Characterising visual context effects: active, pervasive, but resource-limited

Pia Knoeferle

(Cognitive Interaction Technology Excellence Cluster Bielefeld University)

June 11, 2015

Correspondence regarding this article
should be addressed to:

Dr. Pia Knoeferle
Inspiration 1, CITEC
Bielefeld University
D-33615 Bielefeld
Germany
Email: knoeferl@cit-ec.uni-bielefeld.de
Fax: +49 521 106-6560

Introduction

Over the past two decades, researchers in the area of language and cognition have shown an increasing interest in examining language comprehension in relation to the visual context (henceforth 'visually situated' language comprehension). And within twenty to thirty years, the field has gone from postulating strict procedural modularity (e.g., Fodor, 1983; Frazier & Fodor, 1978; Friederici, 2002), according to which visual context information cannot affect incremental language comprehension, towards finding clear evidence to the contrary (e.g., Chambers, Tanenhaus, & Magnuson, 2004; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). By now, some approaches have even postulated a highly "active" influence of at least some kinds of information in the visual context (e.g., action events), depicting who-does-what-to-whom and thus imposing thematic role structure on language (Knoeferle, Crocker, Scheepers, & Pickering, 2005; Knoeferle, Habets, Crocker, & Münte, 2008). This active influence was also revealed through informational preferences: When an unfolding utterance was compatible with the possible mention of two different agents (of which one was associated with the sentential verb based on stereotypical role knowledge while the other was the agent of the action referenced by the verb), comprehenders preferred to rely on the action depiction and its agent (Knoeferle & Crocker, 2006, 2007; Knoeferle, Carminati, Abashidze, & Essig, 2011). Overall, numerous findings support the view that the immediate situation and our perceptual-motor experience play an important role in language comprehension and cognition (see for instance the contributions by Spivey and Huette, Brown-Schmidt, Chambers, and Farmer *et al*, this volume, see also the embodied approaches to language by Arbib, 2005; Barsalou, 1999; Pulvermüller, Härle, & Hummel, 2001).

But how precisely can we characterise the role of the visual context in language comprehension and is it really as important as we may want to believe? A first section in this chapter characterises its role by exemplarily reviewing visual context effects while identifying factors that may delimit them (section 1): For instance, comprehenders exploit not only objects but also action depictions in the immediate and the recent visual context within a few hundred milliseconds, but these effects are limited by the decay of representations in working memory (section 1.1). Moreover, visual context effects emerge time-locked to when words in the utterance identify relevant visual cues and they may be reduced or eliminated if that coordination is strained (section 1.2). Thus, while visual context effects emerge rapidly during comprehension, they - perhaps unsurprisingly - depend upon working memory and are sensitive to (strains on) the temporal coordination of visual perception with linguistic processes.

Should we conclude from this that the immediate visual context is of limited importance for comprehension? When we additionally consider that much of our everyday conversation is about past (vs. ongoing) events, that many of us spend more time writing and reading texts on a computer than engaging in communication about

the immediate environment, and that much of the language we process is about abstract content, then we might conclude that the role of the visual context in communication is very limited. In addition, it will become clear that our limited cognitive resources conspire with stimulus characteristics (e.g., their timing, complexity, and information density, e.g., speech rate and spacing of object mention), in imposing a limit on the mental representations that we can build and that can inform situated language comprehension. While I concede limitations on the effects of the visual context (e.g,. as a function of our cognitive capacities among others see Knoeferle and Crocker (2007)), the second section of this chapter also argues that the importance of visual context effects[1] is evident in their pervasiveness across

(i) reading and spoken comprehension

(ii) different types of scenes (cliparts, photographs, and real-world)

(iii) different aspects of the visual context (a speaker's eye-gaze, mimics, and gestures); and

(iv) both concrete and abstract language processing.

In light of their broad coverage ((i)-(iv)), the existing findings provide a solid basis for developing a relatively comprehensive theory of situated language comprehension and for beginning to specify in more detail the mechanisms of how utterance comprehension interacts with (visual) attention and visual context effects. While a first step towards this goal has been undertaken by the existing accounts of situated language processing (e.g., Altmann & Kamide, 2009; Knoeferle & Crocker, 2006, 2007), these are currently underspecified (see Crocker, Knoeferle, & Mayberry, 2010; Knoeferle, Urbach, & Kutas, 2014, for a relevant computational model and a more precise specification of comprehension sub-processes respectively). One challenge in further specifying model predictions, as I will argue in a third section, is the development of more detailed linking hypotheses[2] between comprehension processes and one of the key measures used to examine situated comprehension (visual attention to objects across time).

1 Characterizing visual context effects in situated language comprehension

This section reviews findings on visual context effects during comprehension, indicates potential delimitations of these effects, and characterises the role of the visual context in language comprehension. For examining visually situated language comprehension, scientists have largely relied on the so-called 'visual-world' paradigm in which we monitor participants' object-directed gaze during spoken comprehension to gain insight into their comprehension processes (see Spivey and Huette, this volume). The present chapter will draw on results from the visual-world paradigm but will also review results from the monitoring of brain activity time-locked to the

presentation of visual and linguistic stimuli (event-related brain potentials, ERPs, e.g., Ganis, Kutas, & Sereno, 1996, for early evidence). ERPs complement insights from eye-tracking studies since they can index qualitatively distinct processes. For instance, variation regarding semantic interpretation in strictly linguistic contexts typically manifests itself as a modulation of the so-called 'N400'. The N400 is a negative peak in the ERP signal, approximately 400 ms after the onset of a stimulus; the larger its amplitude the greater the difficulty of integrating a word in the linguistic context (Kutas & Hillyard, 1980, 1984). By contrast, structural disambiguation in ERPs is typically indexed by an increase in mean amplitude P600s (a positive deflection in the average electrical activity approximately 600 ms after the onset of a disambiguating stimulus, Osterhout & Holcomb, 1992; Hagoort, Brown, & Groothusen, 1993)[3]. Studies on visually situated language comprehension have exploited ERPs precisely examine the nature of the implicated comprehension processes, thus complementing insights into the content of the interpretation, indexed by where comprehenders look.

1.1 Beyond objects: Effects of (recent) action representations and the role of working memory

The first studies to reveal rapid effects of a visual referential context on language comprehension were conducted by Michael Tanenhaus and his group in Rochester (Tanenhaus et al., 1995)[4]. In their study, participants inspected real-world objects, and followed instructions to perform simple actions such as to *Put the apple on the towel in the box*. When hearing *on the towel*, participants could either attach that noun phrase as a modifier to *the apple*, specifying its location, or insert it into the verb phrase, specifying the destination of the putting action. Comprehenders typically prefer the destination interpretation, but when they noticed two apples in the visual context only one of which was on a towel, they abandoned their preferred structure and interpretation, and instead used *on the towel* to identify the correct apple. This became clear because they mostly inspected the apple on the towel and because they (unlike when only one apple was present) did not inspect another (empty) towel which could serve as a destination for the putting action.

One could argue that these findings were unsurprising given that there were only few immobile objects and that comprehenders had ample time to inspect these as the experimenter placed them on a table. Indeed, comprehenders could in principle have differentiated the two apples by virtue of their features (on a towel vs. not on a towel) even before hearing any language and settled on how to interpret the context. Interpreting the visual context and relating it swiftly to the utterance is arguably not that challenging under these circumstances. What if comprehenders faced a similar linguistic ambiguity, but instead of objects, the context depicted two agent-action-patient events and upon hearing the verb people had to relate it on the fly to one of these two events? Would they still be able to rapidly exploit the visual context for utterance interpretation?

In the first study to examine the effects of depicted action events on real-time language comprehension, peo-

ple listened to German NP-V-NP sentences (e.g., *Die Prinzessin malt offensichtlich der Fechter*, 'The princess (amb.) paints apparently the fencer (object / agent)' Knoeferle et al., 2005). The initial noun phrase in these sentences was ambiguous (it could be either the subject / agent or the object / patient) but is often interpreted as the subject of the sentence; the local structural ambiguity was resolved at the case-marked post-verbal noun phrase. Crucially, listeners could rely on one of two depicted events for earlier disambiguation. But which of these depicted events (one showing the princess as being painted by a fencer, the other showing her as washing a pirate) was relevant for comprehension only became clear as they heard the verb. The verb identified either the washing or the painting action, and accordingly established the princess as event agent (subject) or patient (object). Participants rapidly related the verb to the matching action and its associated agent. For instance, they inspected the fencer more often as soon as the verb in OVS sentences had identified him as the agent through verb-action reference (*malt*, 'paints' - fencer-painting). This gaze pattern suggested that comprehenders had assigned an agent role to the fencer, and a patient role to the initially ambiguous noun phrase and its referent, the princess, thus informing structural disambiguation on the fly during auditory sentence comprehension.

