Situated sentence processing: The coordinated interplay account and a neurobehavioral model[*]

Matthew W. Crocker
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany

Pia Knoeferle
Centre for Research on Language
UC San Diego, California, USA

Marshall R. Mayberry
School of Social Sciences, Humanities and Arts
UC Merced, California, USA

Correspondence should be addressed to:

Matthew W. Crocker
Department of Computational Linguistics
Saarland University
66123 Saarbrücken
Germany
Tel: +49 681 302 6555
Fax: +49 681 302 6561
Email: crocker@coli.uni-sb.de

[*] This research was supported SFB 378 project "Alpha" awarded to MWC and a post-doctoral fellowship awarded to PK, both funded by the German Research Foundation (DFG). The authors would like to thank Ina Bornkessel-Schlesewsky and two anonymous reviewers for their helpful comments and suggestions.

Abstract

Empirical evidence demonstrating that sentence meaning is rapidly reconciled with the visual environment has been broadly construed as supporting the seamless interaction of visual and linguistic representations during situated comprehension. Based on recent behavioral and neuroscientific findings, however, we argue for the more deeply rooted coordination of the mechanisms underlying visual and linguistic processing, and for jointly considering the behavioral and neural correlates of scene-sentence reconciliation during situated comprehension. The *Coordinated Interplay Account* (CIA; Knoeferle & Crocker, 2007) asserts that incremental linguistic interpretation actively directs attention in the visual environment, thereby increasing the salience of attended scene information for comprehension. We review behavioral and neuroscientific findings in support of the CIA's three processing stages: (i) incremental sentence interpretation, (ii) language-mediated visual attention, and (iii) the on-line influence of non-linguistic visual context. We then describe a recently developed connectionist model which both embodies the central CIA proposals and has been successfully applied in modeling a range of behavioral findings from the visual world paradigm (Mayberry, Crocker, & Knoeferle, in press). Results from a new simulation suggest the model also correlates with event-related brain potentials elicited by the immediate use of visual context for linguistic disambiguation (Knoeferle, Habets, Crocker, & Münte, 2008). Finally, we argue that the mechanisms underlying interpretation, visual attention, and scene apprehension are not only in close temporal synchronization, but have co-adapted to optimize real-time visual grounding of situated spoken language, thus facilitating the association of linguistic, visual and motor representations that emerge during the course of our embodied linguistic experience in the world.

Introduction

Much of our linguistic experience relates to the people, objects, and events in the world, sometimes even in our immediate environment. It is not surprising that people reconcile the language they hear with the world around them, and with their knowledge of the world as they have experienced it. Indeed, without such a grounding of linguistic expressions in our representations of the world it is unclear how language could have meaning. Grounding serves to both enrich our representations of sentence meaning, and draw our attention to those things in the world around us that are currently important. The observation that linguistic representations are somehow reconciled with non-linguistic perceptual representations is perfectly consistent, however, with accounts in which language understanding and visual perception work largely independently to construct (possibly incomplete) representations of their respective inputs, and in which these representations are subsequently reconciled with each other and with our general knowledge. Indeed, accounts that postulate the reconciliation of language understanding and visual perception via their autonomously constructed *representations* have a long tradition in cognitive science (Fodor, 1983; Jackendoff, 2002).

A broad range of behavioral and neuroscientific studies investigating both situated and embodied language processing have conspired to suggest that this modular view is likely inaccurate, and certainly paints a rather impoverished view of the cognitive systems under consideration. Rather, it is increasingly clear that linguistic representations are inextricably intertwined with our prior linguistic and sensorimotor experience (Barsalou, 1999b), and further that visual and linguistic representations are rapidly reconciled during situated language comprehension (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Knoeferle, Crocker, Scheepers, & Pickering, 2005). The *visual world paradigm*, in which participants' eye movements to visually present referents are monitored as participants listen to an unfolding utterance, has revealed that people automatically map the unfolding linguistic input onto the objects in their visual environment in real-time, during situated language understanding (Cooper, 1974; Tanenhaus et al., 1995). Using this method, Allopenna, Magnuson, and Tanenhaus (1998) demonstrated not only that increased inspections of visually present targets can occur within 200 ms of their mention, but also that such utterance-mediated fixations even reveal sub-lexical processing of the unfolding speech stream. Perhaps of even greater theoretical interest are the findings of Tanenhaus et al. (1995), revealing on-line interaction of visual and linguistic information for sentences such as *Put the apple on the towel in the box.* Not only did listeners rapidly fixate the mentioned objects, but their gaze also suggested the influence of the visual referential context in resolving the temporary structural ambiguity in this sentence (namely, whether *towel* is a modifier of, or the destination for, the *apple*).

One might be tempted to raise the objection that the kinds of visually situated language use investigated by such studies represents only a fraction of our linguistic activity. A number of eye-movement studies, however, have investigated how people attempt to access information about objects that are not immediately visible. Spivey and Geng (2001, Exp. 2) observed that when answering a question about either color or tilt of a target object that was no longer on the screen, participants fixated the location that the target object had previously occupied, leading the authors to argue that the external world may function as a memory, associating an object's properties with its location in the external world via *spatial indices* (Ballard, Hayhoe, Pook, & Rao, 1997). While one might worry that such effects are rather task specific, Spivey and Geng (2001, Exp. 1) monitored eye movements in a completely blank screen as people listened to spatial scene descriptions. Their findings extended previous results of Demarais and Cohen (1998), strongly suggesting that people not only "visualize" directional information in these descriptions, but also engage utterance-driven attentional mechanisms, despite the absence of any visual context (see also Johanssson, Holsanova, & Holmqvist, 2006, for further evidence and discussion). While the underlying explanation for some of these behaviors is not entirely clear, what these findings do suggest is that people employ situated comprehension mechanisms –

such as simulation, visual grounding, and spatial indexing – even when they are not engaged in canonical situated language use (e.g., a request to pass an object).

A range of behavioral results also suggest that language comprehension more generally involves the recruitment of sensorimotor patterns of brain activation to mentally represent and simulate events corresponding to sentence meaning (Barsalou, 1999b; Zwaan, 1999; Glenberg & Kaschak, 2002). Zwaan, Stanfield, and Yaxley (2002), for example, found that after reading a sentence such as *The ranger saw the eagle in the sky* participants were faster to judge a picture of an eagle as mentioned versus not mentioned in the sentence when the depicted shape (wings outstretched) matched the shape implied in the sentence (the eagle is in the sky) compared to when it did not match (perched). Stanfield and Zwaan (2001) observed related findings concerning object orientation, while Yaxley and Zwaan (2005) provided evidence that readers simulate even the visibility of described scenes (e.g., as foggy versus clear). The above observations have generally been taken as support for the seamless interaction of visual and linguistic representations (Tanenhaus et al., 1995), on the one hand, and for multi-modal sensorimotor-derived meaning representations on the other (Barsalou, 1999b). The two views are indeed complementary, since multi-modal, perceptually grounded representations in our long-term memory, and in the ensuing simulations we construct during on-line language comprehension, should naturally facilitate reconciliation with ongoing visuomotor processes or episodic traces thereof.

In this article, we take the findings above as a starting point from which to argue for a seamless temporal interdependence between real-time processing of spoken language and visual interrogation of the environment, and for investigating how such language-mediated attention "maps" onto functional brain mechanisms underlying situated language processing. We outline the *Coordinated Interplay Account* (CIA; Knoeferle & Crocker, 2006), and review both behavioral and neuroscientific findings in support of its three processing stages: (i) incremental sentence interpretation, (ii) language-mediated visual attention, and (iii) the on-line influence of non-linguistic visual context. We then describe a recently developed connectionist model which both embodies the central CIA proposals and has been successfully applied in modeling a range of findings from the visual world paradigm. Results from a new simulation with this model extend it beyond accounting for the behavioral correlates of situated comprehension, such as visual attention, to corresponding ERP effects elicited by the immediate use of visual context for linguistic disambiguation (Knoeferle, Habets, et al., 2008). Finally, we relate our account of situated comprehension to relevant neuroscientific findings on embodied language processing and to other accounts of embedded and embodied language comprehension (Barsalou, 1999a; Glenberg & Kaschak, 2002; Zwaan, 2004). We further discuss implications of a situated and embodied perspective for current theories of sentence processing. We suggest that the mechanisms underlying interpretation, visual attention, and scene apprehension are not only in close temporal synchronization, but have also co-adapted to optimize real-time visual grounding of situated spoken language, thus facilitating the association of linguistic, visual and motor representations that emerge during the course of our embodied linguistic experience in the world.

The Coordinated Interplay Account

Current theories of sentence processing have largely ignored the real-time interaction of visual and linguistic processing, despite a growing body of compelling empirical evidence. The traditional reliance of psycholinguistics on reading methodologies has led to an emphasis on purely linguistic processing, in which reading times are interpreted as reflecting processing ease or difficulty. There has been considerable emphasis, as a result, on how syntactic parsing mechanisms explain processing difficulty during ambiguity, reanalysis (Crocker, 1996; Crocker & Brants, 2000; Fodor & Frazier, 1978; Jurafsky, 1996; Vosse & Kempen, 2000; Pritchett, 1992) and linguistic complexity (Christiansen & Chater, 1999; Gibson, 1998; Hale, 2003; Levy, 2008). In addition, several theories emphasize the rapid integration of syntactic, lexical, and semantic constraints (MacDonald, Pearlmutter, & Seidenberg, 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Pado, Crocker, & Keller, in press; Trueswell & Tanenhaus, 1994). Some recent proposals have sought to reconcile and augment sentence processing accounts with ERP findings: Friederici (2002) builds upon the syntax-first proposal of Fodor and Frazier (1978), associating processing stages with specific ERP correlates (see also Bornkessel & Schlesewsky, 2006). Hagoort (2003), in contrast, draws on the model proposed by Vosse and Kempen (2000). Nonetheless, these neurolinguistic accounts of sentence comprehension share a continued emphasis on language processing in isolation.

