Attention and eye movement metrics in visual world eye tracking

Pirita Pyykkönen-Klauck[1,2] & Matthew W. Crocker[1]

[1]Department of Computational Linguistics, Saarland University
[2]Department of Psychology, Norwegian University of Science and Technology

Abstract

This chapter introduces the visual world paradigm, with the aim of identifying both the opportunities and challenges researchers are presented with when using overt visual attention as an index of the cognitive processes and mechanisms involved in a variety of language processing tasks. The chapter also provides an overview of the linking hypotheses that underlie the coordination of visual and linguistic information. Finally, the chapter discusses key properties of the eye-movement metrics used in the visual world studies – and different approaches to their analysis – in order to support sound interpretations with respect to the underlying theories of visually situated language comprehension that these studies are used to investigate.

While people seemlessly integrate spoken language with information from their visual environment, this ability entails the rapid and adaptive coordination of the relevant cognitive systems involved. The visual world paradigm – which allows us to monitor attention in the visual scene during spoken comprehension – exploits this behavior to investigate the mechanisms that underlie incremental language processing. The value of the paradigm arises from two key findings: Firstly, speech-mediated eye movements to relevant visual targets are closely time-locked to the linguistic stimuli – typically emerging after about 200ms. Secondly, such eye movements have been shown to index a number of underlying comprehension mechanisms, including lexical access (Allopena et al 1998), referential processing (Tanenhaus 1995; see also Engelhard & Ferreira, this volume; Van Gompel & Järvikivi, this volume), and anticipatory processes (Altmann & Kamide, 1999; Kamide, 2008 for a review).

While the neural circuitry involved in controlling the eye movements in visual world studies is always the same, however, cognitive systems interact with this circuitry differently under different task requirements (Rayner, 2009). Additionally, studies have shown that a range of eye-movement measures, such as fixation durations and saccade lengths, do not correlate across scene perception, visual search and reading tasks, also when measured within participants (Andrews & Coppola, 1999; Castelhano & Henderson, 2008; Rayner, Li, Williams, Cave, & Well, 2007). This raises the question: How can we best exploit eye movements as a window to the mind in the visual world paradigm?

To address this question, this chapter reviews several established variants of the visual world paradigm, to identify challenges and opportunities that researchers face when using overt visual attention as an index of underlying cognitive processes and mechanisms, in a range of different tasks. The chapter provides an overview of the linking hypotheses that underlie the

coordination of visual and linguistic information. Finally, the chapter discusses the nature of the eye-movement metrics used in various visual world studies and different solutions to analyzing them in order to derive interpretations of processes of visually situated language comprehension, and informing about cognitive mechanisms underlying these processes.

Measuring and interpreting attention in active and passive tasks

As the chapters in this book reveal, the visual world paradigm has been successfully applied to a variety of research questions related to the study of language and attention (see the chapter by Spivey and Huette, this volume). To better understand the use of this paradigm we first consider two variations, as characterized by the tasks used in the experiments: (1) In an active task participants are required to give an explicit response based on spoken instructions such as *put the apple in the box* or *click on the bacon* (e.g., Spivey & Huette, this volume; Tanenhaus et al., 1995). (2) Passive "look-and-listen" tasks mimic what listeners spontaneously do in natural situations such as when watching television, and other more passive communicative settings: People are not required to perform any additional motoric (hand) movements; instead, they are only instructed to listen to the stories for comprehension (e.g., Knoeferle, this volume; Van Gompel & Järvikivi, this volume).

Active task. Active tasks inherently offer two dependent measures of the comprehension process: Firstly, the motor action typically provides an explicit index of the final interpretation and its correctness. Secondly, listeners' incremental comprehension processes are revealed by measuring overt attention to pictures representing the words of the instructions in a highly time-sensitive manner (for an overview, see Spivey & Huette, this volume). As their overview shows,

people tend to find the correct referent around 1 second after it is mentioned in tasks (as the peaked proportions show), with whatever other potential objects are considered attracting overt attention during the first second. This means that measurement of comprehension processes is practically limited to the inspection of 3-4 potential visual locations during this time window (see Tatler, this volume). However, even though this sounds like a limited amount of time, current eye-tracking technology can sample the location of the eyes up to every 0.5 milliseconds (1-4 ms is more typical) permitting high resolution aggregation and comparison of the number of fixations and saccades to relevant scene targets during these time intervals.