However, strictly speaking, one could criticise the eye-tracking study for its interpretation of the gaze record as reflecting structural disambiguation. Can we be certain that comprehenders' gazes to objects index thematic role assignment and structural disambiguation? In a corresponding auditory ERP study, participants inspected similar scenes and listened to similar sentences while ERPs were being recorded. Recall that the so-called P600 indexes syntactic disambiguation (see above). In the face of structural ambiguity, German listeners will interpret a sentence-initial noun phrase as the subject; if, however, verb-action reference clarifies that the first-named referent is the patient and object the sentence, initiating a revision from a subject-first to an object-first structure, then we should see increased mean amplitude P600s time-locked to the verb. And indeed, participants' mean amplitude P600s time-locked to the verb and the post-verbal adverb increased when the verb identified a depicted event that disambiguated towards the disfavored object (vs. subject)-initial structure (i.e., when the verb referenced the event portraying the princess as the patient versus as the agent, see Crocker et al., 2010, for relevant neuro-behavioral modeling research).

Clearly then, these depicted events affected spoken comprehension rapidly once they had been mediated by language (the verb). One concern about these results, however, is that the co-presence of the scene may have implicitly heightened its relevance for comprehension, arguably enhancing its effects (but note that the relevance of the scene varied within the experiments since filler scenes sometimes had no relation whatsoever to the accompanying utterance, Knoeferle et al. (2005)). However, if visual context effects emerge even when scenes are not immediately present, this would speak to their importance for language comprehension more generally. And indeed, objects (Altmann, 2004; Spivey & Geng, 2001) need not be immediately present to rapidly affect visual attention and language comprehension. In one study, people inspected a clipart picture

showing a man, a woman, a newspaper, and cake, and then the screen went blank. After this, people heard, for instance, *The man will eat....* At this point, listeners inspected the location where they had previously seen the cake, a behavior that was interpreted as suggesting that even a mental record of the visual context can influence incremental semantic interpretation (Altmann, 2004).

Information about *actions* and their agents can also influence visual attention and language comprehension when the actions are not immediately present (Knoeferle & Crocker, 2007). People listened to object-initial German sentences describing one of two events (e.g., a pilot depicted as being offered food by a detective and as being spied-upon by a wizard). As in previous studies (Knoeferle et al., 2005), the verb (e.g., *verköstigt*, 'serves-food-to') referred to one of two depicted actions and increased the relevance of its associated agent (e.g., *Den Piloten verköstigt gleich der Detektiv*, 'The pilot (acc. obj) gives-food-to soon the detective (subj)' mediated the food-serving action and its agent). Unlike previous studies, however, the scene was removed prior to utterance presentation. Gaze patterns in the blank screen during comprehension showed that even when the scene had been removed, comprehenders rapidly relied on a recently-inspected event precisely at the point in time when the verb identified that event as relevant. This finding generalized to a quasi-dynamic action presentation, whereby the two depicted actions were presented one at a time in sequence, and then both removed prior to utterance presentation and only the characters remained on-screen (Knoeferle & Crocker, 2007, Exp 2).

The same experiments revealed further an informational preference: When the utterance contained a different verb (*bespitzelt*, 'spies-on') and was compatible with the possible mention of two different agents (a detective was associated through stereotypical role knowledge while a wizard was depicted as performing a spying action), comprehenders preferred to rely on the action depiction and its agent (Knoeferle & Crocker, 2006). This preference generalised when the scene was presented prior to the utterance (Knoeferle & Crocker, 2007, Experiment 1) but was eliminated when the two depicted actions were presented one at a time in sequence, and then both removed prior to utterance presentation and only the characters remained on-screen (Knoeferle & Crocker, 2007, Experiment 2). Plausibly, the recent action had experienced some decay whereas the competing stereotypical agent received support through its continued on-screen presence.

Recent actions are sometimes also preferred over expectations of (uncertain) future events. In Knoeferle and Crocker (2007, Exp 3.) people inspected a first clipart event depiction in three frames: A character was depicted as moving towards an object, interacting with it (e.g., polishing candelabra), and moving away from it. An ensuing utterance was ambiguous between referring to that recent action (and its target) and referring to an equally plausible future action and its different target (e.g., polishing crystal glasses). As they heard the verb (*poliert*, 'polish'), comprehenders preferentially inspected the target of the recent (vs. future) action. Even ensuing disambiguation through temporal adverbs (referencing the past vs. future), did not eliminate

this inspection preference. This finding suggest an informational preference, viz. that all else being equal, comprehenders prefer to rely on the immediate visual context over their expectations of (uncertain) future events. In section 2 will discuss whether this finding replicates in a real-world setting.

The picture that emerges is one in which comprehension can flexibly and rapidly exploit information from objects and depicted actions in the visual context. Visual context effects further emerged both when relevant objects and events were immediately present and when they were part of a comprehender's recent visual experience, suggesting these effects are somewhat independent of the co-presence of a visual context. However, for the (recent) visual context to affect comprehension, it must be mediated by representations in working memory. Since our cognitive resources are limited, recent scene representations will decay if they receive no further support (e.g., through visual inspection). To the extent that they experience decay, these representations will have a reduced effect on utterance comprehension and visual attention (see Knoeferle & Crocker, 2007, for relevant discussion). Our limited working memory capacity is thus one factor that likely limits visual context effects (see also section 2).

1.2 Temporal coordination: language and visual context effects

Working memory limitations as a bounding factor may also underlie another characteristic of situated language comprehension, viz., the close temporal coordination between utterance comprehension, visual attention and visual context effects. One example of temporally coordinated processing is that comprehenders tend to inspect an object shortly after it has been mentioned. This temporally coordinated, utterance-mediated inspection arguably means that working memory load is minimised (i.e., the relevant representations need not be retained in working memory for a long time). If temporal coordination is a fundamental characteristic of language comprehension (arguably because it reduces cognitive load), then it should be pervasive and robust. And indeed, it seems to be pervasive - in relating scalar adjectives to object size (Sedivy, Tanenhaus, Chambers, & Carlson, 1999), when interpreting prepositions in relation to action and object affordances (Chambers, Tanenhaus, Filip, & Carlson, 2002), and when relating verbs to either object affordances (Chambers et al., 2004), or depicted action events (Knoeferle et al., 2005; Knoeferle & Crocker, 2006). In all of these experiments, visual attention began to shift towards relevant aspects of the visual context from approximately 200 ms after these aspects had been identified as relevant by the utterance[5].

Comprehenders maintain this temporal coordination even when a speaker talks fast and rapidly mentions objects. Andersson, Ferreira, and Henderson (2011) manipulated speech rate (slow vs. fast) and object mention (in rapid succession versus spaced out). These manipulations affected how rapidly participants shifted their gaze to referents as they were mentioned, but they did not entirely eliminate the temporal coordination

between understanding a word and inspecting its referent. When the speech rate was fast and four objects were mentioned in rapid succession, participants attended to the relevant referents with some delay (it took them longer to shift attention to the referent) and less frequently compared with the slower speech rate. Even in the highest load situation, however, (with fast speech and rapid succession of mentioned objects), participants' eye gaze showed they still attempted to rapidly relate nouns to relevant referents. To the extent that comprehenders attempt to preserve those aspects of attentional behavior which benefit their language processing, the observed robustness of the temporal coordination speaks to its importance for language comprehension.

Moreover, if temporal coordination is essential in eliciting visual context effects, then these should be reduced or eliminated when two cues appear asynchronously. This is precisely what has been found in a recent ERP experiment in which comprehenders failed to semantically integrate an iconic gesture with its corresponding linguistic expression when these two cues were not presented in close temporal coordination (Habets, Kita, Shao, Özyürek, & Hagoort, 2011). Based on the observation that an iconic gesture often precedes (and overlaps with) its corresponding linguistic expression(s), Habets et al. (2011) manipulated the onset of the gesture relative to speech (speech was either delayed by 160 or by 360 ms or presented at the same time as the gesture). Participants saw videos of a person gesturing and making a statement that was either semantically congruent or incongruent with the gesture. In both the simultaneous and the 160 ms delay conditions (but not when the delay was 360 ms), mean amplitude N400s in the ERPs time locked to speech onset increased for mismatches compared with matches, indicating speech-gesture integration. These results thus revealed that speech and gesture are integrated most efficiently when their onsets are closely temporally coordinated[6].

If individual aspects of the visual context (e.g., an action) are recruited temporally coordinated with utterance comprehension, then their effects should further emerge time-locked to when they are identified as relevant by the utterance. Knoeferle (2007) compared the processing of structurally unambiguous spoken German sentences (in which case marking identified the first noun phrase as either the subject or the object of a sentence) with locally structurally ambiguous ones (in which the initial noun phrase was case and role ambiguous). For initially structurally ambiguous spoken sentences, depicted events should permit disambiguation shortly after hearing the verb (as had been shown by Knoeferle et al., 2005), and indeed, this result replicated. By contrast, for the unambiguous sentences, case marking on the first noun phrase together with the action depiction could in principle clarify the role relations prior to the verb, and thus elicit earlier thematic role assignment. The gaze pattern during the verb corroborated this expectation, suggesting thematic role assignment occurs as soon as the utterance identifies relevant role relations (see also Zhang & Knoeferle, 2012).