Among current theories of sentence processing, the interactive constraint-based theories (e.g., MacDonald et al., 1994; Tanenhaus, Spivey-Knowlton, & Hanna, 2000) are often cited as a natural framework to account for the on-line interaction of visual and linguistic representations. Implemented models, however, address only the high-level dynamics in which a set of pre-specified interpretations is activated (McRae et al., 1998; Tanenhaus et al., 2000; Spivey, 2007). Thus, even when such models include visuomotor constraints (Farmer, Anderson, & Spivey, 2007), they still shed little light on how the mechanisms of incremental sentence understanding interact with visual perceptual processes, and vice versa. Embodied accounts of language processing, while explaining the resonance that exists between language and visuomotor representations (Zwaan et al., 2002; Glenberg & Kaschak, 2002) as well as temporal aspects of the simulation of events (Claus & Kelter, 2006), have neither paid much attention to the compositional mechanisms of language comprehension and their time course (but see Glenberg & Robertson, 1999; Zwaan & Taylor, 2006) nor to the development of implementable computational models with broader linguistic coverage (see Crocker, 2005, for discussion).

To provide an account of the mechanisms that enable language comprehension to seamlessly draw on non-linguistic visual information in real time, as well as to bridge the traditional sentence processing perspective with that of embodied accounts, Knoeferle and Crocker (2006) sketched the *Coordinated Interplay Account* of situated utterance comprehension. Based on findings from the visual worlds paradigm the CIA accords a central role to visual attention as a mechanism for grounding language understanding: Incremental interpretation of unfolding speech directs visual attention towards mentioned and anticipated objects and events in the scene, and is in turn influenced by the attended scene information. This close temporal interaction entails that visual inspection of relevant objects and events will often occur shortly before/after their mention, maximizing their salience for language understanding (see Knoeferle & Crocker, 2006, for relevant findings). As can be seen from Figure 1, the CIA consists of three informationally and temporally dependent stages. *Sentence interpretation* corresponds closely to the processes of incremental sentence comprehension which are the focus of traditional sentence processing accounts. *Utterance mediated attention* identifies those aspects of the current interpretation which contribute to utterance-mediated shifts in visual attention. *Scene integration*, finally, identifies which aspects of visual representations then in turn inform *interpretation*. It is important to note that the CIA itself makes no assumptions regarding the modular status of either the linguistic or visual processes involved. Rather, the CIA outlines the coordinated interaction of linguistic and visual processing, and specifically the temporal constraints resulting from both the unfolding linguistic input, and changes in salient scene information due to utterance mediated shifts in visual attention.

The following subsections bring to bear a range of behavioral and neuroscientific evidence in support of the processing stages of the CIA. Our aim is firstly to show that the interaction of linguistic and visual processing is bidirectional, and pervades many levels of processing. Furthermore, evidence from eye movements in visual scenes and event-related potentials conspire to provide compelling evidence for the CIA's claim that the cognitive processes and functional brain mechanisms that underlie situated comprehension have adapted so as to enable real-time coordination and interaction.

Sentence Interpretation

We assume that the mechanisms of situated language comprehension intersect with the general mechanisms of compositional sentence processing, as evidenced by eye tracking, electrophysiological and neuro-imaging findings. As a relatively high-level account of situated language processing, the CIA is neutral with respect to any specific account of compositional linguistic processing but rather asserts that sentence processing mechanisms exhibit the characteristics of *incremental*, *predictive* and *integrative* processing, each of which is examined separately below.

That written and spoken language comprehension is highly *incremental* has been established since the

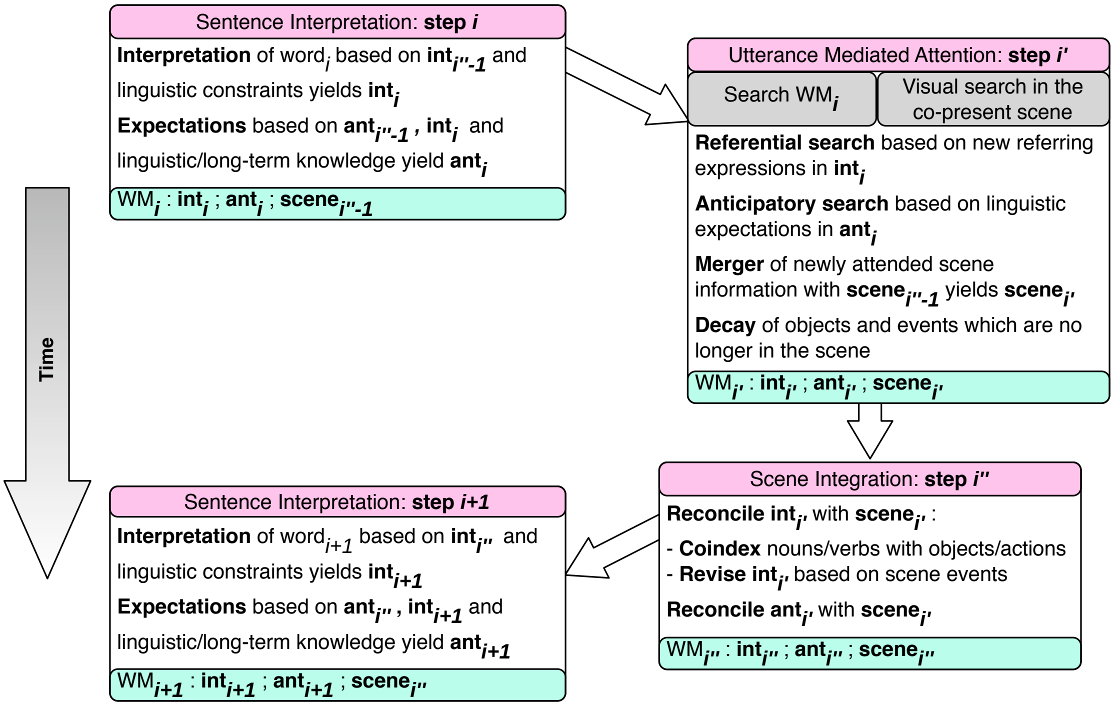
Sentence Interpretation: step i
Interpretation of word$_i$ based on int$_{i''-1}$ and linguistic constraints yields int$_i$
Expectations based on ant$_{i''-1}$, int$_i$ and linguistic/long-term knowledge yield ant$_i$
WM$_i$: int$_i$; ant$_i$; scene$_{i''-1}$
Time
Utterance Mediated Attention: step i'
Search WM$_i$
Visual search in the co-present scene
Referential search based on new referring expressions in int$_i$
Anticipatory search based on linguistic expectations in ant$_i$
Merger of newly attended scene information with scene$_{i''-1}$ yields scene$_{i'}$
Decay of objects and events which are no longer in the scene
WM$_{i'}$: int$_{i'}$; ant$_{i'}$; scene$_{i'}$
Scene Integration: step i''
Reconcile int$_{i'}$ with scene$_{i'}$:
- Coindex nouns/verbs with objects/actions
- Revise int$_{i'}$ based on scene events
Reconcile ant$_{i'}$ with scene$_{i'}$
WM$_{i''}$: int$_{i''}$; ant$_{i''}$; scene$_{i''}$
Sentence Interpretation: step i+1
Interpretation of word$_{i+1}$ based on int$_{i''}$ and linguistic constraints yields int$_{i+1}$
Expectations based on ant$_{i''}$, int$_{i+1}$ and linguistic/long-term knowledge yield ant$_{i+1}$
WM$_{i+1}$: int$_{i+1}$; ant$_{i+1}$; scene$_{i''}$

Figure 1: **The Coordinated Interplay Account (CIA):** Processing of $word_i$ incrementally updates representations related to the current interpretation (int_i), the expectations the interpretation generates (ant_i), and the salience/activation of scene objects and events in memory ($scene_i$). This is accomplished by three temporally dependent processing stages: incremental *sentence interpretation* (step i), *utterance mediated attention* (step i') towards mentioned or anticipated scene entities, and finally *scene integration* which reconciles the interpretation with relevant scene information (step i'') (Knoeferle & Crocker, 2007).

pioneering work of Marslen-Wilson (1975), Bever (1970), and Frazier (1979) (see also Crocker, 1999, for review). By incremental interpretation, we mean that each word is structured into the linguistically well-formed and interpretable (if partial) representation of the sentence fragment that has been read or heard so far. Numerous ERP studies have demonstrated the incrementality of language comprehension as revealed by the on-line detection of semantic (e.g., Kutas & Hillyard, 1980, 1983; van Petten & Kutas, 1990) and syntactic (e.g., Osterhout & Holcomb, 1992, 1993; Matzke, Mai, Nager, Rüsseler, & Münte, 2002) violations, as indexed broadly by deflections in scalp activation such as the so-called N400 and P600 (see also Kutas, van Petten, & Kluender, 2006, for overview and discussion). Incremental processing is also revealed by visual world studies examining the incremental use of adjectival modifiers in narrowing down possible referents (e.g., Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Weber, Braun, & Crocker, 2006), and in studies that show motor representations are activated during incremental (self-paced) reading, precisely in the region where a relevant verb (that mediated a motor action) was being read (Zwaan & Taylor, 2006).