Typically, visual attention to scene objects is determined by fixations – still moments when the eyes are foveating particular points in the scene – and used as an overt measurement of information people are attending to at any given moment during that second (for an overview of measurements, see Holmqvist, Nyström, Andersson, Dewhurst, Halszka, & van de Weijer, 2011; Rayner, 2009). Each fixation is also preceded by a saccade, a fast movement to the location where the fixation is observed. While fast eye movements could in principle be of interest in understanding the allocation of visual attention during active tasks, it has been argued that people do not receive new information during saccades and that – in some situations – the processing of the previously fixated information continues over the saccades (e.g., Irwin, 1998; Irwin & Carlson-Radvansky, 1996). Thus, saccades are typically not employed as measurements of attention in these tasks.

One challenge for the interpretation of the eye movements in active tasks is that performance is dependent on two different processes: Firstly, participants must comprehend the spoken instructions; and secondly, plan and perform a motor action with their hands based on those instructions. Even though these experiments have shown that people tend to fixate the

visual referents of the spoken input, it is difficult to identify which proportion of the fixations/looks that are controlled by these language comprehension mechanisms and which proportion are controlled by the planning mechanisms of a hand-based motor action (see e.g., Boland, 2005, Hayhoe & Ballard, 2005, for a discussion). However, as discussed by Spivey and Huette (this volume), the research questions addressed in active studies typically do not depend on the disentangling of comprehension and motor effects from each other.

Passive task. As in the active tasks, passive tasks elicit people's overt attention to the relevant regions of visual scenes that are related to speech content. Passive tasks are typically applied to study either predictive processes in language comprehension (e.g., Altmann & Kamide, 1999; Kamide et al., 2003; Knoeferle & Crocker, 2007; Knoeferle et al., 2005) or more inferential processes such as referential processing in pronoun resolution (e.g., Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Arnold, Brown-Schmidt, & Trueswell, 2007; Colonna, Schimke, & Hemforth, 2014; Järvikivi, Van Gompel, Hyönä, & Bertram, 2005, 2014; Pyykkönen & Järvikivi, 2010; Pyykkönen, Matthews, & Järvikivi, 2010). Broadly, the measurements of passive tasks do not differ from active tasks, i.e., fixation-based metrics of looks is used over the time to study the nature of comprehension processes (cf. Altmann & Kamide 1999; 2009 for a saccade-based metrics).

Passive tasks are typically used with two kinds of instructions: (1) Explicit instructions which encourage participants to attend to pictures while listening to the stories (e.g., Järvikivi et al., 2005; Knoeferle & Crocker, 2007; Pyykkönen, Hyönä, & Van Gompel 2010; Pyykkönen & Järvikivi, 2010), or (2) implicit instructions, where participants are not given any direct encouragement to relate the depicted pictures to the spoken stimuli (e.g., Altmann & Kamide,

1999, Experiment 2; Knoeferle & Crocker, 2006). Because passive tasks do not require any participant response that could be used to verify the correctness of the final interpretation, different secondary tasks have sometimes been added to improve control over participants' behavior in either explicit or implicit instruction studies. Examples of such secondary tasks are sentence-picture verification task (e.g., Altmann & Kamide, 1999, Experiment 1; Arnold, et al., 2000, Staudte & Crocker, 2011; 2014, Knoeferle, this volume), comprehension questions (Arnold et al., 2007) or a continuation task in which people are asked to continue the story by using the words and pictures occurring in the task (e.g., Järvikivi et al., 2005; Pyykkönen et al., 2010; Pyykkönen & Järvikivi, 2010).