The temporal coordination also characterizes how a listener follows a speaker's gaze. Speakers on average tend to inspect an object approximately 800-1000 ms before they mention it (e.g., Griffin & Bock, 2000), an attentional pattern which could be exploited by a listener. In a first experiment by Richardson and Dale

(2005), a comprehender watched pre-recorded videos in which a speaker talked about characters in a television series (e.g., "Friends"), portrayed on-screen. The comprehender was most likely to inspect a character approximately two seconds after the speaker, and this held both when the speaker named the character and when he had talked about her without naming her (naming shortened the gaze lag by 370 ms). In a further experiment, comprehenders saw the same pictures and these flashed briefly either at the time when the speaker inspected them or in a shuffled order (Experiment 2, Richardson & Dale, 2005). Post-experiment, the comprehenders each responded to eight comprehension questions. The coordination of speech and gaze affected their response latencies with reliably faster responses (by 525 ms) in the synchronised than shuffled gaze condition. A listener's and speaker's eye gaze appear even more closely synchronized in dialogue than in monologue: When Richardson and Matlock (2007) compared gaze patterns in monologue and dialogue, addressees inspected the pictures of mentioned characters approximately two seconds after the speaker in monologues, whereas this lag decreased to zero milliseconds in real-time dialogue.

Summary Together these findings illustrate that the visual context actively imposes thematic role structure onto language, that visual context effects are rapid, and that they emerge for both co-present and recent objects and depicted events. Visual context effects on comprehension are further limited by decay of representations in a comprehender's working memory and characterised by a close temporal coordination with comprehension and (visual) attention. Accordingly they are not invariantly rapid but are sensitive to strains on the temporal coordination of visual and linguistic processing. Overall, the reviewed results are interesting from both a theoretical and a methodological vantage point. From a theoretical viewpoint, they were hailed as clear evidence against strictly modular approaches to cognition and language. Methodologically, they heralded a new era of language studies that used comprehenders' visual attention to objects as a window into the real-time integration of visual and linguistic cues during spoken language comprehension.

The visual-world paradigm has meanwhile branched out: Scientists have also examined how visual context representations modulate reading (e.g., Knoeferle & Crocker, 2005; Knoeferle, Urbach, & Kutas, 2011; Knoeferle et al., 2014); they have examined comprehension across a range of different context types among them real-world (e.g., Tanenhaus et al., 1995), clipart (e.g,. Altmann & Kamide, 1999), real-world photographic (Andersson et al., 2011) and video (e.g., Abashidze, Carminati, & Knoeferle, 2014) contexts. They have assessed the effects of different aspect of the visual context (e.g., of objects, actions, gaze, and a speaker's emotional facial expressions, Carminati & Knoeferle, 2013; Knoeferle & Kreysa, 2012; Kreysa, Knoeferle, & Nunnemann, 2014) and they have examined abstract in addition to concrete language processing (e.g., Guerra & Knoeferle, 2014). These extensions permit us to assess the pervasiveness of visual context effects in real-time language processing (of which more in section 2).

2 The pervasiveness of visual context effects

While visual context appears to play an active role in comprehension as argued in the preceding section, a sceptic might argue that these effects, and the temporally coordinated interplay between visual attention and language comprehension are limited in their generalizability. During spoken comprehension, for instance, our gaze is free to interrogate the scene and rapidly relating objects to language may be relatively straightforward. In reading, by contrast, our visual apparatus is engaged in inspecting words, likely precluding at least the kinds of overt gaze shifts to where an object had been that comprehenders performed during spoken language comprehension. If these overt shifts mediate the effects of the recent visual context, we may not see clear visual context effects in reading. Would (recent) visual context effects also emerge in reading, across different types of visual contexts, and for abstract language among others?

Overall, these effects must generalize to be representative of incremental language comprehension more broadly defined, which includes integrating pictorial information during reading and dealing with all sorts of incongruous language-world relationships. We want to assess visual context effects when language is about dynamic events and cluttered scenes; when comprehenders could integrate an interlocutor's eye gaze, gestures, and mimics with a visual context and language; and when language is about abstract ideas. The review in the preceding section has already hinted at potentially pervasive visual context effects (e.g., of gestures and speaker gaze) during language processing. Section 2 assesses the pervasiveness of visual context effects more systematically (in reading and picture-sentence verification; for different types of visual contexts; for different aspects of a visual context; and for the interpretation of abstract language, see i-iv in section 1).

2.1 Visual context effects during reading: picture-sentence verification

To the extent that visual context effects are pervasive across language modality, we should see them also in reading, closely time-locked with comprehension. One tradition that has examined visual context effects on reading, is picture-sentence verification. In this task, participants verify whether a picture matches or mismatches a sentence ("true" or "false") and visual context effects are indexed by longer response times for picture-sentence mismatches than matches. Such congruence effects were, however, not reliably present for serial picture-sentence presentation (e.g., Goolkasian, 1996; Underwood, Jebbett, & Roberts, 2004). Perhaps then the effects of a recent visual context are less robust in reading than in spoken comprehension? However, perhaps the failure to observe effects of recent visual context on reading is an artefact of the measure used in verification tasks: Post-comprehension response latencies may fail to capture the sort of incremental context effects that we have seen in the visual-world studies. But if incremental effects exist in reading, then they should emerge in continuous eye movement and ERP measures.

Support for incremental visual context effects during reading comes from a recent picture-sentence verification study (Knoeferle et al., 2011). Participants read a subject-verb-object sentence and verified at sentence end whether or not the verb matched a previously inspected action event depiction. ERPs recorded during reading provided insight into potential incremental verb-action congruence effects. Recall that difficulty in integrating a word with its linguistic context elicits an increase in mean amplitude N400s (Kutas & Hillyard, 1980, 1984). If participants rapidly integrate the verb with the preceding action, then difficulty in semantic integration for the mismatches should emerge in increased mean amplitude N400s at the verb. And indeed, participants' verb N400s over centro-parietal scalp were larger and ERPs to the object noun more negative for verb-action mismatches than matches (Knoeferle et al., 2011). In addition, the study replicated the congruence effect in the RTs which had sometimes (but not always) been reported in prior research (e.g., Goolkasian, 1996).

These results highlight the importance of the recent visual context also for reading. In addition, they revealed a modulation of visual context effects through comprehenders' verbal working memory capacity, which we had identified as one bounding factor in section 1. Participants with higher verbal (but not visual) working memory capacity showed earlier verb-action congruence N400 effects. Thus, inter-individual differences in working memory may modulate visual context effects (for instance, high-working memory individuals may retrieve visual-context representations more rapidly, yielding earlier context effects). Overall, individual comprehenders differ in their working memory and attentional capacity such that the representations they glean (and retain) from a visual context will differ in level of detail, in how long they remain active, and in how quickly they are accessed from working memory (see also Carminati & Knoeferle, 2013; Knoeferle, in pressb; Nation & Altmann, 2003; Huettig, Rommers, & Meyer, 2011, for research on further individual differences in situated language comprehension).

If visual cues are recruited temporally coordinated also in reading, then their effects should further emerge time-locked to when they are identified as relevant by the sentence. In a study by Knoeferle et al. (2014) participants read a subject-verb-object sentence (rapid serial visual presentation), and verified whether or not it matched different aspects of a recently viewed clipart depiction (the picture fully matched the sentence, or mismatched in either the action, depicted role relations, or both). Verb-action mismatch effects should emerge at the verb, as could role relations mismatch effects (see Wassenaar & Hagoort, 2007); but the latter could, in principle, occur even earlier, if people rapidly relate the pre-verbal sentence subject to the depiction of a character as an agent or patient. In the ERP data, verb-action congruence effects appeared immediately at the verb: N400s over centro-parietal scalp to the verb (300- 500 ms) were larger for verb-action mismatches relative to matches. ERP effects to the role-relation mismatches differed qualitatively from and occurred prior to the verb-action congruence N400 (during the subject noun), and this finding generalised across different word presentation rates (500 and 300 ms stimulus onset asynchrony). Congruence effects in the response

times emerged only for verb-action but not role relations mismatches (vs. matches), and only when words were presented relatively slowly (with 500 ms but not with 300 ms stimulus onset asynchrony). Thus, relevant aspects of the visual context began to inform reading closely temporally coordinated with when they were identified as relevant, echoing the findings from spoken language comprehension while response latencies did not consistently mirror these incremental effects (see Knoeferle, 2007).