In fact, comprehension is not just incremental, but *predictive* (e.g., Federmeier, 2007). Altmann and Kamide (1999) demonstrated that listeners exploit the selectional restrictions of verbs like *eat* to anticipate edible objects in the scene. While it might be speculated that such prediction only takes place in highly circumscribed visual contexts (i.e., when language restricts the domain of potential referents to a subset of the displayed objects), evidence from reading (Altmann, 1999) and ERP studies (Kutas & Hillyard, 1984) suggests that generation of semantic expectations is a more general characteristic of on-line comprehension. Indeed, predictive mechanisms are common to many recent processing accounts (Crocker, 1996; Gibson, 1998; Konieczny, 2000; Levy, 2008). Furthermore, these general predictive mechanisms appear to be at work even in highly circumscribed visual contexts: Weber, Crocker, and Knoeferle (in press) presented auditory sentence fragments such as *The woman bakes* as listeners viewed a display containing a woman and several objects, and then performed a lexical decision task on a word presented on the screen. Facilitated lexical decision times for appropriate items (e.g., *pizza*), regardless of the scene, provide evidence for a purely linguistic anticipation of upcoming verb arguments. However, they also report a slowing of lexical decision times when there was a plausible depicted referent (e.g., a picture of *cake*) which differed from the lexical decision target, suggesting that visual grounding of general expectations with a specific object instantiated the linguistically-driven expectations concerning which object would follow as the verb argument; when the visually-supported expectations were not met by the target word, lexical decision times were slowed regardless of whether the word was appropriate. Such lexically specific prediction is also elicited by linguistic contexts: DeLong, Urbach, and Kutas (2005) found that when an indefinite article (e.g., *an*) mismatched the noun (e.g., *kite*) that people expected, the amplitude of the N400 to the article varied as a function of the cloze probability of the expected noun. Similarly, van Berkum, Brown, Zwitserlood, Kooijman, and Hagoort (2005) showed that in a linguistic context (but not in the absence of such context) information provided by the determiner about the expected gender of an upcoming noun was used as early as 50 ms after participants heard a (gender-incongruent versus congruent) inflectional ending of a pre-nominal adjective (see also Otten, Nieuwland, & van Berkum, 2007).

A third characteristic of comprehension is the rapid on-line *integration* of diverse sources of information – lexical, syntactic and semantic – as evidenced by findings from numerous reading and visual world studies, (e.g., Kamide, Scheepers, & Altmann, 2003; McRae et al., 1998; Pickering & Traxler, 1998; Trueswell, Tanenhaus, & Garnsey, 1994), and as reflected by an increasing number of processing models (MacDonald et al., 1994; Pado et al., in press; Spivey, 2007; Tanenhaus et al., 2000). The visual world paradigm has further revealed the rapid influence of prosodic information in disambiguating word-order ambiguity (Weber, Grice, & Crocker, 2006) as well as of stress for identifying contrasting referents (Weber, Braun, & Crocker, 2006). Perhaps most important with regard to the present proposal is the evidence regarding the on-line influence of non-linguistic information, such as visual referential context (Tanenhaus et al., 1995), object affordances (Chambers, Tanenhaus, & Magnuson, 2004), depicted events (Knoeferle et al., 2005), and motor resonance (Zwaan & Taylor, 2006). In addition, fMRI findings suggest that when participants relate information from pictures to a sentence, both typical language processing areas (the left posterior temporal gyrus) and visual-spatial processing areas (the left and right parietal areas) are activated (Carpenter, Just, Keller, Eddy, & Thulborn, 1992). Further evidence for the rapid contribution of non-linguistic visual representations to incremental sentence comprehension comes from ERP studies (e.g., Ganis, Kutas, & Sereno, 1996; Wassenaar & Hagoort, 2007; Knoeferle, Habets, et al., 2008; Knoeferle, Kutas, & Urbach, 2008). It is precisely this influence of non-linguistic, particularly visual, constraints that have not been reflected in current accounts of compositional sentence processing. Since this is a central part of the CIA, we return to the issue of scene influence in greater detail below.

As mentioned above, the CIA is not intended as a theory of the compositional mechanisms underlying sentence comprehension *per se*, but rather outlines the framework within which such mechanisms operate interdependently with our visual interrogation of the environment. The hallmark characteristics identified above strongly delimit the space of candidate sentence processing mechanisms, however. Specifically, probabilistic mechanisms, whether formulated in statistical or connectionist terms, likely play an important overarching role. Such accounts have been shown to offer elegant explanations for lexical frequency effects, ambiguity resolution (Jurafsky, 1996; Crocker & Brants, 2000; Crocker, 2005), and processing difficulty (Hale, 2003; Levy, 2008). They also provide a potential means for expressing the strong role that both linguistic (Chater & Manning, 2006), and associated visuomotor (Barsalou, 1999b; Fischer & Zwaan, 2008), experience play in comprehension, and also offer a natural means for characterizing non-deterministic probabilistic inference, which may underlie at least some aspects of prediction. We return to these issues below in our discussion of the CIANet implementation of the CIA (Mayberry et al., in press).

Utterance-Mediated Attention

Utterance processing and interpretation has been shown to rapidly direct visual attention in a related scene, suggesting an overarching strategy of continuously mapping unfolding speech onto the world that surrounds us. Such utterance-mediated attention shift often occurs within 200 ms of hearing the linguistic stimulus (Tanenhaus et al., 1995) – barely enough time to program a saccade (see Appendix I in Altmann & Kamide, 2004; Matin, Shao, & Boff, 1993). In line with existing findings, the CIA assumes that multiple levels of linguistic processing including, but not necessarily limited to, lexical access, reference and pronoun resolution, and the anticipation of predicted role-fillers, rapidly drive visual attention in the scene. Beyond the direct mapping of names onto displayed objects (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001), there is substantial evidence for more general referential attention processes. Tanenhaus et al. (1995), for example, observed the on-line resolution of referential ambiguity via a prepositional phrase modifier, while numerous other studies have demonstrated the incremental use of adjectival modifiers in narrowing down possible referents (Sedivy et al., 1999; Spivey et al., 2001; Weber, Braun, & Crocker, 2006), as well as the rapid identification of likely referents for pronouns (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000). Further, is has been shown that referential eye movements do not only occur in response to mention of people or objects, but also to events (Knoeferle et al., 2005).

Utterance-mediated attention can further reflect prediction in comprehension: People actively anticipate likely thematic role fillers as revealed by shifts in attention to appropriate objects in the scene before they are mentioned (Altmann & Kamide, 1999). While inspecting a scene that showed, for example, a boy, a toy train and car, a ball, and a birthday cake, people's inspection of the cake was increased when they heard *The boy will eat ...* versus *The boy will move ...*, indicating that people rapidly use verb semantics to anticipate likely patients of the verb. Such anticipatory looks are also sensitive to linguistic (Kamide et al., 2003) and prosodic (Weber, Grice, & Crocker, 2006) constraints on grammatical function and thematic roles imposed by compositional interpretation. Knoeferle et al. (2005) also found evidence for the anticipation of role fillers based on participants in verb-mediated depicted scene events.

Perhaps most surprisingly, attention shifts in response to utterance interpretation even occur when relevant objects are no longer present. As we noted in the introduction, it has been shown that people fixated the previous locations of absent objects when asked questions about them (Spivey & Geng, 2001; Richardson & Spivey, 2000). The findings were extended to passive comprehension by Altmann (2004): Using stimuli similar to Altmann and Kamide (1999), people first inspected the scene, which was then removed before people heard *The boy will eat* . . . . Eye movements in the "blank screen" closely resembled the time-course of inspections found in the original study. These findings point to the general deployment of spatial indexing mechanisms (Spivey, Richardson, & Fitneva, 2004), suggesting that mental representations of a prior scene are still indexed in the world during comprehension (see Knoeferle & Crocker, 2007, for studies on events) .

One function served by language-mediated visual attention likely is to ground the described objects and events with the visual environment, thus enabling situation-enriched interpretation. This view is supported by the findings that visual search processes are made more efficient by this close coordination with linguistic processes. Spivey et al. (2001) argue, for example, that a concurrent spoken instruction such as *Is there a red vertical?* facilitated conjunctive visual search more than when the target was identified prior to image onset. Related work by Lupyan (2007) further shows that visual search, where participants already know the target, is facilitated by a preceding statement labeling the distractor (e.g. *ignore 5*) or the target (e.g. *find* *the 2*). Lupyan argues that the category labels and their associated visual features, as provided by unfolding speech, may in fact facilitate visual processing of member categories in the scene by influencing the object-selective regions of the cortex to improve response to desired visual targets. Taken together these findings lead us to speculate that visual processes do not simply respond to linguistic directives, but that visual search and perception mechanisms may be directly modulated by ongoing interpretation processes and active linguistic representations. Equally, visual search and perception are likely to be highly influential in triggering perceptuo-motor simulation based on scene information, and thus constitute an essential link between visually situated and embodied aspects of language comprehension.

Scene Integration

A consequence of the utterance-mediated attention mechanisms is that our attention is drawn rapidly to relevant aspects of a possibly quite complex scene. A central claim of the CIA is that such salience of relevant scene information, synchronized with the unfolding utterance, results in both the use, and relatively high importance, of scene information. Our goal in this section is to review eye-tracking and ERP evidence for the influence of more general and natural visual contexts on incremental comprehension, including the referential context, scene events and affordances.

As mentioned above, Tanenhaus et al. (1995) demonstrated the rapid influence of visual referential context on ambiguity resolution in on-line situated utterance processing. Listeners were presented with a scene showing either a single apple or two apples, and the utterance *Put the apple on the towel in the box*. Eye-movements revealed that the interpretation of the phrase *on the towel* as either the location of the apple versus its desired destination was influenced by the visual context manipulation. Sedivy et al. (1999) further demonstrated the influence of a visual referential contrast: Listeners looked at a target referent (e.g., *the tall glass*) more quickly when the visual context displayed a contrasting object of the same category (a small glass) than when it did not. As noted above, Weber et al. (in press) also found evidence in a visually situated lexical decision task, suggesting that anticipatory inspection of the appropriate object based on preceding linguistic input sets up contextually grounded expectations concerning which object would follow as the verb argument; when the visually-supported expectations were not met by the target word, lexical decision times were slowed across the board.