An important question to ask is whether participants' eye-movement patterns change when different secondary tasks are used in passive visual world studies. So far, there is no evidence to suggest that this is the case. For example, Altmann and Kamide (1999) showed that when people were given a picture verification task (Experiment 1) the pattern of eye movements were the same as when they were not specifically required to relate the pictures to the speech (Experiment 2). In other words, people related the spoken language to the pictures even without specific instructions. This robustness of language-mediated gaze behavior is also shown by Altmann (2004), who found similar eye-movement behavior in the Blank Screen Paradigm, where the visual scene was presented only prior to the spoken stimulus and thus absent when people were listening to the spoken language stimuli (though see Knoeferle & Crocker 2007, for evidence of reduced fixations in the blank screen version of their studies). Relatedly, Pyykkönen-Klauck & Järvikivi (2015) conducted a passive visual world pronoun resolution study in which they asked participants to judge the referent for an ambiguous pronoun at the end of the stories such as (1).

(1) PASSIVE VISUAL WORLD (*pictures of a rabbit, a fox, a river*)

There is the rabbit and the fox. The rabbit strokes the fox near the river. ***He*** wanted to tell about his last night's dream.

COMPREHENSION QUESTION:

Who wanted to tell his last night's dream?

They found that in the condition where visual information stayed constant (showing the rabbit, the fox and the river throughout the story), the answer to the comprehension question correlated with eye movements: The most preferred referent was the first-mentioned subject "rabbit". This finding is also mirrored in the eye-movement analyses showing a strong and persistent "rabbit" preference soon after the onset of ***He***. However, when one of the referents walked away from the screen just before listeners heard the ambiguous pronoun, adults' eye-movement patterns no longer directly reflected their final interpretation. While adults did look at the antecedent candidate that stayed on the visual scene, they selected the first-mentioned subject ("rabbit") in the final interpretation.

This study also replicated the findings of previous pronoun resolution studies showing a first-mentioned subject preference for ambiguous pronoun selection among adult participants using a different secondary task (Järvikivi et al. 2005, 2014). This further implies that when the visual scene continuously shows both antecedents, the final comprehension question is not needed to verify the results interpreted from eye movements only. However, if the scene manipulates the availability of the potential antecedents, then the final question is needed to

determine how people resolve the pronoun and how this resolution relates to the eye-movement patterns during comprehension (see also Burigo & Knoeferle, 2015).

Comparing the findings from active and passive tasks. Most studies that discuss mechanisms and processes involved in visually situated language comprehension do not make any differentiations on whether each study was conducted with an active or passive task. As the above descriptions show, there are lots of commonalities among these studies. However, some studies have shown that eye movements are differentially sensitive across different tasks, and thus this issue should not be completely ignored across different versions of the visual world studies either. Thus, the differences are important to be considered in experiment design; and would deserve space for further discussions also in the literature reviews.

One particularly revealing set of studies by Weber and Crocker contrasted the influence of "bottom-up" word frequency – which is known to reflect lexical access as revealed in visual world studies that us an active clicking task (Dahan, Magnuson, & Tanenhaus, 2001) – with the "top-down" semantic influence of restrictive and non-restrictive verbs (Altmann & Kamide, 1999). In a passive listening study, (Weber & Crocker, 2012), participants listened to sentences like "The woman finds/iron the blouse" pair with a scene containing a woman, blouse (low frequency), distractor, and a high-frequency phonological competitor to blouse. In the unrestrictive condition ("finds"), listeners first inspect the higher-frequency phonological competitor as they heard "blouse", and only later inspected the image of the blouse. In the restrictive condition ("irons"), most looks were at the target (blouse) immediately following the verb, with looks to high-frequency competitors nonetheless still showing a significant advantage over the distractor. In a further study, reported in Weber et al (2010), the same experiment was

conducted again, but with an active response task that is standard in lexical access studies (participants were instructed to click on the second noun mentioned in the sentence). This study revealed an enhanced sensitivity of eye movements to frequency, suggesting that task may differentially modulate competing generators of gaze behavior.