Clearly then the failure to robustly observe congruence effects in reading seems to have been an artefact of the (response time) measure. A further explanation for the variability of visual context effects in picture-sentence verification response latencies is that encountering frequent incongruence between language and the world may have discouraged comprehenders from integrating these two information sources. By contrast, in most visual-world studies, referential success was above chance, and it's possible that visual context affected comprehension in real time because it could be successfully related to language. Incongruence can appear as outright mismatches, or as nuances in how different individuals describe the same object or event (e.g., one person sees and thinks *couch* while another refers to it as *sofa*). How people talk about their world may also depend on their age, gender, and social status. As a result of this variation, another language user's utterances and written text may not always match a comprehenders own representation of the non-linguistic visual context. But even when mismatches were frequent, comprehenders seemed to attempt rapid (rather than delayed) reconciliation of linguistic and visual context information, as has become clear from the studies discussed above (see also Vissers, Kolk, van de Meerendonk, & Chwilla, 2008; Wassenaar & Hagoort, 2007). In summary, visual context effects appear robust in reading even when language and pictures are frequently mismatched, but they are sensitive to a comprehenders' verbal working memory capacity.

2.2 Different types of visual contexts, complexity and preview time

The active visual context effects, and more generally research on situated comprehension, have been criticized for the prevalent use of impoverished visual contexts, and the associated risk that findings won't generalize to more complex scenes (Henderson & Ferreira, 2004). To which extent existing findings on situated sentence comprehension generalize to different visual contexts is still unclear, although first insights are beginning to emerge. Recall, for example, the study by Andersson et al. (2011) which manipulated information load such that a speaker mentioned several objects in photographs of cluttered real-world scenes temporally spaced out or in rapid succession and with slow or fast speech. While listeners were slower to shift their gaze to relevant referents when speech was fast and objects mentioned in rapid succession, they still attempted to inspect them, suggesting that the closely temporally coordinated interplay of language understanding and visual attention generalizes. That interplay was also apparent in comprehenders' close shadowing of an interlocutor's gaze

shifts in dialogue interaction (Richardson & Matlock, 2007). In fact, a close shadowing of dynamic visual cues exists from early infancy (six months of age Richardson & Kirkham, 2004).

Given these findings it is plausible that our comprehension system can also rapidly exploit dynamic real-world events. Abashidze, Knoeferle, Carminati, and Essig (2011) used the design from Experiment 3 by Knoeferle and Crocker (2007, see section 1) but replaced the clipart depictions with real world events in which an experimenter faced the participant and performed actions on objects (e.g., strawberries and pancakes) located on a table in front of him. In an example trial, the experimenter sugared pancakes (see Fig. 1). When that action had been completed, the experimenter gazed straight ahead and a sentence was played about either the recent action (e.g., *Der Versuchleiter zuckerte soeben die Pfannkuchen*, literally: 'The experimenter sugared recently the pancakes'), or about a potential future action on the other available object (e.g., *Der Versuchleiter zuckert demnächst die Erdbeeren*, literally: 'The experimenter sugars soon the strawberries'). At issue was whether comprehenders would - just as in the clipart version of the experiment - exhibit a preferred inspection of the recent (vs. future) action target during comprehension of the verb. Gaze pattern during and after the verb confirmed that this was the case. The time course of this gaze pattern for real-world actions was approximately the same as for the clipart studies reported by Knoeferle and Crocker (2007). This suggests that – at least for these kinds of actions – both clipart and real-world versions can affect spoken language comprehension in real time with a highly similar time course.

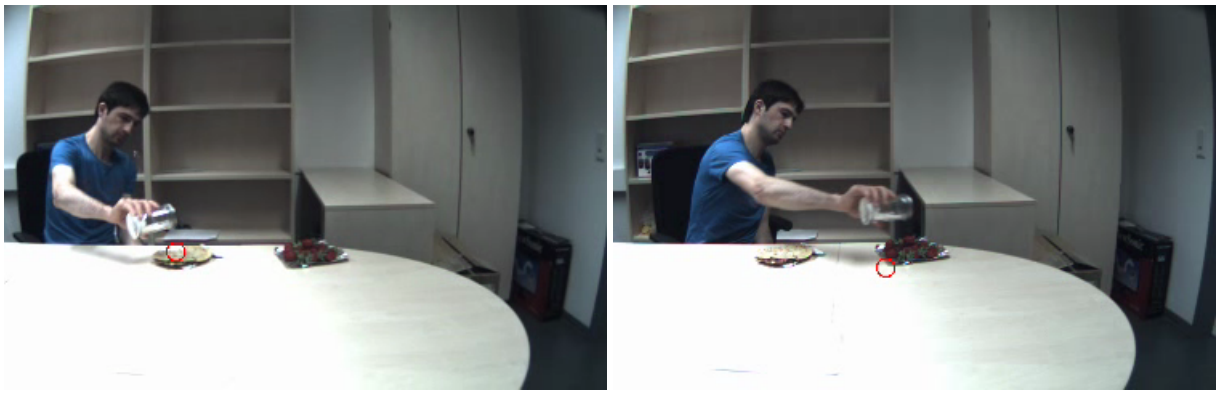

Figure 1: Participants saw the experimenter sugar the pancakes. Then they heard either (a) *Der Versuchsleiter zuckerte kürzlich die Pfannkuchen* 'The experimenter sugared recently the pancakes' or (b) *Der Versuchsleiter zuckert demnächst die Erdbeeren. 'The experimenter sugars soon the strawberries*. After sentence presentation, they saw the experimenter sugar the pancakes.

The results from the study by Abashidze et al. (2011) further corroborate the view (obtained from the experiments with clipart materials) that visual context plays a highly active role in language comprehension.

Ensuing studies put this claim to a more stringent test. In the clipart studies and in Experiment 1 by Abashidze et al. (2011), people had always seen one action per trial, prior to sentence comprehension This within-experiment frequency bias towards recent events may have caused the preferred reliance on recent (vs. future) events. The recent-event preference persisted, however, when participants saw equally many recent real-world actions (i.e., performed before sentence presentation and referenced in the past tense) as future actions (i.e., described as occurring soon and performed immediately after sentence presentation, Experiment 2, Knoeferle et al., 2011).

It also persisted when the stimuli were videos, presented on a computer display, and when the recent actions were pitted against a strong short-term (within-experiment) frequency manipulation: Abashidze et al. (2014) introduced a frequency bias in favor of the future over the recent event (Experiment 1: 88% future vs. 12% past events in combination with future and past sentences; Experiment 2: 75% future vs. 25% past events). Analyses of the gaze data from their experiments revealed that increasing the frequency of the future event did result in earlier fixations to the target of the future event than previously observed (in Experiment 2 of Knoeferle et al., 2011). However, they replicated the same *overall* preference to look at the target of the recent event from the verb and throughout sentence presentation.

In the experiments reviewed in this subsection, the real-world actions were dynamic but otherwise the visual context contained only two objects and an actor. While this sort of variation in the stimuli did not noticeably modulate the time course of visual context effects on language comprehension, variation in the preview time and complexity of a visual context did. Ferreira, Foucart, and Engelhardt (2013) examined visual context effects on the resolution of a local structural ambiguity. Participants listened to sentences such as *Put the book on the chair in the bucket* in which *on the chair* can either serve as a modifier of *the book* or as a destination of the putting action. Tanenhaus et al. (1995) had reported clear effects of a referential visual context on the resolution of this ambiguity (when two books were present, participants should prefer the modifier attachment; when one book was present, they should interpret *on the chair* temporarily as the destination). Ferreira et al. (2013) replicated this effect with displays of four objects and a three-second preview time for the visual context. By contrast, when no preview time was given (Experiment 2), or when twelve objects were displayed, these visual context effects were eliminated.

The results by Ferreira et al. (2013) can be interpreted in terms of resource limitations (whereby reduced visual context effects result from incomplete scene representations if time is scarce or the scene complex, given limited resources). Effects of decay in working memory are also apparent in other studies: Chambers and San Juan (2008) examined comprehenders' visual attention in response to a sequence of instructions. In their first experiment, participants saw four objects which were placed on a grid of nine numbered fields. They were instructed to move one object (e.g., a chair) to area 2 in a first instruction. Then they either immediately received the target instruction (e.g., *Now return / move the chair to area five*, its initial location) or they were

first instructed to manipulate another object, thus receiving the target instruction with some delay. In this setting, participants made more saccades to the target object (the chair) when they had heard *return* than *move* in the target instruction, and more for *return* (but not *move*) sentences when the target instruction immediately followed the first instruction than when it was delayed. These results suggest that discourse constraints (the 'returnable' status of an object) can affect visual attention (the target object had previously been moved and thus fulfilled the returnable constraint, eliciting more inspection for *return* than *move*); they crucially also corroborate that once an object is deemed relevant (the chair for a returning action), then decay of visual context representations in working memory modulated its inspection (the also Chambers, this volume for relevant discussion).

In summary, the rapid visual context effects seem on the one hand relatively robust and generalize from clipart to photographic scenes, to real-world action events and to videos of real-world events. On the other hand, the amount of time that comprehenders were given for previewing the visual context, the time that had passed after seeing a relevant action, as well as the complexity of the visual context, all modulated visual context effects. These finding re-emphasize the importance of comprehenders' resource limitations (e.g., in attention and working memory) for modelling visual context effects.