Knoeferle et al. (2005) further demonstrated that people are able to resolve initially structurally ambiguous simple subject-verb-object (SVO) and object-verb-subject (OVS) sentences based purely on depicted scene event information. Listeners heard an utterance beginning (*Die Prinzessin malt* . . . , “The princess paints . . . ”) where the verb – per the event depictions – identified the princess as either the patient or agent of a painting event in a co-present visual scene. Anticipatory eye movements to yet-to-be-mentioned role-fillers (e.g., the person painting, or being painted by, the princess, respectively) revealed that they rapidly integrated information from depicted actions. In a corresponding auditory ERP study, Knoeferle, Habets, et al. (2008) found a P600 time-locked to the verb when it identified a depicted event that forced disambiguation towards the dispreferred OVS (when the event referred to by the verb showed the princess as patient) rather than SVO (when the relevant event showed the princess as agent) interpretation. This finding strongly corroborates the claim that listeners can immediately use a co-present depicted event to inform structural disambiguation during auditory sentence comprehension. Differences in the time course with which depicted events are used (early vs. late) varied as a function of when the utterance identifies them as relevant (early vs. late) provide further evidence for the temporal coordination of utterance processing, visual attention and the use of depicted events for comprehension (Knoeferle, 2007). Evidence from ERPs in a picture-sentence verification study have confirmed the rapid use of depicted actions. Participants read NP1-VERB-NP2 sentences, word by word, after viewing an agent-action-patient event that matched or mismatched semantic/referential aspects (e.g., verb-action reference) of the sentence; their task was to verify whether or not the sentence matched the prior scene. Verification latencies were slower, and N400 (300-500 ms) mean amplitudes to the verb larger, when the verb did not match vs. matched the depicted action (Knoeferle, Kutas, & Urbach, 2008). N400 latency and centro-parietal maximum closely resembles that elicited by words in lexical, sentential and discourse contexts (Kutas et al., 2006).

Knoeferle and Crocker (2006) conducted a series of experiments that focussed on the time course with which linguistic knowledge (Altmann & Kamide, 1999) and information from the scene interacted (Knoeferle et al., 2005), and, crucially, the *relative importance* of these sources of information. Their findings confirmed that people quickly use whatever information is available to anticipate thematic role fillers. Additionally, when they pitted the relative importance of depicted events for thematic role assignment against stereotypical thematic role knowledge, people showed a clear preference for the depicted information over their world knowledge. Knoeferle and Crocker (2007) present three experiments investigating the temporal interdependency between *dynamic* visual context and utterance comprehension. Exploiting the "blank screen paradigm" discussed earlier, event scenes were presented prior to the onset of an utterance and then replaced by a blank screen either before or during the utterance. Additionally, two of the experiments featured scenes involving dynamic events, i.e., actions were depicted as occurring over time, introducing an aspectual dimension to the depicted events, which were furthermore coupled with verb and adverb tense manipulations in the utterances used in the third experiment. The findings suggested that people do exploit scene event information even when it is no longer present, but that the relative priority with respect to other information sources is strongest when events are co-present.

Beyond referential context and scene events, there is also evidence that other aspects of the scene and affordances inform comprehension. Chambers et al. (2004), for example, found that listeners were sensitive to the affordances of task-relevant objects (e.g., whether an egg was broken, and hence pourable, or not) with respect to the action required by the instruction (e.g., *Pour the egg in the bowl over the flour*). These findings suggest that utterance interpretation is guided by the listener's situation-specific evaluation of how to achieve the behavioral goal of an utterance.

In sum, there is considerable behavioral and neuroscientific evidence suggesting that comprehension is not only influenced by the visual context, but further that utterance-mediated visual attention precisely serves to optimize the exploitation of the scene, and in some cases affords it priority over our experience-based expectations. The kinds of visual influences on comprehension are further highly varied, ranging from visual grounding of linguistic expectations to the more inferential exploitation of referential and event context. The CIA here crucially asserts that visual information, particularly that identified by the utterance as relevant, becomes highly salient for the unfolding interpretation and disambiguation of situated spoken language.

A Neurobehavioral Computational Model

At the heart of the CIA is the claim that utterance-mediated attention in the visual context is not only driven by incremental and anticipatory linguistic processing, but crucially that it is this modulation of visual attention that underpins both the use and salience of the relevant visual context. CIANet was developed to instantiate the central claims of this proposal and evaluate them computationally (Mayberry et al., in press). The architecture is based on a simple recurrent network (SRN; Elman, 1990) that produces a case-role interpretation of the input utterance. The choice of a connectionist approach was motivated by the requirements developed in our discussion of linguistic interpretation above. Processing in an SRN is incremental, with each new input word interpreted in the context of the sentence processed so far, as represented by a copy of the previous hidden layer which serves as additional, contextual input to the current hidden layer. The model is able to exploit distributional information accrued during training to learn syntactic constraints such as constituent order and case marking, lexical constraints on likely role-fillers for particular verbs, as well as correlations between utterance meaning and the characters and events in the visual context. The integration of both kinds of knowledge – long-term experience and immediate visual context – contributes both to interpretation and and the non-deterministic probabilistic anticipation of likely role-fillers. The network architecture is also adaptive, in that it is trained to perform incremental thematic interpretation both with and without a scene. Scene contexts contain characters and actions that explicitly depict relationships between them, but which may not always be fully relevant to the utterance being processed.

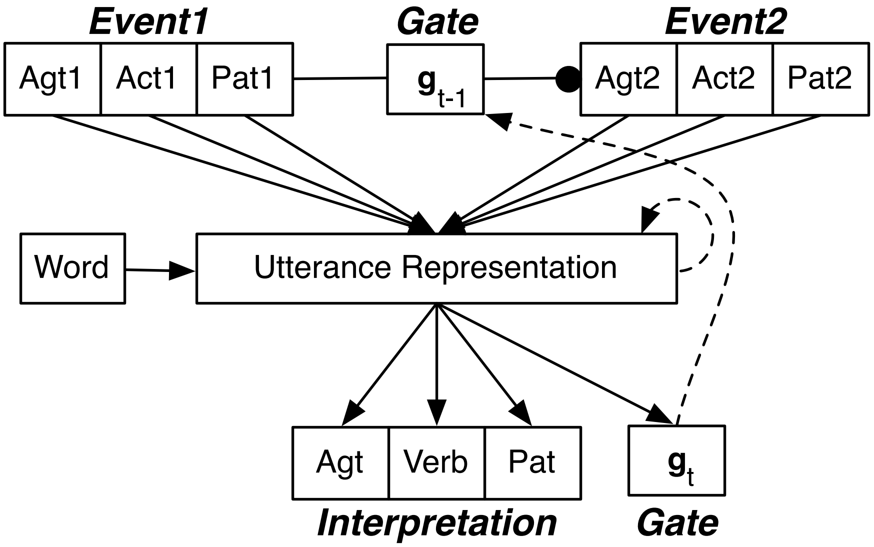

Figure 2: **CIANet**: A network featuring scene-language interaction with a basic attentional gating mechanism. As a sentence is processed one word at a time, the developing utterance representation in the hidden layer activates the attentional gating mechanism to select the event in the scene most relevant to the interpretation of the situated unfolding utterance.

Importantly, however, we do not claim CIANet is a model of the mechanisms that underlie sentence comprehension: The processing architecture, lexical representations, and linguistic scope are too limited (see Mayberry, 2003; Mayberry & Miikkulainen, submitted, however, for a related architecture with greater coverage). Rather, we propose CIANet as a model of the mechanisms underlying the interaction of interpretation, attention, and scene integration. Just as with the CIA, CIANet should not be taken as making any claims regarding the modularity of linguistic and visual processes: the unrestricted interaction of linguistic and visual information is rather intended to enable direct investigation of the temporal dependencies that emerge from incremental sentence processing and resulting shifts in visual attention. In this section we briefly sketch the model, its embodiment of the CIA, and the central behavioral findings accounted for by the model. We then consider the relevant ERP findings of Knoeferle, Habets, et al. (2008), and provide a linking hypothesis for CIANet which accounts for those findings, and also those of Matzke et al. (2002) (see Kluender & Kutas, 1993; Rösler, Pechmann, Streb, Röder, & Hennighausen, 1998, for related ERP findings on syntactic processing of word order variations).

As shown in Figure 2, CIANet incorporates visual input through an additional input representation of a scene as (optional) visual context for the input utterance. Scenes contain two events, only one of which is relevant to the input utterance. Each of the two scene events has three constituents (*agent, action* and *patient*) that are propagated to the SRN's hidden layer through shared weights representing a common post-visual-processing pathway. Since the constituents of the two events are superimposed, the network must learn to extract the information from the event that is most relevant to the target interpretation by modulating the event representations. Lexical representations are 144-unit, random binary vectors that fill the Word input and each constituent in both events and interpretation. The gate is similarly a 144-unit vector in order to permit elementwise modulation of the two event constituents' activation. The hidden layer consists of 400 units. A more detailed presentation the CIANet architecture, and how it is trained, is provided by Mayberry et al. (in press).