In another example, Hayhoe and Ballard (2005) conducted an experiment using tasks that did not include language comprehension, but simply required participants to either search for an object or reach for an object. The overall probability of fixations was higher when people were reaching towards an object than when they were searching it (see also Castelhano, Mack & Henderson, 2009 for eye-movement differences between searching task and memorizing the scene). Similarly, Huestegge (2010) found that for tasks in which participants are reading text and simultaneously asked to articulate, eye movements to the text differed from the task in which participants were simply asked to read a text without articulation. As Rayner (2009) had also pointed out, there are measurable differences in oral and silent reading: In the oral task readers are not only reading for comprehension but also need to make the effort to articulate the words. This leads, for example, longer fixations in the oral than in the silent reading task. Bridging these findings back to the active and passive tasks in visual world studies, we can conclude that even though similar measurements can be used in both kinds of tasks, often supporting similar interpretations, there are some limitations when comparing the results across active and passive tasks – both at the measurement level, and regarding what processes eye movements are sensitive to.

Linking visual environment and language

In order to understand how looking patterns are indicative of underlying cognitive processes in the interaction of spoken language and visual perception, previous research has posited a number of linking hypotheses. These hypotheses aim to explain in detail how attention in language and attention in visual scenes are coordinated, and how people form meaningful representations of events in visually situated language comprehension. Two of the more fully articulated proposals for linking hypotheses and their corresponding computational models are presented by Crocker and colleagues (Crocker, Knoeferle, & Mayberry, 2010, Knoeferle & Crocker, 2006, 2007; Mayberry, Crocker, & Knoeferle, 2009) and by Altmann and colleagues (Altmann & Kamide, 2007, 2009; Altmann & Mirkovic, 2009).

The Coordinated Interplay Account (CIA: Knoeferle & Crocker, 2006, 2007) specifies three separate processes involved in establishing interpretation of visually situated spoken language: (i) searching for visual referents of spoken referring expressions, (ii) grounding referring expressions with objects and events in the scene, and (iii) use of visual scene to confirm or inform the linguistic interpretation. In the first two phases of searching for referents and grounding referring expressions, speech input guides overt attention in the visual environment. Here, people search for objects and events in the visual scene (including episodic representations of recent visual information that may no longer be visible) to identify likely visual referents, as well as to anticipate potential or likely up-coming referents. In the last phase, after people have overtly attended to the (anticipated) referents, the visual environment influences the comprehension process by confirming or altering the interpretation incrementally online. The closely time-locked process for identifying relevant scene referents, and then altering comprehension depending on the foveated scene regions, crucially emphasizes the active

influence of scene information, such as depicted actions, on comprehension processes. In doing so, the model highlights the influence of the scene itself, going beyond the notion that gaze in the scene is simply a passive index of comprehension. One motivating factor for this, is evidence that scene information may even take priority over linguistic and world knowledge of typical events in driving anticipatory language-mediated attention (Knoeferle & Crocker, 2006; 2007). Even though the phases (i) and (ii) can be conceptually distinguished, they do not need to occur serially. Instead, the processes may overlap and occur in parallel during the comprehension process.

Importantly, the CIA has been instantiated in a computational model of situated language processing (Mayberry et al., 2009; Crocker et al, 2010). CIANet is based on a simple recurrent network (SRN; Elman, 1990) that produces a case-role interpretation of the input utterance. Processing in an SRN takes place incrementally, with each new input word interpreted in the context of the sentence processed so far, as represented by a copy of the hidden layer at the previous word/time-step. Additionally, CIANet incorporates visual input through an additional input representation of the scene, thus providing (optional) visual context for the input utterance.

The model exploits distributional information accrued during training to learn syntactic constraints such as constituent order and case marking, semantic constraints on likely role-fillers for particular verbs, as well as correlations between utterance meaning and the characters and events in the visual context. The integration of both kinds of knowledge – long-term experience and immediate visual context – contributes to interpretation and the non-deterministic anticipation of likely role-fillers in the manner outlined by the CIA. Attention to scene events is allocated via a gating vector, which is determined by the networks current interpretation of the unfolding utterance. This gating vector thus implements the shifts in visual attention that are

elicited by the utterance, and the increased importance of the attended scene region on subsequent interpretation (see Mayberry et al., 2009 for details).

