguishes between successive increments of time is fixed. However, the nature of the recurrent links at the hidden layer (Fig. 1A), coupled with the delay of one time step with which their activation is fed back into the hidden layer, means that activation of the hidden layer at time t is contingent both on the current input and on the activation state of the hidden layer at time $t - 1$, which, in turn, was contingent on its state at time $t - 2$, and so on. Thus, the activation patterns across the hidden layer are contingent on a *history* of prior states, and there is no bound (in principle, although not necessarily in practice) on the influence that previous states can have on the current state. Thus, if a contingency exists between the input at time $t - n$, and the output at time t, a network with time-delay recurrency could in principle learn this contingency. Thus, although the learning task is presented to the network with respect to the contingency between the input at time t and the input at time $t + 1$, the network can learn contingencies across different temporal resolutions. In principle, therefore, the pattern of activation at the hidden layers can simultaneously encode the anticipation of inputs at different time steps in the future (although one can also think of this as encoding its own state

at different times in the future). In this sense, the consequence of a task based on a single grain size (a single temporal resolution) is a system able in principle to anticipate across multiple grain sizes—as required to capture the different kinds of contingency, each across different time frames, between the successive inputs to which the network is exposed. Again in principle, these different contingencies will give rise to emergent representations that capture their differing temporal dynamics: a system exposed to successive acoustic–phonetic segments should learn the contingencies between, for example, subtle variations in a vowel due to co-articulation and the identity of the following consonant, and in so doing develop emergent representations that capture distinctions among phonetic segments as a function of the context in which they occur. But in principle such a system should also capture the ''higher-order'' contingencies (i.e., at different temporal resolutions) between one word and another, and in so doing develop emergent higher-order representations that capture distinctions among words as a function of the contexts in which *they* occur—resulting in multiple hierarchical representations, which emerge through exposure to systematic variation occurring across different time frames (and hence the inextricable relationship between linguistic structure and unfolding time). In this respect, there is an important distinction between prediction as a learning task with its single, given, unit of temporal incrementation, and prediction/anticipation as the consequence of this task with its multiple varying-sized units of temporal incrementation. From the perspective of human language comprehension, this bodes well for understanding how an adult system, able to represent multiple levels of hierarchical representation and anticipate what may happen next across multiple time frames, could emerge through the operation of a predictive mechanism that is initially bound to just a single temporal resolution (from time t to time $t + 1$) and just a single level of representation (the uninterpreted pattern of activity across the input units at each of time t and time $t + 1$).

3. Conclusions

Earlier, we identified four principles, enshrined within Elman's SRN (Elman, 1990, 1993) and its variants, which we believe underlie human sentence comprehension. *Mapping across domains* referred to the principle that language is not processed in isolation of the world it describes; instead, comprehension consists in realizing a mapping between the unfolding sentence and the event representation corresponding to the real-world event that is being described. *Prediction* referred to the principle that the realization of this mapping manifests as the ability to predict both how the language will unfold, and how the real-world event would unfold if it were being experienced directly. *Context* referred to the principle that sentence fragments are not interpreted in isolation from the linguistic and nonlinguistic contexts in which they unfold, and that the concurrent input *and* the internal state of the system each drive the predictive process. *Representation across time* referred to the ability to make predictions that span variable time frames and variable levels of representational abstraction (e.g., phonemes, words, event-specific roles, etc.). The data we have reviewed exemplify the operation of these principles during human language comprehension. This is