In line with the language-mediated visual attention mechanisms of the CIA, the unfolding linguistic input to CIANet modulates the salience of the relevant scene event based on the emerging interpretation in the hidden layer. A gating vector implements the attentional mechanism in CIANet, transforming the architecture into a basic recurrent sigma-pi network (Rumelhart, Hinton, & Williams, 1986), in which nodes may be multiplied as well as added together. The units of the gate are multiplied element-wise with the corresponding units in each of the three lexical representations (agent, action, and patient) of an event (see Figure 2). Each unit of the gate is subtracted from 1.0 to derive a vector complement that then modulates the other event. This means that more attention to one event in the model entails less attention to another. Crucially, the network is never explicitly taught which event in the scene it should attend to. Rather, the gate is optimized to increase the contrast between the constituents of the two events. This is achieved during training by recurrently back-propagating error information from the multiplicative connections of the gate to the modulated constituent representations of each event. Consequently, the average activation of the gate's units directly correlates with greater activation of the attended event in a scene. Accordingly, attention is driven by correlations with the roles of arguments in the scene events and the linguistic aspects of the input utterance, such as case-marking and stereotypicality.

Linking hypothesis: behavioral

CIANet was intended to account for several of the key behaviors that have been found within the visual worlds paradigm, including incremental interpretation, anticipation of role fillers, and the influence of the scene (Mayberry et al., in press). In order to computationally investigate these phenomena, CIANet was developed to model the use of depicted events (Knoeferle et al., 2005, Experiment 1), and the relative priority of depicted versus stereotypical events (Knoeferle & Crocker, 2006, Experiment 2). Regarding the use of depicted events, CIANet models observed gaze behavior concerning how people used depicted relations about who-does-what-to-whom to facilitate prediction of thematic role relations in the sentence at a point when the linguistic input was ambiguous (see discussion of Knoeferle et al., 2005, Experiment 1, above). In both experiment and model, the first noun phrase of the linguistic input (e.g., *The princess*) could serve either as the subject (agent) or the object (patient) of the sentence. Recall that for subject-verb-object sentences the verb *painting* mediated a depicted event that showed the initially ambiguous noun phrase as an agent (princess-paints-fencer); for object-verb-subject sentences, in contrast, the verb *washes* mediated a different depicted event in which that referent was the patient (pirate-washes-princess). Shortly after hearing the verb, human participants inspected the *fencer* (the patient of the princess-painting event) more than the *pirate* for structurally ambiguous subject-initial sentences, and inspected the *pirate* (the agent of the pirate-washes-princess event) more than the *fencer* for object-initial sentences. The findings were taken to reflect the rapid influence of scene events for incremental and anticipatory role assignment to both the initial noun phrase, the *princess*, and the expected noun phrase (the *fencer*-patient and *pirate*-agent for subject- and object-initial sentences, respectively). In the model, the activation of target output representations of the *fencer*-patient and *pirate*-agent was taken as a comparable index for "anticipation" of role fillers and underlying assignment of a thematic role to the first noun phrase. As with the human participants, the model displays increased activation of the *fencer* for structurally ambiguous subject-initial relative to object-initial sentences and increased output activation of the *pirate* for object-initial relative to subject-initial sentences. Furthermore, the correct role (as indexed by the model's case-role representation) is assigned to both the initial and the anticipated NP. CIANet's performance thus qualitatively agreed with the empirical results: The network accurately predicted the appropriate role fillers at the same point in time during the sentence (immediately after the verb) as people did (as reflected by their gaze).

CIANet also modeled the findings of Knoeferle and Crocker (2006, Experiment 2), concerning the relative priority of (non-stereotypical) depicted events (Knoeferle & Crocker, 2006) versus verb-based semantic knowledge (Altmann & Kamide, 1999). As noted earlier, Knoeferle and Crocker (2006) found not only that people could use either source of information equally well, when only one was relevant, but crucially that when the two information sources conflicted, depicted event information took priority. CIANet was only trained on two conditions where the verb was unambiguously compatible with one of the two available agents (either through stereotypical verb-agent association or per the depicted event an agent performed). When participants heard a German sentence that described the pilot as being *enchanted*, they quickly (post-verbally) looked at the wizard because that agent is closely associated with the verb (*enchants*). The model correspondingly showed the agent constituent for *the wizard* as more activated post-verbally than the representation of the competing target agent (*the detective*). In contrast, when humans/CIANet processed a sentence that described the pilot as being served food to, people preferentially looked at *the detective*, who was the agent of the food-serving depicted event corresponding to the verb *serves* and the model showed a higher activation of the target output for *the detective* than for *the wizard*. When tested on conflicting conditions, which the network had never been exposed to, CIANet revealed the same pattern of behavior as that revealed by eye-movements in the human study: When processing a verb that identified two different scene agents, both humans and CIANet revealed a clear preference for predicting the agent that was depicted as performing the verb action (even though that action was implausible) rather than a different agent that was a stereotypical for the verb (Knoeferle & Crocker, 2006). The reason the model predicts this behavior is because of the attentional vector which increases the salience of the relevant depicted event, precisely as outlined by the CIA.

In sum, CIANet demonstrated the hallmark characteristics of the human eye-gaze data as described above: the interpretation is developed incrementally as each word is processed; likely upcoming role fillers are accurately anticipated; utterance and scene are integrated in real-time; and the model is able to adapt to the presence as well as the absence of the scene. Analysis of the gating vector shows that it basically acts like a scalar, because most of its units are either highly activated (i.e., close to 1.0) or not activated (close to 0.0). When the gating vector is then multiplied with each element of each constituent in an event, the result is that the most relevant event is selected (or "attended to"), in keeping with the empirical evidence.

Linking hypothesis: neural

Knoeferle et al. (2005) interpreted the eye-movement data as reflecting disambiguation of the local structural ambiguity on the first noun phrase of their sentences (i.e., whether the first noun phrase is the subject/agent or object/patient of the sentence). Eye movements, while providing behavioral evidence for the claim that scene information affects the incremental disambiguation of initially structurally ambiguous utterances, may also reflect various other underlying linguistic and nonlinguistic processes such as semantic interpretation (Sedivy et al., 1999), thematic interpretation (Altmann & Kamide, 1999), and visual search (Spivey & Geng, 2001). Furthermore, the findings do not clarify whether the resolution of local structural ambiguity through scene information exploits the same functional brain mechanisms as when disambiguation occurs through linguistic cues.

To address these two points, Knoeferle, Habets, et al. (2008) recorded ERPs during the processing of stimuli that were virtually identical to those of Knoeferle et al. (2005, Experiment 1, see above). Prior research (Matzke et al., 2002) had observed a positivity time-locked to linguistic cues such as case marking that disambiguated a temporarily ambiguous, sentence-initial noun phrase towards the dispreferred, object-initial structure. That positivity occurred approximately 600 ms after the onset of the disambiguating word, was larger when case marking disambiguated towards the dispreferred (e.g., object-initial) rather than the preferred (e.g., subject-initial) structure in German, and had a centro-parietal maximum ('P600'). Knoeferle, Habets, et al. (2008) relied on the findings by Matzke et al. (2002) as an index of structural disambiguation for locally structurally ambiguous German sentences (see also beim Graben, Schlesewsky, Saddy, and Kurths (2000) and Frisch, Schlesewsky, Saddy, and Alpermann (2002)).[1] Knoeferle, Habets, et al. (2008) predicted that if people structurally disambiguate the initial structural ambiguity of the first noun phrase and verb as soon as the verb has identified a scene event that shows who-does-what-to-whom, then they should observe a P600 time-locked to the verb: Specifically, when the verb identified a *visually disambiguating* scene event that forced the disfavored object-verb structure, ERPs should show a larger amplitude of the P600 relative to ERPs when the verb mediated a scene event that confirmed the subject-verb interpretation. Not only did they find the predicted P600 at the verb (suggesting disambiguation through verb-mediated events) in conditions when the scene was present, but they also replicated previous results when no scene was present, namely a P600 at the linguistically disambiguating second noun phrase, but none at the verb.

One interesting question concerning these findings in light of CIANet is how to link the model not only to the eye-tracking but also to the ERP data, thus taking a first step towards turning CIANet into a neurobehavioral model. Given that the P600 is typically associated with a substantial revision of the current interpretation of the utterance, we would expect this revision to be reflected by the internal representation of the network. Specifically, when such a major revision takes place, we would expect the hidden-layer activation of the network to undergo a substantial change within a condition c (e.g., unambiguous with scene) from one time step t_{n-1} to the next time step t_n (denoted by δ_n^c). As an initial, qualitative linking hypothesis relating the hidden-layer activation with the ERP findings by Knoeferle, Habets, et al. (2008), we therefore hypothesized that if the model performs structural revision at the point in time when Knoeferle, Habets, et al. (2008) observed a P600 (at the verb), then the vector distance Δ_{adv} between δ_{adv}^{ovs} and δ_{adv}^{svo} (i.e., between verb and the adverb that immediately follows it) should be greater for the dispreferred OVS structures (when the verb mediates events that disambiguate towards the dispreferred OVS structure) compared to SVO sentences (when the verb and its associated events confirm the initially preferred subject-first interpretation).