Altmann and Mirkovic (2009) present a *Joint representation of linguistic meaning and visual information* account that differs somewhat from the CIA. According to Altmann and Mircovic, the anticipated linguistic meaning and visual scene information are not distinguishable from each other. Instead, these two modalities interact with each other and updating occurs on the joint representation of linguistic meaning and visual information. They note that typically in visual world studies, the visual environment or scene is available to people from the onset of the spoken stimuli or even prior to it. While visual information used in these studies often contains static representations of the environment, speech unfolds in time and becomes available to comprehenders only incrementally. Thus, it is possible that identification and representation of the visual information precede spoken language. This later speech input then activates features of visualized objects in the hearer's mind, including affordances (information of how the objects interact with other objects in the real world), resulting in early eye movements to relevant objects either in an integrative or predictive manner (Altmann, & Kamide, 2007, 2009). Their account proposes that the nature of the visual world is not visual per se, but rather an interwoven mental representation formed by coordinating visual perception with spoken language and which is updated on the basis of the unfolding spoken language (see also Allopenna et al., 1998; Tanenhaus et al., 2000; Smith et al, 2013).

About the importance of coordinating visual information and language. These models for relating situated comprehension processes to observed language-mediated gaze behavior, provide first linking hypotheses to explain how and why people come to coordinate the information

received from the two modalities. Recently, Huettig, Rommers, and Meyer (2011) raised the question whether it is always important or necessary to coordinate visual attention and linguistic attention. They suggested that maybe one reason to coordinate visual and linguistic stimuli is that often these two modalities provide complementary information and therefore their coordination is beneficial for comprehension. However, in natural everyday behavior and communication, people also face situations in which coordination of linguistic and visual input does not necessarily facilitate the task people are performing. This is a topic that deserves further investigation. As for now, we know relatively little about the conditions under which visual attention and language comprehension processes are not coordinated, i.e., situations in which the location of the eyes are not indicative of underlying language comprehension process.

One example in which the coordination of language comprehension processes and gaze is challenged was found by Pyykkönen-Klauck and Järvikivi (2015). As mentioned earlier, when the potential referents for an ambiguous pronoun were continuously kept in the visual scene, participants did select the same character they fixated as an antecedent for the pronoun. However, the visual availability of one of the referents was also modulated, such that when people heard the anaphoric pronoun "he" in the context *"…The rabbit strokes the fox near the river. He…*" either rabbit or fox had walked away from the visual scene. When the disappearing animal was "rabbit" in the above example, adults tend to look at fox after hearing the pronoun, as this was the only potential referent left on the screen. There were very few fixations to the empty place where the rabbit used to be. Nonetheless, judgments for the referent of the pronoun, indicated that listeners tended to select the first-mentioned subject (rabbit). This suggests that there are situations in which the overt visual attention and language comprehension are not necessarily correlated. The finding was different with children: When 4-year-old children were

exposed to the same task, they were likely to shift their final selection to "fox" when rabbit was moved from the visual scene; when both animals were kept constantly on the screen, they were likely to select rabbit much like the adults.

However, there is also evidence that adults are sensitive to visual manipulation: For example, another study that manipulated the presence of target objects articulated in the spoken language as well as their distractors also showed that adults prefer looking at objects that remain on the screen instead of shifting the attention to the location at which the object currently being articulated was located earlier before its removal (Burigo & Knoeferle, 2015). In addition, studies in which visual and linguistic information are in conflict – pushing the interpretation to different directions – emphasize the relative importance of cues and modalities guiding the comprehension processes. For example, Knoeferle and Crocker (2006) presented participants German Object-Verb-Subject (OVS) sentences such as (2) when they saw visual scenes depicting three characters, a pilot in the middle, with a detective on one side and a magician on the other side. The scene also showed atypical actions, such as a magician spying on a pilot (rather than a detective spying on the pilot).