2.3 Effects of speaker-based visual information

A further test case for the coordinated interplay and visual context effects is how comprehenders deal with visual complexity of a different sort (e.g., when not only language and a referential context but also visual aspects of the speaker contribute relevant information). Most visual world studies have examined referentially mediated visual context effects on real-time language comprehension without showing the speaker (and even in dialogue studies the interlocutors are often separated by a divider, see, Brownschmidt, this volume). Relatedly, in many studies on speaker gaze effects, the gaze cue was implemented as a moving cursor overlaid on a video while the speaker herself was not shown (Brennan, Chen, Dickinson, Neider, & Zelinsky, 2007; Kreysa, 2009). Research on real-time speech-gesture integration has often used paradigms in which only either visual context and gestures (Wu & Coulson, 2005), or gestures and speech (but no referential visual context) were present (e.g., Kelly, Barr, Church, & Lynch, 1999; Kelly, Creigh, & Bartolotti, 2010). Meanwhile, however, first studies have examined how the presence of a speaker (her gaze, gestures, or mimics) affects real-time language comprehension when language is about objects and actions in visual context (e.g., Hanna & Brennan, 2007; Carminati & Knoeferle, 2013; Cook & Tanenhaus, 2009; Knoeferle & Kreysa, 2012; Nappa & Arnold, 2009). These recent investigations permit us to assess to which extent speaker-based information informs comprehenders' visual attention and language comprehension.

A first question is whether people can use a speaker's gaze and head movements to anticipate referents before they are mentioned. In a collaborative task, Hanna and Brennan (2007) found that listeners shifted attention to an object in their own workspace as soon as they saw the speaker attend to the corresponding object in her workspace (see also the discussion by Brown-Schmidt, this volume). Listeners can use speaker gaze and head movements flexibly to anticipate referents, even when the speaker is not human but a robot (Staudte & Crocker, 2009). Gaze does seems to modulate a listener's visual attention in a similar manner as other visual cues (e.g., an arrow), suggesting those two cues contribute at least somewhat similarly to language comprehension (Staudte, Crocker, Heloir, & Kipp, 2014).

That gaze and head shifts of a speaker rapidly affect processes of establishing reference may not be too surprising. But would they - much like information from objects or depicted actions (see Knoeferle et al., 2005, 2008) - also inform processes of thematic role assignment? In a recent study, a video-taped speaker referred to two out of three virtual (Second Life) characters on a computer screen, using either a canonical German subject(NP1)- verb-object(NP2) or a non-canonical object(NP1)-verb-subject(NP2) sentence (Knoeferle & Kreysa, 2012; Kreysa & Knoeferle, 2011b, see Fig. 2). She shifted gaze once from the pre-verbal to the post-verbal referent, a behavior which could, in principle, allow the listener to anticipate which character will be mentioned post-verbally. Post-comprehension, participants verified either the sentence referents (Experiment 1) or their role relations (Experiment 2). When participants had seen the speaker's gaze shift, they anticipated the NP2 referent before its mention and earlier than when the speaker had been obscured. This anticipation was stronger for subject- than object-initial sentences in both tasks. The difficulty associated with verifying thematic role relations for object-initial sentences was, however, eliminated entirely if listeners had followed the speaker's gaze shifts to the NP2 referent. Thus, gaze effects on visual attention are robust; they vary depending on the syntactic structure and thematic role relations conveyed by a sentence; and they can eliminate the difficulty in processing non-canonical thematic role relations, suggesting they are actively contributing towards structure building.

Gaze effects appear to resemble action effects in that both rapidly interact with syntactic structure building. But these two cues differ in that referents and actions are rapidly inspected during comprehension as they are mentioned, while the speaker is hardly fixated, even at the moment when she initiates a gaze shift. And indeed, this inspection difference has consequences for how gaze (vs. action depictions) affects anticipation of an upcoming referent. Kreysa et al. (2014) directly compared the effects of action depictions with that of speaker gaze and head shifts in the paradigm from Knoeferle and Kreysa (2012, see Fig. 2). In a first experiment, one action (depicted as a tool, e.g., balloons for 'congratulate') appeared between the middle character and the outer character together with the onset of the speaker's gaze shift from the middle (first-named) to the outer character. In a second experiment, two action tools appeared at the same time, thus forcing comprehenders to

Figure 2: The speaker inspected Second Life characters displayed on a 20inch iMac and shifted gaze from the first-mentioned middle character to one of the other characters as she described the scene, e.g., *Den Kellner beglückwünscht der Millionär ausserhalb des Geschäfts*, literally: 'The waiter (obj) congratulates the millionaire (subj) outside the shop'

process the action in more depth (they had to establish reference from the verb to the correct action). Processing was supported either by both speaker gaze and the action depiction, either of the two on its own, or neither. Comprehenders' eye movements to the target character increased immediately after the speaker had initiated her gaze shift to the target (the speaker herself was hardly inspected when she shifted gaze). The onsetting action, by contrast, attracted attention in both experiments, eliciting a somewhat delayed anticipation of the target character relative to the effects of the speaker's gaze shift. These results suggests that as we move towards richer visual contexts, we will want to consider the nature of the cue (e.g., how it relates to language and which semantic contribution it makes to comprehension).

Further studies have begun to examine recent action effects in relation to speech and gesture interpretation. In one study, people saw a video of an action (e.g., chopping vegetables) and subsequently verified whether a target trial consisting of gesture and speech was congruous or incongruous (Kelly, Özyürek, & Maris, 2010). Incongruence with the action for critical trials could either result from a mismatch with the gesture or with the speech, both of which could be weakly (e.g., *cut*) or strongly (*twist*) incongruous with the prime action (chopping). Participants' verification times were shorter and their responses were more accurate for action congruent than action-incongruent speech and gestures. The effect was more pronounced the stronger the incongruence (strong vs. weakly incongruous). Clearly then, recent actions can be semantically integrated with speech-gesture pairs.

It has also been shown that a speaker's emotional facial mimics can rapidly inform a listener's semantic in-

terpretation and visual attention. Emotion recognition changes across the lifespan with younger adults attending more to negatively than positively valenced stimuli (e.g., facial expressions or pictures); oder adults (above 60 years of age), by contrast, attend more to positively than to negatively valenced material (Carstensen, Fung, & Charles, 2003; Isaacowitz, Allard, Murphy, & Schlangel, 2009). Carminati and Knoeferle (2013) revealed similar qualitative differences in the effects of a speaker's emotional facial expressions on visual attention to pictures during sentence processing. In their study, younger and older adults inspected either a happy or an unhappy speaker face and subsequently listened to a sentence describing either a positive or negative event, both portrayed through photographs on the screen. Older adults looked at the photograph of the positive event during comprehension of the positive sentence more often than when they had inspected a negative (vs. positive) face. Younger adults, by contrast, showed such facilitation only for negative (vs. positive) prime faces and sentences. Visual attention and language comprehension in older compared with younger adults thus did not differ substantially concerning the time course. Rather, differences between the two age groups emerged in preferential eye-movement responses to positive compared with negative prime faces.

Overall, speaker-based visual information can thus rapidly affect language comprehension even in complex settings that include language, a visual referential context, and a visually portrayed speaker. However, the differential effects of speaker gaze and action depictions on visual anticipation of a relevant target character and the qualitatively distinct emotional priming effects also clarify that much remains to be learned about the interplay between attention to objects and actions, visual cues of the speaker (her gaze, gestures, and mimics), and ongoing utterance comprehension.

2.4 Abstract language-world relationships

I have so far discussed evidence that supports an active, temporally coordinated, and robust influence of all sorts of information in visual context during both spoken comprehension and reading. At the same time, a comprehender's cognitive capacities (e.g., attentional and working memory resources) can represent a bottleneck on visual context effects. This became apparent when the time course of visual context effects differed as a function of comprehenders' verbal working memory; when asynchronous cue presentation eliminated their semantic integration of gesture and speech; and when a lack of preview time or complexity of the visual context eliminated visual context effects on ambiguity resolution.

In many (but not all) the reviewed studies, language referred to, or was associated with, visual context information, thereby identifying it as relevant for comprehension. One exception are the effects of speaker-related information. However, while speaker-based cues are not directly referenced, they are relevant by virtue of the speaker's communicative role and her reference to objects. Accordingly, the speaker's gaze, gesture,

and mimics can be assumed to "point" to relevant visual context information. Moreover, a clear referential or associative relationship was present in all of the studies in that language related to and often directly referred to visual context (sometimes with varying degrees of congruence). Perhaps visual context effects are limited to situations in which language is about the visual context, and tasks in which participants are asked to try to understand both utterances and the related visual context. Alternatively, adults draw on visual context for "non-referential" language use also (e.g., when communicating abstract ideas), a finding that would speak to the importance of the visual context for language comprehension more broadly.