As CIANet also processes sentences correctly in the absence of scenes, we further examine a possible correlation between the hidden layer activation changes of CIANet and the ERP findings of Matzke et al. (2002). Specifically, Matzke et al. (2002) found a larger negativity with a left anterior maximum (LAN) to the first word of unambiguous OVS vs. SVO sentences. In addition, they observed a further negativity to the second determiner of unambiguous sentences: the second determiner of OVS sentences elicited a larger

[1] Importantly, however, while the P600 is reliably associated with the object-initial main clause (verb second) structures investigated here, that association is not observed for all constructions involving object-initial disambiguation (e.g., in complement clauses: Bornkessel, McElree, Schlesewsky, & Friederici, 2004). See also Haupt, Schlesewsky, Roehm, Friederici, and Bornkessel-Schlesewsky (2008) for a review of ERP findings, as well as new evidence, suggesting that an N400 may often be elicited by object-initial disambiguation in embedded clauses.

negativity than that of SVO sentences. This sustained negativity is often associated with an increased working memory load.[2] For initially ambiguous German sentences, in contrast, they found a P600 when case marking on the second determiner disambiguated towards OVS vs. SVO. Based on these findings, we would thus expect to see a greater difference in hidden layer activation to the first noun phrase for unambiguous OVS vs. SVO sentences in the absence of scenes, and similarly a bigger change in hidden layer activation when the model processes the second noun phrase of unambiguous German OVS vs. SVO sentences. In contrast, for initially ambiguous German sentences in the absence of scenes, the only difference in SVO vs. OVS sentences should appear on the second noun phrase.

To evaluate these hypotheses, new simulations were conducted with CIANet in order to first ensure an overall bias towards subject-verb-object structures.[3] CIANet was trained on both SVO and OVS sentences, with and without a scene, for initially ambiguous as well as fully unambiguous German sentences. We presented the network optionally with a scene in which characters in one event corresponded to the utterance, and the characters in the other event were randomly selected subject to the constraint that the Noun1 character filled the role opposite to the one it lled in the relevant event. Half of the time the network was trained with a scene and half of the time without. A subject-verb-object (SVO) bias was introduced by training CIANet network on subject-verb-object sentences 75% of the time, and object-verb-subject (OVS) in the remaining 25%. As in the original model (Mayberry et al., in press), each scene, when present, had only one event relevant to the sentence being processed.

Results from contrasting incremental changes δ_w^c between hidden layer activations for SVO vs. OVS word orders across the four test conditions c (initially ambiguous/unambiguous with/out scene, identified as (a) AVS/O; (b) ANS/O; (c) UVS/O; and (d) UNS/O for ease of presentation) are shown in Figure 3. As predicted by the P600 found at the verb for initially ambiguous OVS sentences when the scene was present (Knoeferle, Habets, et al., 2008), we observed a significant Δ_{Adv} (immediately after the verb) for the OVS vs. SVO interpretation shown in Figure 3 (a). In the absence of scenes, the model also corroborated the predictions outlined above, thus supporting our linking hypothesis between ERP data and hidden layer activation distances: When linguistic case marking on a determiner disambiguated initially ambiguous sentences towards OVS in the absence of a scene (see Knoeferle, Habets, et al., 2008; Matzke et al., 2002) as shown in Figure 3 (b), analysis of the model revealed an appreciably larger Δ_{Det2} for the OVS. Moreover, the graph for the unambiguous sentences without a scene (d) showed a larger Δ_{NP1} and Δ_{NP2} that may reflect the two anterior negativities

[2] The pattern is somewhat different for unambiguous object-inital *mittelfeld* constructions, which sometimes elicit a transient negativity (Rösler et al., 1998; Bornkessel, Schlesewsky, & Friederici, 2002, 2003).

[3] The model presented by Mayberry et al. (in press) purposely avoided such a bias to eliminate potential confounds with other phenomena under investigation in that work.

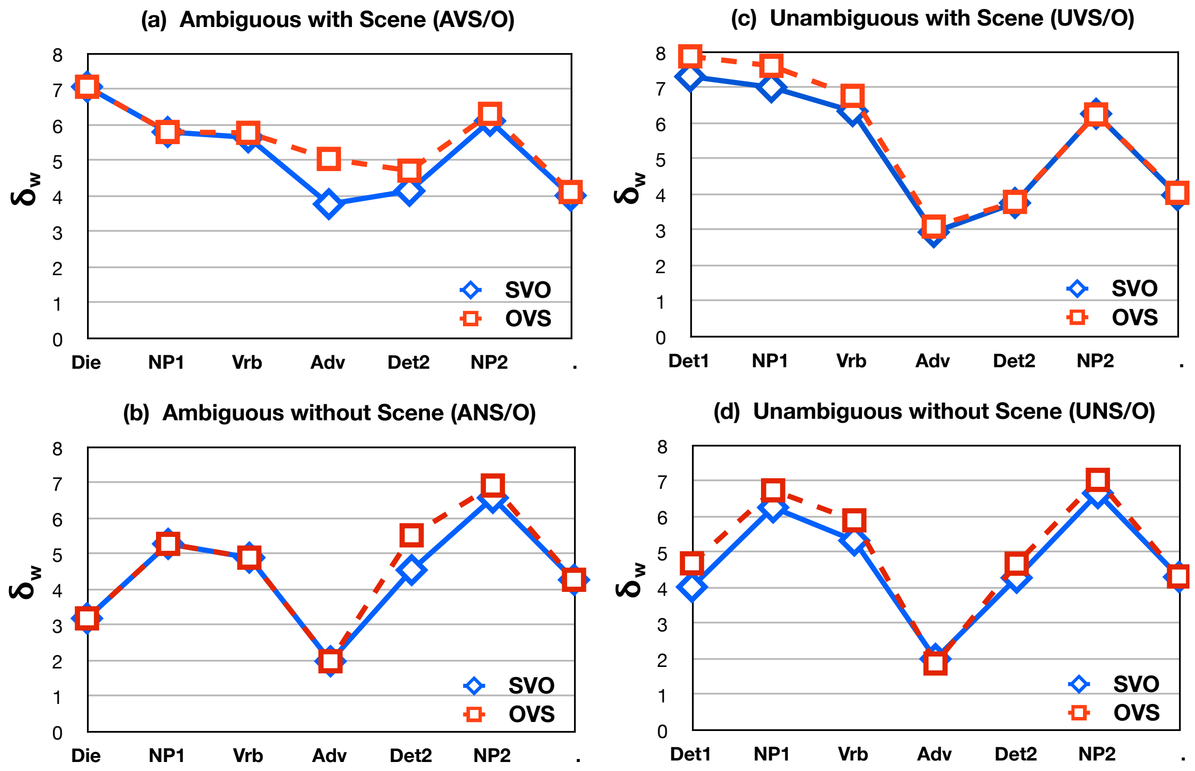

Figure 3: **Hidden layer activation differences.** Each graph plots the incremental change, δ_w^c, in hidden layer activations as each word w is processed for SVO and OVS sentences for each condition c shown in plots (a)-(d). The distance, Δ_w^c between OVS and SVO, correlates with ERP data from Knoeferle, Habets, et al. (2008) and Matzke et al. (2002).

that Matzke et al. (2002) found at the first and second noun phrase of unambiguous German OVS vs. SVO sentences. In contrast to the unambiguous sentences without a scene, the Δ_{NP2} difference between OVS vs. SVO was less pronounced when a scene was present (c) in line with findings by Knoeferle, Habets, et al. (2008).

Further analysis of the hidden layers for each of the conditions strongly suggested that each of the hidden layer activation changes (i.e., the changes indexing the P600 vs. LAN) was caused by a different subset of the hidden layer units: For the initially ambiguous sentences both with and without a scene, the greater hidden layer activation differences for the OVS versus SVO interpretation that indexed the experimentally observed P600 effect in the model was brought about by 40 of the 50 most active components of Δ_{Adv} (with scene) and Δ_{Det2} (without scene). Similarly, roughly 40 of the 50 most active units in the hidden layers of the unambiguous conditions with and without a scene accounted for the LAN effect indicated by the Δ_{NP1}, again showing a greater activation difference for the OVS sentences. It is important to highlight that the sets of units in the hidden layer that were responsible for the P600 vs. the LAN effects predicted by our linking hypothesis hardly overlapped: There was only an overlap of five units for the two phenomena, suggesting that CIANet has learned to functionally organize the hidden layer space in a way that distinguishes the P600 and LAN effects. In likening the hidden layer analyses of CIANet to the ERP data, this functional organization suggests qualitative differences at the level of the hidden layer for the activation changes that were argued to index the experimentally observed qualitatively different P600 versus LAN.

In extending the empirical coverage of the model, there are several directions which might be pursued. As noted earlier, the general pattern of ERP results for disambiguation towards object-initial structures is much more complex when one considers embedded and subordinate clauses (Haupt et al., 2008), as are the anterior negativities associated with unambiguous object-initial structures. To investigate these findings would entail substantially extending the linguistic coverage of the model, and ensuring that relevant distributional properties (e.g. frequency and predictability) of the language are reflected in the training regime. Further, by training the model on high- and low- predictability contexts, we would predict that CIANet should index an N400 when anticipated lexical material is not encountered (van Petten, Coulson, Rubin, Plante, & Parks, 1999; van Berkum et al., 2005; DeLong et al., 2005). Further, it would be interesting to extend the neural linking hypotheses to the combined early negativity/late positivity sequence elicited in response to scene-sentence relational mismatches (Wassenaar & Hagoort, 2007; Vissers, Kolk, van de Meerendonk, & Chwilla, 2008), and the N400 elicited by a verb-action mismatch (Knoeferle, Kutas, & Urbach, 2008). More systematically addressing the question of which unit subgroups in the hidden layer of the CIANet architecture correspond to which ERP components (including N400 eventually; Kutas & Hillyard, 1980, 1983; van Petten & Kutas, 1990) will be essential for enabling a precise linking hypothesis to such diverse ERP findings.