(2)

Scene: Magician (spying-on) Pilot (serving-food) Detective

Den Piloten bespitzelt gleich der Detektiv.

'The pilot (OBJ) spies-on soon the Detective (SUBJ)'

When actions remained depicted during listening, participants robustly anticipated the magician after hearing the verb – the character that was performing the action – rather than the detective, which would be anticipated based on world knowledge. Indeed, the same pattern was observed,

although somewhat reduced, in two subsequent studies where either (a) the entire scene was removed during listening (blank screen paradigm), and (b) when the actions were initially animated, but then only characters (without actions) remained in the scene during listing (Knoeferle & Crocker, 2007). These results indicate that when the (recently) visible scene provides an explicit action cue regarding how to interpret and predict likely continuations, such cues can temporarily override expectations arising from people's general world knowledge.

Eye movements in the analyses

Up to this point, our discussion has relied heavily on generalizations regarding the measurement of eye movements used in visual world eye tracking studies. In order to understand both the potential and constraints of eye movement metrics and analyses, we turn our attention to the nature of eye-movement data in different tasks. As noted earlier, most visual world studies draw conclusions regarding the underlying cognitive processing mechanisms based on fixation-based metrics, and to a lesser extent, saccade-based metrics. What is typically measured is the number of inspections into some pre-defined areas of interest (AOIs) during specific time windows. This measurement is categorical: When the participant is looking at the target, she cannot fixate any other object simultaneously. Thus, the looks into the AOIs are typically calculated as "inspected, not inspected". This is done each time the eye tracker samples the participant's gaze coordinates; alternatively, down-sampling is used and the fixation information is calculated for larger time bins of e.g., 10 or 20 milliseconds. For statistical analyses these time windows are further aggregated into even larger time bins.

When the visual world paradigm first became widespread, the standard method for hypothesis testing was by using linear models of repeated measurements (ANOVA). The

solution was to transform the categorical looks into a continuous variable by calculating proportions of aggregated fixations over multiple trials and over time. Time, in turn, was turned into categorical by breaking the analyses into distinct time regions. Separate ANOVAs were calculated for each time bin. Recently, this method has been criticized because the transformations violated the following assumptions of ANOVAs: (i) continuously dependent variables, (ii) independence of observations, (iii) unbounded range of the continuous dependent variable and (iv) normal distribution of observations (see e.g., Barr, 2008; Jaeger, 2008).

In order to resolve these violations, several different approaches have been proposed. Arai et al. (2006), for example, proposed calculating log ratios (ln(P(a)/P(b))) to be analyzed with ANOVA. Knoeferle et al. (2005) applied hierarchical log-linear regression models that do not make the assumptions of independence and normality. Barr (2008) suggests a multilevel logistic regression model with log odds that can handle eye movements categorically and time continuously as well as differentiate between anticipatory and integrative/rate effects. Mirman, Dixon, & Magnuson (2008) advocate assessing change over time by using a growth curve model that allows polytomous dependent variables while comparing means across different conditions (averaged over participants or items). Recently, the scientific community has settled (for the most part) on the use the multilevel logistic regression when analyzing the time course of the effects. However, there is ongoing debate regarding the best practices for computing these models, e.g., inclusion of random effects structure (see a recent discussion in Barr, Levy, Scheepers, & Tily, 2013).

Challenging the analyses

Beyond such statistical considerations, different experimental decisions crucially influence the interpretation of the data and what we can expect to find in the statistical analyses in the first place. These decisions pertain to aspects such as the amount of visual preview prior to the speech onset, the point in time selected as an onset for the time course analyses and the duration of the time bins for which the analyses are carried out.