Visual-world evidence on the processing of concrete relative to abstract words comes from a study by Du*ñ*abeitia, Aviles, Afonso, Scheepers, and Carreiras (2008). Spanish participants listened to spoken sentences containing a critical (concrete vs. abstract) word. The critical spoken words were semantically associated with a visual target (e.g., concrete: 'crib' is associated with the depiction of a baby; abstract: 'smell' is associated with the depiction of a nose). The authors found that on hearing an abstract word, healthy adults rapidly inspected a target picture representing an associate of that word. While this was also the case for concrete words, inspection to the associate target picture was reduced and delayed for concrete relative to abstract words. These findings contribute to the mounting evidence that not just concrete words serve to relate object depictions to language but that visual attention to objects can - at least for associative language-world relationships - be guided even more strongly by abstract words.

Findings from a different, similarity judgment, task suggest further that abstract concepts (e.g., of semantic similarity) are linked with experiential concepts such as spatial distance (Casasanto, 2008). Participants in the study by Casasanto (2008) rated the similarity of words (Experiment 1) and faces (Experiment 2) that were either presented far away from, or in close proximity to, one another on a computer display. Participants' similarity ratings were higher when semantically similar words were presented closer together (vs. farther apart), and they were lower when dissimilar words were presented farther apart (vs. closer together). Face distance also affected similarity ratings but in the opposite direction such that similar faces were rated as more dissimilar when presented close to (versus far from) one another; dissimilar faces were rated as more similar when presented far (vs. close) from one another.

While these findings (Casanto, 2008) suggest a link of some sort between semantic similarity and spatial distance, the extent to which such a relationship would impact real-time sentence comprehension was, until recently, unclear. Two eye-tracking reading studies investigated whether the findings by Casasanto (2008) extend from similarity judgements to incremental semantic interpretation (Guerra & Knoeferle, 2014). Participants inspected two playing cards that were presented either far or close from one another on a computer display. The two cards either each showed a word (e.g., *Entspannung*, 'relaxation', and *Erholung*, 'recreation') which reappeared in an ensuing written sentence (Experiment 1), or they were blank (Experiment 3). Participants then

read a sentence implying either similarity (e.g., *Entspannung und Erholung sind fast äquivalent...*, 'Relaxation and recreation are almost equivalent...') or dissimilarity (*Aufregung und Entspannung sind eher andersartig ...*, 'Agitation and relaxation are rather different...'), and judged sentence veracity based on their world knowledge. After verifying the sentence, participants decided whether the pre-sentence card depiction matched (vs. mismatched) a target picture. Cards were thus irrelevant for the sentence comprehension task in both experiments.

First-pass reading times at the adjective 'equivalent / different' were modulated by the distance between words on the two playing cards and by the distance between two blank cards. Reading times were faster when card distance (far vs. close) matched (vs. mismatched) the semantic relationship of the nouns in the sentence (dissimilar vs. similar respectively), and this effect emerged at the adjective (Experiment 1) or second noun phrase (e.g., 'recreation', Experiment 3). Even when visual context was irrelevant for a comprehension task, and when there was no overt referential relationship between that context (e.g., blank playing cards) and an ensuing sentence, spatial characteristics of the context influenced language comprehension on a first pass through the sentence.

Summary The review and discussion in this section illustrate that key characteristics of situated language comprehension (such as the closely temporally coordinated interplay between comprehension and visual attention) hold up robustly across different language modalities, different types of visual context; for objects, actions, as well as speaker-based visual cues; and for abstract language. In fact, the coordinated interplay could be one of the causes of the observed robustness of visual context effects, since a closely temporally coordinated integration of linguistic and non-linguistic representations likely minimizes working memory load and maximises the impact of pictorial representations on language. Indeed, when cues were presented asynchronously, when information density was high (fast speech and rapid mention of objects), when scenes were complex or comprehenders could not preview them, then visual context effects were reduced or sometimes even eliminated altogether. Notwithstanding these limitations in light of our cognitive resources, the pervasiveness of the rapid coordinated interplay and of visual context effects corroborated the important role of visual contexts for language comprehension.

3 Accounting for (situated) language comprehension

Existing accounts accommodate the interplay between comprehension, (visual) attention to relevant visual context information, and subsequent feed-back of visual context information into comprehension processes. The 'Coordinated Interplay Account' (CIA) achieves this via three informationally and temporally dependent stages (Knoeferle & Crocker, 2006, 2007). A first stage, sentence interpretation, covers the processes of incremental

sentence comprehension. The resulting interpretation feeds into the second stage (utterance-mediated attention) in which aspects of the current interpretation contribute to shifts in visual attention. Attended aspects of the immediate scene and their resulting representations, or of recent scene representations in working memory can then feed back into interpretation processes (see Crocker et al., 2010; Mayberry, Crocker, & Knoeferle, 2009)[7].

3.1 Complementing eye movements with ERP measures

Based on the evidence reviewed in section 2, it seems obvious that more comprehensive accounts will need a model of speaker-based information, including, her gaze, gestures, and mimics. What is also noticeable is that the extant accounts include a relatively coarse-grained model of different linguistic processes. Of course, the level of grain at which comprehension processes can be accommodated depends - amongst other things - on how well we can infer them from available measures (e.g., object-directed visual attention). For visually situated sentence comprehension, researchers have largely focused on qualitative linking hypotheses, and have established two key links between utterance comprehension and visual attention. A first is that comprehenders shift their visual attention to objects as they interpret an utterance, and these shifts reflect processes of establishing reference and lexico-semantic associations (Cooper, 1974; Tanenhaus et al., 1995; Dahan & Tanenhaus, 2005; Huettig & Altmann, 2005). Furthermore, comprehenders tend to anticipate objects before they are mentioned if the linguistic context is sufficiently constraining, and these anticipatory looks reflect their expectations (Altmann & Kamide, 1999; Cooper, 1974). However, an increase in visual attention to objects has also been interpreted as reflecting semantic interpretation (e.g., Sedivy et al., 1999) or syntactic structuring (e.g., Tanenhaus et al., 1995) depending on specifics of the experimental design. While good experimental design can isolate behavior that indexes different comprehension processes, the fact that the same measure was taken to index these two different comprehension processes highlights the need for more detailed linking assumptions.

Extant studies have exploited neuroscientific measures such as ERPs to gain complementary insight into the type of comprehension processes implicated in different visual context effects (see Knoeferle, in pressa). Recall, for instance, the temporally distinct brain responses observed in response to verb-action compared with thematic role relations mismatches (Knoeferle et al., 2014). The N400 congruence effects observed in response to verb-action mismatches (vs. matches) had a centro-parietal distribution reminiscent of the topography of N400 effects in strictly linguistic contexts (see Kutas & Hillyard, 1980; Kutas & Federmeier, 2011, for discussion of the N400). N400 congruence effects to thematic role mismatches, by contrast, had a more anterior distribution, similar to N400s effects observed during more pictorial-based semantic processing (Ganis et al., 1996).

The CIA has in the meantime been extended to accommodate these distinct brain responses and can also model verification response latencies (Knoeferle et al., 2014). In the 2014 version of the account, a response

index, set to true or false, tracks congruence; sentential and scene-based representations are also indexed for (in)congruence as well as for the type of process (establishing reference from the verb to an action vs. from depicted to sentential-thematic role relations). This new version of the CIA has also begun to model characteristics of the comprehender such as his verbal working memory capacity as well as the timing of stimulus presentation. These two parameter can then impact the time course of different picture-sentence matching processes (and visual context effects more generally). More cognitive resources and / or more time to process a stimulus would lead to more detailed and highly active scene-based representations which could lead to more pronounced role congruence effects. These parameters can accommodate the variation in picture-sentence congruence effects observed in the ERP studies (Knoeferle et al., 2011, 2014). But they can also accommodate the elimination of visual context effects in visual-world studies when no preview time was given (Ferreira et al., 2013). In the CIA this would be accommodated by a lack of detail in the visual representations since a short preview does not permit encoding a complete model of the visual context. The elimination of visual context effects with complex scenes in Ferreira et al. (2013) would similarly be accommodated through incomplete scene representations given resource constraints (and thus no clear effects of these representations on language processing).

3.2 Task constraints as a means to refine linking hypotheses

A further promising way of further developing models of situated language processing would be to refine the visual-world linking hypotheses such that we know precisely which gaze pattern reflects specific comprehension sub-processes. To the extent that such a unique link exists, we could derive more specific predictions about the deployment of visual attention and associated comprehension.

Existing research has begun to dissociate some of the more frequently analyzed measures and has confirmed that different gaze measures can index different underlying processes. Altmann and Kamide (2004) compared proportions of trials with fixations to target objects in a given time window with the proportion of trials on which participants launched a saccade to these objects. Their analyses showed that these two measures can be dissociated, suggesting they may reflect only partially overlapping cognitive and comprehension processes. For instance, when participants have heard *The man*, they may initially launch saccades to the man and then continue to inspect it. As time passes, however, they will be less likely to make additional saccades to the man and will begin to saccade to another object. The probability of fixating the man may, however, at the same time remain high, resulting in a dissociation of saccade launch and fixation probabilities.