not to say that these data are the only examples of such operation. Most of the evidence we have cited in support of the principles identified earlier has involved studies of how language is mapped onto a concurrent or prior visual world. This is not the only experimental paradigm from which relevant data can be gleaned in support of these principles. For example, the findings from studies presented here overlap to a great extent with findings from studies on situation models in text comprehension (see, for example, Zwaan & Radvansky, 1998, Zwaan & Rapp, 2006; for review), including studies that have explored how spatial and temporal changes can influence the accessibility of discourse entities (e.g., Glenberg, Meyer, & Lindem, 1987; Rinck & Bower, 2000). Similarly, our account of the dynamically changing activation states of internal representations is not incompatible with accounts of discourse processing based on focus or foregrounding (e.g., Chafe, 1976; Sanford & Garrod, 1981). The paradigm discussed here has the advantage of permitting the moment-by-moment investigation of how, when, and on what basis language is mapped onto an *external, nonlinguistic, world*. It is used as a surrogate for exploring the mapping of language onto event structures. Nor are these principles the *only* principles that can be identified as central to human language comprehension (we have not discussed, for example, *constraint satisfaction* as an explicit principle [e.g., MacDonald, Pearlmutter, & Seidenberg, 1994], although it pervades the four principles we *have* discussed). Our purpose has been to identify just those principles that were captured within Elman's seminal work on the SRN (e.g., Elman, 1990, 1993) and to explore how pervasive they may be in human language comprehension. Nor have we identified all the principles which, embodied within Elman's SRN, influence the adult system; we have not, for example, expanded in any detail on the manner in which the SRN learns nor on how this leads to emergent representations. The implications of such emergence for acquisition more generally are beyond the scope of this article. And yet they are crucial (as we alluded to above). And nor are our observations particularly unique; there is a growing body of computational work under the banner of *expectation-based parsing*, drawing on insights from information theory, which is predicated on parsing as a probabilistic and inherently predictive process (e.g., Hale, 2001; Levy, 2008). Bayesian approaches to parsing (see Jurafsky, 2003; for a brief review) allow multiple sources of information, at different levels of representational abstraction, to combine to influence the unfolding probability structure of a sentence. They have the attraction that formulations of Bayes' rule at one level of description (e.g., words) can be expanded to incorporate formulations of the rule at a different level of description (e.g., phonemes). Unlike the SRN, which is usually trained on small-scale stimulus sets, these computational systems can learn from large-scale corpora. But to do so they need to know in advance what kind of input patterns exist within the world onto which they will be unleashed—Bayesian models relate the probabilities (both prior and conditional) of two units occurring (e.g., a relative clause in the context of a relative pronoun such as ''that''), but to do so, the units must be known—that is, as with any corpus-based learner, the learner must be told which things in the corpus to count. The SRN is slightly different, because it need know only about itself (which are its input units, which are its output units, and so on); it need not know a priori what kind of structure exists in the dynamically changing environment to which it is exposed. And it need not know this for one reason: structure is relevant only insofar as it is predictive of other

structure, and so long as it is, the SRN will, in principle at least, develop the appropriate representation (see Altmann, 1997, for details).

The view we are left with is of a comprehension system that is ''maximally incremental''; it develops the fullest interpretation of a sentence fragment at each moment of the fragment's unfolding. We use the term ''maximally'' to refer to the predictive component of this interpretation—the fullest interpretation should include not only all possible levels of interpretation (describable in terms of a hierarchy of structure) but also an encoding of all possible continuations of the fragment (again, described at all possible levels of interpretation). Of course, conversational goals (including participants' goals while engaged in psycholinguistic studies, as well as other nonlinguistic goals) will necessarily change the attentional state of the system (cf. Cohen et al., 2004), leading to changes in what constitutes the fullest possible interpretation of a sentence (cf. Ferreira, Ferraro, & Bailey, 2002). The ''maximal'' in ''maximal incrementality'' is thus situation dependent. However, incrementality itself arises within a system of the kind we have described because language unfolds in systematic ways across time. Incrementality merely mirrors that systematic unfolding.