Visual preview. Huettig and McQueen (2007) studied timing of phonological, semantic and visual shape competitor effects and found that the preview time affected the likelihood of inspecting time differences across the competitor effects. With a 1000 millisecond preview, they found that looks to the phonological competitors preceded looks to the semantic or visual shape competitors and no timing difference with the latter two. However, when the preview was only 200 milliseconds (Experiment 2), an early phonological competitor effect was not found, and the visual shape competitor effects arose earlier than semantic competitor effects (see also Dahan & Tanenhaus, 2005, with 300 millisecond preview for visual shape competitors). These findings illustrate how varying the visual preview can have important consequences for the effects inspected in the visual world studies and should thus be carefully selected and motivated prior to the experiment.

Onsets for time course analyses. Visual world studies typically contain spoken stimuli that extend over several words, phrases, clauses and even sentences. Speech unfolds rapidly and fluently in time, rarely containing acoustic breaks between words, and sometimes even the boundaries between clauses and sentences can also be very minimal. While it has been established for reading studies, that people slow down their reading rate when they encounter difficulty (see a review, Rayner 2009), this is not possible in visual world studies that utilize pre-

recorded spoken stimuli. In these experiments, the participants have no control over the speed at which the content is presented to them, and thus in cases of difficult comprehension phases, it is unclear how the system adapts, whether it proceeds to the upcoming information and leaving previous material in some underspecified state, or continues processing the difficult information at the cost of the understanding the subsequent speech signal (Ferreira, Engelhard, & Jones, 2009). Thus, it is important to carefully select the linguistic material in the way that the onset of the critical word of interest would appear in a position that is assumed to be cognitively and structurally equal across conditions.

A related issue for selecting the onset for the time course analyses concerns the continuous nature of fixating objects in the visual environment. In rather simple visual scenes with only a few objects, it is likely that people will randomly inspect target or competitor objects in sentences that do not identify any particular object prior to the target word such as instructional tasks of "move/put the TARGET". In those studies, participants often have a 1000 millisecond preview as well as time of the sentence start "move the" to move their eyes freely prior they start processing the target word. Thus, when calculating the looks for the first time bin after the target onset, it is likely that a certain proportions of the looks are already at the target and no further eye movements are necessary. In such cases it is difficult to determine when exactly listeners have landed on that interpretation.

One practical solution is to remove all trials in which the participants are already fixating the objects modeled in the statistical analyses and analyze only those trial in which participants were not already fixating the objects at the onset of the time course analyses (e.g., Järvikivi et al., 2005, 2014; Pyykkönen et al., 2010). However, this may lead to relatively large data loss in studies that use only a few objects on the computer screen. An alternative (or a combined)

solution that has been implemented in pronoun resolution studies is to attract participants' eye movements away from potential targets just prior to the pronouns. This can be achieved by referring to a (depicted) location or situation just prior to the pronoun (see a review of these studies in Van Gompel and Järvikivi, this volume).

The length of time bins. In order to study how interpretation develops over time, the time course analyses are often of interest. As mentioned earlier, the continuous time is typically broken into separate time bins in the statistical analyses. However, there is no clear consensus how long these bins should be. While including lots of small bins may be desirable in order to explore the unfolding dynamics of comprehension, it also increases the number of statistical models in a manner that is generally not recommended. However, the challenge of making the bins very large is the dynamic nature of the eye movements: When the bins are very long and the data is averaged over lots of shorter samples, it is possible that some changes are hidden in averages.

Summary

In this chapter, we have consider three fundamental methodological aspects of the visual world paradigm: the nature of the task that people are engaged in during visually situated comprehension, the linking hypothesis that related underlying models of situated comprehension with eye-movement behavior, and challenges related to the nature of eye-movement data and statistical analyses. We have not attempted to offer a particularly detailed or comprehensive treatment of these topics. Instead, we want to make the reader sensitive to the importance of clearly identifying what people are doing, what mechanisms are thought to generate language-

mediated eye movements, and how data are analyzed and reported – both when evaluating findings in the literature, and designing and analyzing one's own experiments. These issues are taken up throughout the chapters of this book, with different researchers adopting varying assumptions and methods allowing each reader to pick up the techniques best suitable for their experimental questions.

References

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

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

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

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

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.

Altmann, G.T.M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33,* 583-609.

Andrews, T. J., and Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research, 39,* 2947–2953.