While we are still in the process of better understanding how these (and other aspects) of the eye-movement record relate to comprehension processes, additional measures have been introduced. In a study by Arai, van

Gompel, and Scheepers (2007) participants read out aloud a prime sentence in either direct object or prepositional object structure after which they listened to a spoken target sentence that was temporarily ambiguous between these two structures, and inspected a related scene. Arai et al. reported first gaze duration (the duration of consecutive fixations for the first inspection to an object in a given time window) as an index of expectations about which object would be mentioned next, and target objects symbolized alternative sentence interpretations and structures. Longer first gaze durations on a primed target object / sentence structure co-varied with increased fixations to that object and were taken to index anticipation of the target object (see also Scheepers & Crocker, 2004).

In other tasks such as sentence-picture verification, by contrast, overall longer fixation durations have been reported when picture and sentence were incongruous (vs. congruous), and have been associated with additional picture-sentence comparison operations (e.g., Underwood et al., 2004). Pupil size measures have also been reported as an index for processing difficulty in situated sentence comprehension. Scheepers and Crocker (2004) had participants read out an object-initial, a subject-initial, or a neutral sentence; participants then listened to an initially structurally ambiguous spoken target sentence in either subject-verb-object or object-verb-subject order that related to depicted events. Pupil size increased when the spoken target sentence was disambiguated towards the non-canonical object-initial (relative to subject-initial) structure, a finding that was interpreted as indexing processing difficulty for non-canonical relative to canonical sentence structures. Relatedly, Engelhardt, Ferreira, and Patsenko (2010) reported pupil size as a measure of processing incongruence between visual context and prosody. Participants in their study listened to a spoken sentence, inspected a visual context and answered a comprehension question about the sentence. When visual context and prosody were incongruous (vs. congruous), pupil size increased.

While these studies have begun to explore how different gaze measures relate to cognitive and comprehension processes, much remains to be learned about the linking of specific gaze pattern to specific comprehension processes. One way of better understanding this linking could be by systematically manipulating the comprehension task. Research on scene perception, for instance, has shown that task plays an important role in guiding visual attention (see Tatler, this volume). When participants had to determine the age of characters in a painting (compared with when they estimated their wealth), they were more likely to inspect the faces of the portrayed characters (see Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010; Yarbus, 1967). Similarly, the allocation of visual attention in an image differed for a visual search vs. memorization task (Castelhano, Mack, & Henderson, 2009). It has been argued that task could also play an important role for visually situated comprehension [see Salverda et al., 2011 for a review of task effects in scene perception and task-based variants of the visual world paradigm]. To the extent that this holds, task could constrain, and help to refine, linking hypotheses. Improving (i) the linking assumptions between language comprehension processes and one of the

key measures used to examine situated comprehension (visual attention to objects across time, see, Altmann & Kamide, 2007; Burigo & Knoeferle, 2011; Tanenhaus, Magnuson, Dahan, & Chambers, 2000; Tanenhaus, 2004) and (ii) our understanding of how different tasks affect visual attention during language comprehension (see Kreysa & Knoeferle, 2011a; Salverda, Brown, & Tanenhaus, 2011), will be an important step in further developing existing accounts of situated comprehension.

Good evidence that task can affect eye gaze in situated language comprehension comes from a study that compared picture-sentence verification and passive listening tasks (Altmann & Kamide, 1999). That study is best known for reporting anticipatory gaze effects to objects before their mention. When a sentential verb restricted later reference to a single target object (vs. to four objects), people began to shift attention to that singled-out target object before its mention. While this finding held independent of task, a comparison of the picture-verification and passive listening task also revealed task-specific aspects such that the target object was inspected later and less often during passive listening compared with verification.

Such a direct comparison of how different tasks affect language comprehension has so far been the exception. Furthermore, existing task-based approaches to situated language comprehension discuss task effects at a relatively coarse level of grain - such as when comparing verification with passive listening, or act-out tasks (see Salverda et al., 2011). This is an important first step, inspired by task effects in scene perception research. It risks underestimating, however, the role of task in elucidating the mechanisms underlying situated language comprehension. If we in addition examine more subtle task manipulations that bear on different levels of linguistic structure (e.g., processes of establishing reference versus thematic role assignment), then we can compare visual attention deployment in response to subtly differing sub-processes in comprehension. At the same time, we can leverage task constraints to refine our linking assumptions between comprehension sub-processes and visual attention. Note that the idea here - of introducing a task-based approach - is similar to the proposal by Salverda et al. (2011). The difference is in the level of grain and in choosing linguistically-meaningful task constraints that can isolate sub-processes of language comprehension..

One recent example compared how a referent verification compared with a thematic role verification task modulated the allocation of visual attention (Knoeferle & Kreysa, 2012; Kreysa & Knoeferle, 2011a). If participants verify reference versus thematic role assignment, eye gaze on correctly answered trials can be assumed to reflect the respective foci of these two tasks. In these studies, a videotaped speaker referred to depicted characters, using either a German subject-verb-object or a non-canonical object-verb-subject sentence (see section 2 and Fig. 2). At the verb, the speaker shifted gaze from the pre-verbal to the post-verbal referent. In conditions where the speaker was visible, speaker gaze and head shifts made it possible to anticipate the upcoming post-verbal (object or subject) referent. Following the video, participants verified whether a schematic depiction correctly highlighted different aspects of the immediately preceding video and sentence. The experi-

ments differed only in which aspects of the video had to be verified. In a first study, the template showed three stick figures two of which were circled. People had to judge whether the circled referents corresponded (via their position) to the sentential referents shown in the video. In a further experiment, the template circled one out of three characters and people verified whether the circled character was the sentential patient (Kreysa & Knoeferle, 2011a). In a third study, template characters weren't circled but an arrow between two out of three stick figures represented directionality (agent-patient) of thematic role relations and people verified whether the schematically indicated role relations matched those of the preceding sentence.

Speaker gaze rapidly influenced the allocation of visual attention during comprehension such that participants inspected the post-verbal referent earlier in all tasks. When the post-sentence task focused attention on the patient, people anticipated the post-verbal referent more often when it was the patient (as was the case in subject-verb-object sentences) than when it was the agent (as was the case in object-verb-subject sentences). This suggests that visual interrogation of the scene focused on task-relevant aspects. For referent and thematic role verification, by contrast, the gaze pattern was reversed. Listeners anticipated the post-verbal referent more strongly when it was the agent (in object-initial sentences) than the patient (in subject-initial sentences). Thus, different post-sentence verification tasks affected which character participants inspected most, and modulated the effects of sentence structure and speaker gaze on visual attention allocation to the target character.

3.3 Conclusions

Even a review of a small part of the literature clarifies that accounts of visually situated comprehension must cover a variety of visual context effects. Visual context effects were observed both during spoken comprehension and in reading; with different degrees of match between language and visual context (ranging from a perfect match to incongruence); for different language-world relationships (referential and associative links but also links between abstract language and unrelated visual context); and for contexts that show objects, events and speakers with their gaze, gestures, and mimics. It has become clear that key characteristics such as the robustness of real-time visual context effects, as well as the temporally coordinated interplay of visual attention and language comprehension emerged across-the-board even in complex contexts (real-world events, cluttered photographs; when both a speaker and referential visual context were shown; when representations of non-linguistic content had to be retrieved from working memory; and when language was abstract and entirely unrelated to visual context).

In light of these pervasive visual context effects, it is becoming increasingly important that we arrive at a better understanding of how continuous measures such as visual attention (but also complementary measures such as event-related brain potentials) reflect different sub-processes of real-time comprehension such as estab-

lishing reference, assigning thematic roles, or constructing a temporal and spatial model of the comprehension situation. When considering visual attention, it has become clear that a single stream of eye gaze can be related to many different comprehension sub-processes. Complementing eye-movement with ERP measures is one solution to improving our insight the type of processes implicated in visual context effects. Ultimately, however, we must obtain a better model of how different sub-processes in language comprehension are reflected in visual attention, perhaps via direct comparison of visual attention deployment across subtle comprehension sub-tasks.

Notes

[1]Elsewhere it has been pointed out that the visual context plays an important role at the early stages of child language development and that this primary role speaks to its importance for communication at the adult life stage (see Knoeferle, in pressb). The present chapter complements this argument with an overview of the pervasiveness of visual context effects.

[2]A linking hypothesis relates patterns in the data to cognitive processes.

[3]The distinction between the N400 and P600s is not always clear-cut; a 'semantic' P600 has been observed in response to what looked like semantic violations (Kolk, Chwilla, Van Herten, & Oor, 2003; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003). Ambiguity in the linking assumptions leads to ambiguity in comprehending and modeling the implicated comprehension processes (see Kutas, Van Petten, & Kluender, 2006; Tanenhaus, 2004). For eye movements during spoken word recognition, Allopenna, Magnuson, and Tanenhaus (1998) provide a formal linking hypothesis but overall, the linking between the eye-gaze record and language comprehension processes is relatively underspecified (see section 3)

[4]Cooper (1974) conducted studies using the same method but at the time the potential of this new paradigm was not recognised

[5]There are only few reports of delayed responses in a listener's visual attention, and these have been interpreted as reflecting time-consuming comprehension processes (e.g, Huang & Snedeker, 2009, the computation of scalar implicature)

[6]It will be interesting to see to which extent this finding extends to other kinds of world-language relations

[7]For a discussion of alternative accounts, see Knoeferle et al. (2014)

Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Center (DFG) at Bielefeld University, Germany.