One final concern: much hinges in Elman's SRN on the prediction task and its consequences. But is prediction a general property of neural systems, pervading cognition generally, or does it support only certain specific cognitive abilities? The hippocampus, other subcortical structures, and the neocortex contain recurrent excitatory connections, which support a form of (asymmetric) Hebbian learning called ''spike-timing dependent plasticity'' (STDP); a synapse is strengthened if an input spike arrives before an output spike, and it is weakened if the input spike arrives after the output spike (see Rao & Sejnowski, 2003; for review). Rao and Sejnowski (2001) demonstrated how such asymmetric learning could implement a learning rule that would result in both the prediction of neural input and the generation of temporally ordered sequences. They further proposed that cortical neurons may develop sensitivity to temporal events occurring at different timescales (cf. Montague & Sejnowski, 1994), reminiscent of our earlier discussion of prediction across multiple timescales. Most likely, therefore, prediction has a neural basis that pervades cortical function.

In summary, Elman's (1990) proposal for processing structure in time, as embodied within the SRN, captures significant insights with respect to principles that appear to underlie human language comprehension. Indeed, these principles are most likely not unique to comprehension alone (consideration of how these principles apply to language *production* is beyond this paper's remit, although language production, as with other forms of action and action planning, is predicated on the anticipation of future states). As stated at the outset, our intention is certainly not to argue that the language system is an SRN. Nonetheless, the neurophysiology of brain structures that appear to underpin human (and other species') ability to interact with the world and, in the human case, describe it, suggests that time-delay recurrence and prediction across multiple timescales are characteristic of these structures. This raises the real possibility that the high-level cognitive abilities we have reviewed here may well, someday, receive concrete grounding in neurally inspired, even neurally accurate, models of brain structure and function. For now, and most likely for a long time to come, Elman's SRN remains the most concise and succinct statement of some of the most fundamental principles of human cognition.

Notes

1. This property of Elman's SRN—the prediction task—distinguishes it from other types of recurrent network which, although including recurrence through time, do not include prediction of the input at the next moment in time (e.g., Harm & Siedenberg, 2004; Mayberry, Crocker, & Knoeferle, 2009).
2. This does not imply that the predictions need to be accurate. The principle only states that prediction is part-and-parcel of language comprehension (as it is of many other aspects of human behavior); its accuracy will depend on the extent to which previous experience with the language and the world, as encoded in the system, can generalize to the current state of the real world, linguistic and nonlinguistic.
3. The notion of affordances originates in the work of Gibson (1977) and has recently been adopted in the field of embodied cognition (Glenberg, 1997; Glenberg & Kaschak, 2002) and semantic cognition in general (e.g., Rogers & McClelland, 2004). One of the basic tenets of embodied cognition is that of a representational substrate shared between perception and action. In some accounts of embodied cognition (e.g., Barsalou, Simmons, Barbey, & Wilson, 2003), language comprehension occurs through a process of ''simulation'' (a re-enactment of the perceptuo-motor experience that would arise through directly experiencing the event described by the language). Within the framework outlined here, simulation can be equated with changes to the internal state of the substrate shared between language and different sensorimotor domains that enable predictions regarding the changes in real-world (or indeed, bodily) states that are entailed by the event structures described by that language.

Acknowledgments

The production of this article was supported by a grant from The Wellcome Trust (ref. 076702/Z/05/Z) awarded to the first author. We are grateful to Silvia Gennari for useful discussion of the theoretical issues surrounding event structure, and to Yuki Kamide for her continuing involvement in the program of empirical work described herein. We thank also Matt Crocker and two anonymous reviewers for their helpful comments on an earlier version of this paper. Finally, we thank Jeff Elman, whose friendship and support continue to inspire our work.

References

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

Altmann, G. T. M. (1997). *The Ascent of Babel: An exploration of language, mind, and understanding*. Oxford, England: University Press.

Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory and Language*, *41*, 124–145.

Altmann, G. T. M. (2002). Learning and development in neural networks: The importance of prior experience. *Cognition*, *85*, 43–50.

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

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, *284*, 875.

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

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action* (pp. 347–386). Hove: Psychology Press.