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

Arnold, J.E., Brown-Schmidt, S., & Trueswell, J.C. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes, 22,* 527-565.

Arnold, J.E. Eisenband, J.G., Brown-Schmidt, S., & Trueswell, J.C. (2000). The immediate use of gender information: Eyetracking evidence of the time-course of pronoun resolution. *Cognition, 76,* B13-B26.

Barr, D.J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59,* 457-474.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255-278.

Boland, J.E. (2005). Visual arguments. *Cognition, 95,* 237-274.

Burigo M. & Knoeferle, P. (2015). Visual Attention during Spatial Language Comprehension. *PLoS ONE 10*. Doi: 10.1371/journal.pone.0115758

Castelhano, M. S., and Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psycholology 62,* 1-14.

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

Colonna, S., Schimke, S., & Hemforth, B. (2014). Information structure and pronoun resolution in German and French: Evidence from the visual world paradigm. In B. Hemforth, B. Mertins, & C. Fabricius-Hansen (eds). *Psycholinguistic Approaches to Meaning and Understanding across Languages, pp 175-195.* Springer, Switzerland.

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

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

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

Ferreira, F., Engelhard, P.E., Jones, M.W. (2009). Good enough language processing: A satisficing approach. In N. Taatgen, H. Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 413-418.* Austin, TX, US.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188-194.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka, J. & van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, UK.

Huestegge, L. (2010). Effects of vowel length on gaze durations in silent and oral reading. *Journal of Eye Movement Research, 3,* 1-18.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57,* 460-482.

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.

Irwin, D. E. (1998). Lexical processing during saccadic eye movements. *Cognitive Psychology, 36,* 1-27.

Irwin, D. E., & Carlson-Radvansky, L. A. (1996). Cognitive suppression during saccadic eye movements. *Psychological Science, 7,* 83-88.

Jaeger, F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446.

Järvikivi, J., Pyykkönen-Klauck, P., Schimke, S., Colonna, S., & Hemforth, B. (2014). Information structure cues for 4-year olds and adults: Tracking eye movements to visually presented anaphoric referents. *Language, Cognition & Neuroscience, 29,* 877-892.

Järvikivi, J., Van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. *Psychological Science, 16,* 260-264.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass , 2*, 647-670.

Kamide, Y., Altmann, G.T.M., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.*f*

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

Knoeferle, P., & Crocker, M.W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye-movements. *Journal of Memory and Language, 57,* 519-542.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering, M.J. (2005). The influence of the immediate visual context on incremental thematic role assignment: Cross-linguistic evidence from German and English. Cognition, 95, 95-127.

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

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

Pyykkönen, P., Hyönä, J., & Van Gompel R.P.G. (2010). Activating gender stereotypes during online spoken language processing: Evidence from visual world eye tracking. *Experimental Psychology, 57,* 126-133.

Pyykkönen, P., & Järvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology, 57,* 5-16.

Pyykkönen-Klauck, P., & Järvikivi, J. (2015). The influence of visually changing environments on the representation of discourse referents: Comparing 4-year-old children and adults. *In revision.*

Pyykkönen, P., Matthews, D., & Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. *Language and Cognitive Processes, 25,* 115-129.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62,* 1457-1506.

Rayner, K., Li, X., Williams, C.C., Cave, K.R., & Well, A.D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research, 47,* 2714–2726.

Smith, A., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology, 4.*

Staudte, M. & Crocker, M.W. (2011). Investigating Joint Attention Mechanisms through Spoken Human-Robot Interaction. *Cognition, 120,* 268-291.

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.

Weber, A., Crocker, M.W., & Knoeferle, P. (2010). Conflicting constraints in resource adaptive language comprehension. In: Crocker & Siekmann (eds), Resource Adaptive Cognitive Processes, pp. 119-142, Springer Verlag, Berlin.

Weber, A. & Crocker, M.W. (2012). On the nature of semantic constraints on lexical access. *Journal of Psycholinguistic Research, 41*, 195-214.