References

Abashidze, D., Carminati, M. N., & Knoeferle, P. (2014). How robust is the recent-event preference? In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society (p. 92-97). Cognitive Science Society.

Abashidze, D., Knoeferle, P., Carminati, M. N., & Essig, K. (2011). The role of recent real-world versus future events in the comprehension of referentially ambiguous sentences: Evidence from eye tracking. In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), Proceedings of the European Conference on Cognitive Science. New Bulgarian University Press.

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

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

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

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

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

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. Cognition, 111, 55-71.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex spsech and scenes during language comprehension. Acta Psychologica, 137, 208-216.

Arai, M., van Gompel, R., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. Cognitive Psychology, 54, 218-250.

Arbib, M. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioural and Brain Sciences, 28, 105–167.

Barsalou, L. W. (1999). Perceptual and symbol systems. Behavioural and Brain Sciences, 22, 577–609.

Brennan, S., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2007). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. Cognition, 106, 1465-1477.

Burigo, M., & Knoeferle, P. (2011). Visual attention during spatial language comprehension: Reference alone isn't enough. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Cognitive Science Society.

Carminati, M. N., & Knoeferle, P. (2013). Effects of speaker emotional facial expression and listener age on incremental sentence processing. PLoS ONE, 8, e72559.

Carstensen, L. L., Fung, H. H., & Charles, S. T. (2003). Socioemotional selectivity theory and the regulation of emotion in the second half of life. Motivation and Emotion, 27, 103-123.

Casasanto, C. (2008). Similarity and proximity: When does close in space mean close in mind? Memory & Cognition, 36, 1047–1056.

Castelhano, M., Mack, M., & Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. Journal of Vision, 9(3), 1?15.

Chambers, C. G., & San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. Cognition, 108, 26–50.

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

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

Cook, S. W., & Tanenhaus, M. K. (2009). Embodied communication: Speakers? gestures affect listeners? actions. Cognition, 113, 98–104.

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., Knoeferle, P., & Mayberry, M. (2010). Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. Brain and Language, 112, 189-201.

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

Duñabeitia, J. A., Aviles, A., Afonso, O., Scheepers, C., & Carreiras, M. (2008). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. Cognition, 110, 284-292.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. Quarterly Journal of Experimental Psychology, 63, 639–664.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. Journal of Memory and Language, 69, 165-182.

Fodor, J. (1983). Modularity of mind. Cambridge, MA: MIT Press.

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

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences, 6, 78-84.

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.

Goolkasian, P. (1996). Picture-word differences in a sentence verification task. Memory & Cognition, 24, 584–594.

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

Guerra, E., & Knoeferle, P. (2014). Effects of object distance on incremental semantic interpretation: similarity is closeness. Cognition, 133, 535-552.

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. Journal of Cognitive Neuroscience, 23, 1845-54.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. Language and Cognitive Processes, 8, 439–483.

Hanna, J., & Brennan, S. (2007). Speakers's eye gaze disambiguates referring expressions early during face-to-face conversation. Journal of Memory and Language, 57, 596–615.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), The interface of language, vision, and action: eye movements and the visual world (pp. 1–58). New York: Psychology Press.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics?pragmatics interface. Cognitive Psychology, 58, 376?415.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. Cognition, 96, 23-32.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica, 137, 151–171.

Isaacowitz, D. M., Allard, E. S., Murphy, N. A., & Schlangel, M. (2009). The time course of age-related preferences toward positive and negative stimuli. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 64B, 188-192.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. Journal of Memory and Language, 40, 577–592.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. Journal of Cognitive Neuroscience, 22, 683–694.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. Psychological Science, 21, 260–267.

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. (in pressa). Language comprehension in rich non-linguistic contexts: combining eye tracking and event-related brain potentials. In Towards a cognitive neuroscience of natural language use. Cambridge: Cambridge University Press.

Knoeferle, P. (in pressb). Visually situated language comprehension in children and in adults. In R. K. Mishra, N. Srinivasan, & F. Huettig (Eds.), Attention and vision in language processing. Springer Language and Cognition series.

Knoeferle, P., Carminati, M. N., Abashidze, D., & Essig, K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. Frontiers in Psychology, 2, 376.

Knoeferle, P., & Crocker, M. W. (2005). Incremental effects of mismatch during picture-sentence integration: Evidence from eye-tracking. In Proceedings of the 27th Annual Meeting of the Cognitive Science Conference (pp. 1166–1171). Mahwah, NJ: Erlbaum.

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

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. Journal of Memory and Language, 75, 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, 789–795.

Knoeferle, P., & Kreysa, H. (2012). Effects of speaker gaze on syntactic structuring. Frontiers in Psychology, 2, 376, doi: 10.3389/fpsyg.2011.00376.

Knoeferle, P., Urbach, T., & Kutas, M. (2014). Different mechanisms for role relations versus verb?action congruence effects: Evidence from erps in picture?sentence verification. Acta Psychologica.

Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental

sentence comprehension: insights from ERPs and picture-sentence verification. Psychophysiology, 48, 495–506.

Kolk, H., Chwilla, D., Van Herten, M., & Oor, P. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. Brain and Language, 85, 1-36.

Kreysa, H. (2009). Coordinating speech-related eye movements between comprehension and production. Unpublished doctoral dissertation, University of Edinburgh, UK.

Kreysa, H., & Knoeferle, P. (2011a). Effects of speaker gaze on spoken language comprehension: task matters. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society. Cognitive Science Society.

Kreysa, H., & Knoeferle, P. (2011b). Peripheral speaker gaze facilitates spoken language comprehension: Syntactic structuring and thematic role assignment in German. In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), Proceedings of the European Conference on Cognitive Science. New Bulgarian University Press.

Kreysa, H., Knoeferle, P., & Nunnemann, E. (2014). Effects of speaker gaze versus depicted actions on visual attention during sentence comprehension. In M. M. . B. S. Paul Bello Marcello Guarini (Ed.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (p. 2513-2518). Cognitive Science Society.

Kuperberg, G., Sitnikova, T., Caplan, D., & Holcomb, P. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. Cognitive Brain Research, 17, 117–129.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). Annual Review of Psychology, 62, 621-647.

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

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.), (2nd Edition ed., p. 659-724). New York: Elsevier.

Mayberry, M., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. Cognitive Science, 33, 449–496.

Nappa, R., & Arnold, J. (2009). Paying attention to intention: Effects of intention (but not egocentric attention) on pronoun resolution. In Proceedings of the CUNY Conference (p. 262).

Nation, K., & Altmann, C. M. M. G. T. M. (2003). Investigating individual differences in childrens real-time sentence comprehension using language-mediated eye movements. Journal of Experimental Child

Psychology, 86, 314–329.

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

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

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers? and listeners? eye movements and its relationship to discourse comprehension. Cognitive Science, 29, 1045–1060.

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

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual-world studies. Acta Psychologica, 137, 172–18.

Scheepers, C., & Crocker, M. W. (2004). Constituent order priming from reading to listening: a visual-world study. In M. Carreiras & J. C. Clifton (Eds.), The on-line study of sentence comprehension: Eyetracking, ERP, and beyond. United Kingdom: Psychology 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.

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

Staudte, M., & Crocker, M. (2009). The effect of robot gaze on processing robot utterances. In N. Taatgen & H. van Rijn (Eds.), Proceedings of the Cognitive Science Conference (pp. 431–436). Cognitive Science Society, Inc.

Staudte, M., Crocker, M. W., Heloir, A., & Kipp, M. (2014). The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. Cognition, 133, 317-328.

Tanenhaus, M. K. (2004). On-line sentence processing: Erps, eye movements, and beyond. In M. Carreiras & J. Charles Clifton (Eds.), (p. 209-228). Psychology Press.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken- language comprehension: evaluating a linking hypothesis between fixations and linguistic

processing. Journal of Psycholinguistic Research, 29, 557?580.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 632–634.

Tatler, B., Wade, N., Kwan, H., Findlay, J., & Velichkovsky, B. (2010). Yarbus, eye movements, and vision. i-Perception, 1, 7-27.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: eye movements in general encoding and in focused search. The Quarterly Journal of Experimental Psychology, 56, 165–182.

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, 967-982.

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

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. Psychophysiology, 42, 654–667.

Yarbus, A. L. (1967). Eye movements and vision. New York: Plenum Press.

Zhang, L., & Knoeferle, P. (2012). Visual context effects on thematic role assignment in children versus adults: Evidence from eye tracking in german. In . R. P. C. Naomi Miyake David Peebles (Ed.), Proceedings of the annual meeting of the cognitive science society (p. 2593-2598). Boston, USA: The Cognitive Science Society.