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. (2009). Discourse-mediation of the mapping between language and the visual world: Eye-movements and mental representation. *Cognition*, *111*, 55–71.

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

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, *34*, 774–806.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*, 27–38.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. N. Li (Ed.), *Language comprehension and the acquisition of knowledge* (pp. 25–56). Washington, DC: Winston.

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

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation vs. exploration. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 71–90). New York: Guilford Press.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 84–107.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507–534.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121.

Dienes, Z., Altmann, G. T. M., & Gao, S.-J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, *23*, 53–82.

Dowty, D. (1979). *Word meaning and Montague grammar*. Dordrecht, The Netherlands: Kluwer Academic.

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

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

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press/Bradford Books.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General*, *120*, 339–357.

Federmeier, K., & Kutas, M. (1999). A rose by another name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situations schemas, and thematic role concepts. *Journal of Memory and Language*, *44*, 516–547.

Fischler, I., & Bloom, P. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*, 1–20.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*, 161–187.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing*. Hillsdale, NJ: Lawrence Erlbaum.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Gleitman, L. R., & Gillette, J. (1995). The role of syntax in verb learning. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 413–428). Oxford, England: Blackwell Publishers Ltd.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1–19.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, *9*, 558–565.

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, *26*, 69–83.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL* (Vol. 2, pp. 159–166). Pittsburgh, PA.

Harm, M. W., & Siedenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Henderson, J. M., & Ferreira, F. (2004) Scene perception for psycholinguists. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action* (pp. 1–58). Hove: Psychology Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoover, M. A., & Richardson, D. C. (2008). When facts go down the rabbit hole: Contrasting features and objecthood as indexes to memory. *Cognition*, *108*, 533–542.

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.

Jackendoff, R. (2002). *Foundations of language*. Oxford, England: Oxford University Press.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach (Tech. Rep. No. 8604)*. San Diego, CA: University of California, Institute for Cognitive Science.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. Cambridge, MA: MIT Press.

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

Knoeferle, P., & Crocker, M. (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.

Laeng, B., & Teodorecu, D. (2002). Eye scan-paths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, *26*, 207–231.

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

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

Mayberry, M. R., Crocker, M., & Knoeferle, P. (in press). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*, 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, 33–42.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997a). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130.

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

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Journal of Memory and Language*, *33*, 1174–1184.

Montague, P. R., & Sejnowski, T. J. (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learning mechanisms. *Learning & Memory*, *1*, 1–33.

Morton, J. (1964a). The effect of context on the visual duration threshold for words. *British Journal of Psychology*, *55*, 165–180.

Morton, J. (1964b). The effects of context upon speed of reading, eye movements, and eye-voice span. *Quarterly Journal of Experimental Psychology*, *16*, 340–354.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance*. Chicago: The University of Chicago Press.

Rao, R. P. N., & Sejnowski, T. J. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Computation*, *13*, 2221–2237.

Rao, R. P. N., & Sejnowski, T. J. (2003). Self-organizing neural systems based on predictive learning. *Philosophical Transactions of the Royal Society*, *361*, 1149–1175.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*, 613–643.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood squares: Looking at things that aren't there anymore. *Cognition*, *76*, 269–295.

Rinck, M., & Bower, G. H. (2000). Temporal and spatial distance in situation models. *Memory & Cognition*, *28* (8), 1310–1320.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986): *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51–89.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language*. Chichester, England: Wiley.

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

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes*, *4*, 211–234.

Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, *66*, 319–327.

Tyler, L. K., & Moss, H. E. (2001). Toward a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, *5*, 244–252.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 443–467.

Wanner, E. (1980). The ATN and the sausage machine: Which one is baloney? *Cognition*, *8*, 209–225.

Wittgenstein, L. (1953/2001). *Philosophical investigations*. Oxford, England: Blackwell Publishing.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin*, *133*, 273–293.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185.

Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 725–764). New York: Elsevier.