The extent to which the model can be extended to account for additional ERP findings, however, will likely be modulated by two factors. Firstly, it will be important to improve and extend the hidden layer analyses by examining the temporal interdependencies of hidden layer activation changes in response to (a) hearing a word, (b) increasing the salience of the relevant object in the scene, (c) anticipating the target role filler, and (d) the network's equivalent of language processes such as syntactic revision and semantic interpretation as reflected by changes in hidden layer activation. Secondly, it is again important to note that CIANet is not intended as a model of the precise mechanisms that people recruit for syntactic processing, *per se.* Rather, CIANet is intended to model more general properties of situated comprehension, such as its capacity for incremental, predictive, self-organizing, multi-modal, and experience-driven processing. Thus the proposed linking hypothesis does not assume that people are using an SRN to comprehend language, but rather that whatever language acquisition and processing architecture people possess organizes itself in a manner, such that sub-systems of information processing resources are devoted to the distinct dimensions of language comprehension indexed by observed ERP components. It is unlikely the model will organize itself isomorphically to people for all dimensions of all these dimensions of language processing. As such we see this kind of modeling as complementary to those theories of language processing which seek to relate ERPs with specific linguistic processes (e.g., Bornkessel & Schlesewsky, 2006; Friederici, 2002), only some of which may have counterparts that can be identified in a self-organizing model with limited coverage.

General Discussion

Based on a review of converging evidence for the real-time seamless reconciliation of visual and linguistic representations in situated language comprehension we have argued that the underlying processes are temporally interdependent. The account proposed here is consistent with that of other researchers who have argued for a tight coupling of the linguistic, motor, and visual processes and the representations they yield (Spivey & Richardson, in press; Zwaan & Madden, 2005), and can be viewed as a concrete computational proposal for the mechanisms of situated, or *embedded*, language processing. What is particularly novel about the connectionist implementation of the account is that it provides not only an account of on-line situated comprehension *behavior* (i.e., how visual attention increases the salience of relevant events), but also qualitatively reflects P600 and LAN components via changes in hidden layer activation, with distinct groups of units accounting for the two components. In what follows, we discuss two issues – the implications of online visual-linguistic interdependence for embodied language processing and the advantages of combined neurobehavioral linking hypotheses – in more detail. We conclude with a critical evaluation of the CIA in relation to embodied theories of language processing (Barsalou, 1999a; Glenberg & Kaschak, 2002; Zwaan, 2004).

The benefits of interdependence in visuo-linguistic processing are manyfold. Most obviously, it supports the rapid on-line reconciliation of linguistic and situational information, and the visually grounded enrichment of utterance meaning (see also Glenberg & Robertson, 1999). Crucially as we have observed above, the mutual constraints offered by such mechanisms serve to synchronize and "tune" the processes of both visual and linguistic inquiry: Information from the unfolding speech stream does not simply direct our eyes towards relevant objects and events of the world around us (Tanenhaus et al., 1995; Knoeferle et al., 2005), but appears to directly enhance visual search strategies (Spivey et al., 2001; Lupyan, 2007). Equally, the attended scene information can rapidly influence and even disambiguate the unfolding interpretation. Visual grounding instantiates lexical and thematic expectations (Weber et al., in press; Altmann & Kamide, 1999; Knoeferle et al., 2005), while also providing contextual information about recent and ongoing events (Knoeferle & Crocker, 2007) and referential domains (Tanenhaus et al., 1995) that have been shown to disambiguate sentence interpretation even before linguistic disambiguation occurs. Importantly, the rapid influence of scene event information on disambiguation and interpretation is not only revealed by observed patterns of utterance-mediated attention in the scene, but also confirmed by the ERP findings of Knoeferle, Habets, et al. (2008), using similar concurrent scene-sentence presentation and materials. Of further relevance was their observation that the P600 displayed a similar scalp distribution, regardless of whether disambiguation was triggered linguistically or by a verb-mediated depicted event, suggesting the involvement of similar functional brain mechanisms regardless of visual or linguistic modality.

While current theories of sentence processing exhibit many of the hallmark characteristics of incremental, experience-based, anticipatory, and integrative processing that the CIA assumes, the CIA highlights several challenges that traditional sentence processing theories face. Most obviously they typically lack any account of how such compositional comprehension mechanisms interact with visual processes. As we have argued, addressing this issue is not simply a matter of linking traditional linguistically-centered interpretation mechanisms with visual attention, but also of accounting for the use and relatively high salience of non-linguistic, visual context. Further, while many current theories emphasize the role of probabilistic, experience-based mechanisms, they again assume that relevant experience is linguistic in nature, such as lexical and structure frequencies. While behavioral evidence from visually situated comprehension suggests that linguistic expectations indeed play a pervasive role in comprehension (Altmann & Kamide, 1999), they are in fact overridden by relevant events in the visual context (Knoeferle & Crocker, 2006). This suggests that the probability of particular interpretations and probabilistic expectations are strongly modulated by objects, events and affordances in the immediate scene (or at least, that part of the scene being attended).

The CIANet architecture, while limited in its linguistic coverage, exemplifies how linguistic and non-linguistic processes may be temporally synchronized, with all available and relevant sources of information seamlessly integrated to determine the preferred interpretation. Crucially, the attentional mechanism amplifies relevant scene information, increasing its salience for interpretation. This behavior is further not stipulated, but rather emerges as a natural consequence of *situated* training (Mayberry et al., in press). Probably the most novel aspect of CIANet is that it offers a direct, if still qualitative, index of both behavioral (visual attention to scene events) and neuroscientific (event-related potentials) measures. Originally developed to model behavioral findings from visually situated language comprehension as outlined by the CIA, we have shown here that CIANet also offers a novel account of neuroscientific findings for processing of similar stimuli, via the proposed linking hypothesis that relates changes in activation for sub-groups of hidden layer units to specific ERP components. Specifically, one group of hidden units indexes the late positivity (P600) both for the case disambiguating second noun phrase when no scene was present, and for the earlier verb region when a co-present scene provided disambiguating event information. We assume that the P600 results from the reconciliation of propositional scene information with that of the preferred unfolding linguistic meaning, as outlined in the CIA, which entails a reanalysis in the ambiguous OVS condition. This on-line "reconciliation" of scene and utterance during the *scene integration* stage of the CIA is broadly consistent with the P600 predicted by both monitoring theory (Vissers et al., 2008) and the *generalized mapping* step of the eADM (Bornkessel & Schlesewsky, 2006). For unambiguous sentences, a second, and largely distinct, group of hidden units was shown to index the sustained LAN which is often associated with increased working memory costs (see Rösler et al., 1998; Matzke et al., 2002, and references therein).

We suggest that in the field of embodied and situated language processing, which has benefited greatly from insights of diverse measures and approaches – among them behavioral (response latencies, action execution, and eye tracking), neuro-imaging, and neuro-computational – the mapping of both eye-tracking and event-related brain potential measures onto the output and hidden layer activation respectively of a connectionist model harbors great promise for future research. More speculatively, a model with both visual attention and neural linking hypothesis has the potential to make predictions about the temporal interdependence of ERP and visual attention effects, potentially taking first steps towards modeling and examining phase-locking of visual attention and ERP components that reflect situated language comprehension. These extensions will likely require refining the attentional mechanism of CIANet to modulate not only event salience, but also that of individual scene constituents, and indeed to permit more flexible scene contexts with varying numbers of objects and events (see Mayberry et al., in press, for discussion).

The temporal coordination and interdependence of real-time linguistic and visual processing that we argue for sits comfortably with theories emphasizing the multimodal nature of representations, grounded in visual, sensory and motor experience. Firstly, such close synchronization of utterance and attention serves to support associationist mechanisms seeking to correlate language with visual and sensorimotor experience. Indeed, Knoeferle and Crocker (2006) conjecture that there may be a developmental basis for this synchronization of language and visual attention, arising from the important role that the immediate environment plays as a child learns to ground concepts with visual referents during language acquisition. There is by now general acceptance of the importance that visual grounding of language plays during language acquisition, with experimental findings suggesting that the influence of nonlinguistic information (e.g., objects and events) on language acquisition is highly dependent on the time-lock between a child's attention to, and child-directed speech about, these objects and events (Dunham, Dunham, & Curwin, 1993; Harris, Jones, Brookes, & Grant, 1986; Hennon, Chung, & Brown, 2000; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2007; Smith, 2000). The consistent coordination of comprehension and attention thus directly facilitates the association of verbal interpretations with visuomotor experience, and the representations those experiences yield, as posited in the embodied accounts of Barsalou (1999b) and perhaps most notably Zwaan and colleagues (Zwaan et al., 2002; Zwaan, 2004; Zwaan & Madden, 2005).

Interestingly, however, there are points where our own account and many embodied accounts appear to diverge. Zwaan et al. (2002), for example, accord considerable importance to the specific visual representations that the utterance evokes, based on long-term prior experience. While the CIA assumes that such experience is actively used during interpretation, it postulates that immediate visual context can readily override such experience, even when it is highly atypical. While these findings derive from differing experimental paradigms and phenomena, it will nonetheless be important to understand the balance that people strike when forced to make decisions in the face of such conflicting cues as the immediate visual context and their experience of the world.

At first inspection, the CIA appears to have limited direct bearing on the interaction of linguistic and motor processes (Glenberg & Kaschak, 2002). Let us assume for instance that, just as hearing *eat* triggers eye movements to a piece of cake on the table, so does it activate relevant areas of the motor system in the brain associated with eating (Buccino et al., 2005; Pulvermüller, Härle, & Hummel, 2001; Pulvermüller, 2005; Tettamanti et al., 2005). The CIA, in contrast, claims that utterance-mediated visual attention responds to several levels of the unfolding compositional linguistic interpretation, and has as its goal the rapid mapping and grounding of the utterance to relevant aspects of the scene, which in turn enables a rapid influence of relevant scene information. Activation of the motor system, in contrast, appears to reflect the grounding of particular verb meanings in the actions, and possibly also the simulation of more complex actions based on compositional sentence meaning. While such motor activation and action simulation may indeed partly underlie visual inspection of depicted events (Knoeferle et al., 2005) or anticipation of likely role fillers (Altmann & Kamide, 1999), it would be reckless to suggest the mechanisms are coextensive. Rather, we claim that mechanisms of utterance-mediated attention have adapted to guide visual interrogation of the environment to facilitate reconciliation of visual and linguistic information. What should be clear, however, is that such synchronization of spoken language comprehension with visual attention to relevant objects and events in the scene likely enhances the associationist mechanisms that ground meaning in the observed visual referent tokens and unfolding events. More speculatively, the CIA may act as a catalyst in synchronizing the activation of motor processes that are (a) evoked by the meaning of the unfolding utterance on the one hand, and (b) that occur in response to the time-locked visual inspection of ongoing events, on the other. While such temporally synchronous utterance-event pairs may be relatively rare in adult language experience, they may provide a powerful basis for bootstrapping the association of linguistic, visual and motor representations that emerge during language development.

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

Altmann, G. T. M. (2004). Language-mediated eye-movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*, *93*, B79–B87.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 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: Eye movements and the visual world* (pp. 347–386). New York: Psychology Press.

Arnold, J., Eisenband, J., Brown-Schmidt, S., & Trueswell, J. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*, B13–B26.

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

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

Barsalou, L. W. (1999b). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

beim Graben, P., Schlesewsky, M., Saddy, D., & Kurths, J. (2000). Symbolic dynamics of event-related brain potentials. *Physical Review E*, *62*, 5518-5541.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Bornkessel, I., McElree, B., Schlesewsky, M., & Friederici, A. (2004). Multi-dimensional contributions to garden path strength: Dissociating phrase structure from case marking. *Journal of Memory and Language*, *51*, 495-522.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*, 787–821.

Bornkessel, I., Schlesewsky, M., & Friederici, A. (2002). Grammar overrides frequency: Evidence from the online processing of flexible word order. *Cognition*, *85*, B21–B30.

Bornkessel, I., Schlesewsky, M., & Friederici, A. (2003). The neurophysiological basis of word order variations in german. *Brain and Language*, *86*, 116–128.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolattii, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Cognitive Brain Research*, *24*, 355–363.

Carpenter, P. A., Just, M., Keller, T., Eddy, W., & Thulborn, K. (1992). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *Neuroimage*, *10*, 216–224.

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

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, *10*(7), 335–344.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Claus, B., & Kelter, S. (2006). Comprehending narratives containing flashbacks: Evidence for temporally organized representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1031–1044.

Cooper, R. (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.

Crocker, M. W. (1996). *Computational psycholinguistics: An interdisciplinary approach to the study of language*. Dordrecht: Kluwer.

Crocker, M. W. (1999). Mechanisms for sentence processing. In S. Garrod & M. J. Pickering (Eds.), *Language processing* (pp. 191–232). London: Psychology Press.

Crocker, M. W. (2005). Rational models of comprehension: Addressing the performance paradox. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 363–380). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*(6), 647–669.

Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 361-367.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(89), 1117–1121.

Demarais, A., & Cohen, B. H. (1998). Evidence for image-scanning eye movements during transitive inference. *Biological Psychology*, *49*, 229–247.

Dunham, P. J., Dunham, F., & Curwin, A. (1993). Joint-attentional states and lexical acquisition at 18 months.

Developmental Psychology, *29*, 827–831.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Farmer, T. A., Anderson, S. E., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory and Language (Special Issue on Language-Vision Interaction)*, *57*(4), 570–595.

Federmeier, K. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology*, *61*(6), 825–850.

Fodor, J. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

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

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. Unpublished doctoral dissertation, University of Connecticut.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science*, *6*(2), 78–84.

Frisch, S., Schlesewsky, M., Saddy, D., & Alpermann, J. (2002). The P600 and an indicator of syntactic ambiguity. *Cognition*, *85*, B83-B92.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for ”common sense”: An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience*, *8*, 89–106.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1-76.

Glenberg, A., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, *9*(3), 558–565.

Glenberg, A., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, *28*, 1–26.

Hagoort, P. (2003). How the brain solves the binding problem for language: A neurocomputational model of syntactic processing. *Neuroimage*, *20*, S18–S29.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(1), 101–122.

Harris, M., Jones, D., Brookes, S., & Grant, J. (1986). Relations between the non-verbal context of maternal speech and rate of language development. *British Journal of Developmental Psychology*, *4*, 261–268.

Haupt, F., Schlesewsky, M., Roehm, D., Friederici, A., & Bornkessel-Schlesewsky, I. (2008). The status of subject-object reanalysis in the language comprehension architecture. *Journal of Memory and Language*, *59*, 54-96.

Hennon, E., Chung, H., & Brown, E. (2000). What does it take for 12-month-olds to learn a word? In *Breaking* *the language barrier: An emergentist coalition model of the origins of word learning (monographs of the society for research in child development)* (pp. 62–84). Oxford: Blackwell Publishing.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.

Johanssson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, *30*(6), 1053–1080.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*(1), 37–55.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, *5*, 196–214.

Knoeferle, P. (2007). Comparing the time-course of processing initially ambiguous and unambiguous German SVO/OVS sentences in depicted events. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 517–533). Oxford: Elsevier.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking. *Cognitive Science*, *30*(3), 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 (Special Issue on Language-Vision Interaction)*, *57*(4), 519–543.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*, 95–127.

Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual scenes trigger immediate syntactic reanalysis: evidence from ERPs during situated spoken comprehension. *Cerebral Cortex*, *18*(4), 789–795.

Knoeferle, P., Kutas, M., & Urbach, T. (2008). ERP correlates of verb-action and sentence-scene role relations incongruence in a picture-sentence verification task. In *Annual Meeting of the CNS.* Cognitive Neuroscience Society.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition*, *11*, 539–550.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

Kutas, M., van Petten, C., & Kluender, R. (2006). Handbook of psycholinguistics. In M. Traxler & M. Gernsbacher (Eds.), (pp. 659–724). Elsevier.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lupyan, G. (2007). Reuniting categories, language, and perception. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

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

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*, 226–228.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: Information processing time with and without saccades. *Perceptual Psychophysics*, *53*, 372–380.

Matzke, M., Mai, H., Nager, W., Rüsseler, J., & Münte, T. (2002). The costs of freedom: An ERP-study of non-canonical sentences. *Clinical Neuropsychology*, *113*, 844-852.

Mayberry, M. R. (2003). *Incremental nonmonotonic parsing through semantic self-organization.* Unpublished doctoral dissertation, Department of Computer Sciences, The University of Texas at Austin, Austin, TX.

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

Mayberry, M. R., & Miikkulainen, R. (submitted). *Incremental nonmonotonic parsing through semantic self-organization.*

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly brain potentials. *Psychophysioloy*, *30*, 170–182.

Otten, M., Nieuwland, M. S., & van Berkum, J. J. A. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(8), 1–9.

Pado, U., Crocker, M., & Keller, F. (in press). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*.

Pickering, M. J., & Traxler, M. (1998). Plausibility and the recovery of garden paths: An eye-tracking study.

Journal of Experimental Psychology: Learning, Memory and Cognition, *24*, 940–961.

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

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2007). The birth of words: Ten-month olds learn words through perceptual salience. *Psychological Science*, *18*, 414–420.

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

Pulvermüller, F., Härle, M., & Hummel, F. (2001). Walking or talking?: Behavioural and neurophysiological correlates of action verb processing. *Brain and Language*, *78*, 143–168.

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

Rösler, F., Pechmann, T., Streb, J., Röder, B., & Hennighausen, E. (1998). Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by event-rrelated brain potentials. *Journal of Memory and Language*, *38*, 150–176.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–148.

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford: Oxford University Press.

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

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

Spivey, M. J., Richardson, D., & Fitneva, S. (2004). Thinking outside the brain: Spatial indices to linguistic and visual information. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action* (pp. 161–189). New York: Psychology Press.

Spivey, M. J., & Richardson, D. C. (in press). Language embedded in the environment. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition.* Cambridge, UK: Cambridge University Press.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, *12*(4), 282–286.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of impied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*, 153–156.

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modelling discourse context effects: A multiple constraints approach. In M. Crocker, M. Pickering, & C. Clifton (Eds.), *Architectures and mechanisms for language processing* (pp. 90–118). Cambridge: Cambridge University Press.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., et al. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*, 273–281.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

van Berkum, J. J., 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, and Cognition*, *31*(3), 443–467.

van Petten, C., Coulson, S., Rubin, S., Plante, S., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 394–417.

van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, *18*, 380–393.

Vissers, C., Kolk, H., van de Meerendonk, N., & Chwilla, D. (2008). Monitoring in language perception: Evidence from ERPs in a picture sentence matching task. *Neuropsychologia*, *46*, 967-982.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*, 105–143.

Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence-picture matching electrified. *Neuropsychologia*, *45*, 716–740.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*, 367–392.

Weber, A., Crocker, M. W., & Knoeferle, P. (in press). Conflicting constraints in resource adaptive language comprehension. In M. W. Crocker & J. Siekmann (Eds.), *Resource adaptive cognitive processes.* Heidelberg: Springer Verlag.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities:

A study of anticipatory eye movements. *Cognition*, *99*, B63–B72.

Yaxley, R. H., & Zwaan, R. A. (2005). Simulating visibility during language comprehension. *Cognition*, *105*, B229–B236.

Zwaan, R. A. (1999). Embodied cognition, perceptual symbols, and situation models. *Discourse Processes*, *28*(1), 81–88.

Zwaan, R. A. (2004). The immersed experiencer: Towards an embodied theory of language comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 35–62). New York: Academic Press.

Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 224–245). Cambridge, UK: Cambridge University.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*, 1–11.