Chapter 5: The Visual World Paradigm

Anne Pier Salverda[a]

Michael K. Tanenhaus[ab]

[a] Department of Brain and Cognitive Sciences

University of Rochester

[b] School of Psychology

Nanjing Normal University

Salverda, A.P., & Tanenhaus, M.K. (in press). The visual world paradigm. In A.M.B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics: A practical guide*. Malden, MA: Wiley-Blackwell.

Abstract

The visual world paradigm (VWP) is a family of experimental methods for studying real-time language processing in language comprehension and production that can be used with participants of all ages and most special populations. Participants' eye movements to objects in a visual workspace or pictures in a display are monitored as they listen to, or produce, spoken language that is about the contents of the visual world. Eye-movements in the VWP provide a sensitive, time-locked response measure that can be used to investigate a wide range of psycholinguistic questions on topics running the gamut from speech perception to interactive conversation in collaborative task-oriented dialogue.

Keywords: speech; spoken-word recognition; sentence processing; pragmatics; conversation; language production; eye movements; visual attention; eye-tracking; language-mediated eye movements.

Introduction

The visual world paradigm (VWP) is a family of experimental methods in which participants' eye movements to real objects in a visual workspace, or to pictures on a display are monitored as they listen to spoken language or produce language. Figure 1 shows an example of the experimental setup. The term, coined by Tanenhaus and colleagues (Allopenna, Magnuson, & Tanenhaus, 1998), emphasizes that the visual workspace defines a circumscribed context that the language is *about*.

Insert Figure 1 about here

In 1974, in a remarkable article titled "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", Roger Cooper reported experiments that used a Dual-Purkinje eye-tracker to measure participants' eye movements as they listened to stories while looking at a display of pictures. Participants initiated saccades to pictures that were named in the stories and to pictures associated with those names. Fixations were often generated before the spoken word ended, suggesting a tight coupling of visual and linguistic processing.

More than 20 years later, Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) used a head-mounted video-based eye-tracker to monitor participants' eye movements as they followed experimenter-generated spoken instructions to pick up and move objects arranged on a table (e.g., *Put the apple that is on the towel in the box)*. Their task, or action-based, approach was influenced by pioneering work at Rochester that used eye movements to study vision in natural tasks (see Hayhoe & Ballard, 2005, for a review). Tanenhaus et al. found evidence for rapid integration of visual and linguistic information

in word recognition, reference resolution, and syntactic processing (parsing). The latter was the focus of their report. Allopenna, Magnuson and Tanenhaus (1998) is the first VW study to use a screen-based presentation to study the time-course of spoken-word recognition in continuous speech. Trueswell, Sekerina, Hill and Logrip (1999) demonstrated that the VWP could be used to study sentence comprehension in pre-literate children, using a variant of the set-up in Tanenhaus et al.

Many current VW studies follow the methods and rationale introduced by Cooper (1974), who did not use an explicit task. Altmann and Kamide (1999) is the foundational "look-and-listen" study. They presented displays with clipart of a person (e.g., a boy) and a set of four objects (e.g., a cake, a toy car, a ball, and a toy train) and a spoken utterance, e.g., *The boy will eat the cake* (see Figure 2). Participants were more likely to generate a saccade, that is, to make an "anticipatory eye-movement" to the target object, as the verb unfolded when the semantics of the verb were consistent with only one of the objects (*eat*; only the cake is eatable, as opposed to *move*).

Insert Figure 2 about here

Two seminal studies provided the foundation for using the VWP in language production. Meyer, Sleiderink and Levelt (1998) demonstrated that eye movements are closely time-locked to utterance planning during the production of simple noun phrases. Griffin and Bock (2000) monitored eye movements with schematic scenes that could be described using active or passive constructions (e.g., a picture of lightening striking a house) and demonstrated a tight coupling between scene apprehension and utterance planning.

Assumptions, Logic, and Terminology

All VW experiments use similar logic and variations of the same design. A visual workspace contains real objects or a display depicts an array of objects, a schematic scene, or a real-world scene. With screen displays, pictures are typically used, but some studies use printed words instead (McQueen & Viebahn, 2007). Participants' eye movements are monitored as speech unfolds. Of interest is at what point in time with respect to some acoustic landmark in the speech signal (e.g., the onset of a word) a shift in the participant's visual attention occurs, as measured by a saccadic eye movement to an object or picture.

Behavioral and neuroimaging measures require a *linking hypothesis* that maps the dependent measure, in this case eye movements, onto hypothesized underlying processes. The most general form of the VW linking hypothesis is that as visual attention shifts to an object in the workspace, as a consequence of planning or comprehending an utterance, there is a high probability that a saccadic eye movement will rapidly follow to bring the attended area into foveal vision. Where a participant is looking, and in particular when and to where saccadic eye movements are launched in relationship to the speech, can provide insights into real-time language processing. We return later to considerations about how linking hypotheses affect the interpretation and analysis of VW studies.

Across studies, the characteristics of the language, the contents and structure of the visual workspace, and the instructions and/or task vary. For this discussion, we assume that the potential referents are pictures displayed on a screen. Each picture may be referred to one or more times as the spoken language unfolds. The picture of interest, at a particular point in time, is the *target*. Experimenters are primarily interested in when

looks to the target diverge from looks to the other pictures. The properties of one or more of the non-target pictures are often manipulated such that they are more related to the target than the other non-target pictures along some specified dimension, which could include participation in an implied event. Those pictures are then typically referred to as the *competitors* and the unrelated pictures as *distractors*. Competitors are referred to by the dimension(s) along which they differ from the target. For example, if the names of two of the pictures begin with the same syllable, e.g., *candle* and *candy*, and the participant hears the instruction, *Click on the candle,* then the candle would be the *target* and the candy would be the *phonological competitor* (or alternatively, the *cohort competitor*). Competitors can differ along any number of dimensions, ranging from how their names differ from the target (e.g., cohort, rhyme, or voice-onset time (VOT) competitors) to how similar they are along visual and/or conceptual dimensions. For example, two depicted objects of the same type might differ along a dimension such as size, color, or a feature such as having stripes or stars. In comprehension studies, the point in the speech signal when only one picture is consistent with the integration of information in the sentence and the properties of the objects in the visual world is sometimes referred to as the *Point of Disambiguation* (POD). The POD can serve as a reference point, defining the earliest point in the speech signal where a participant could identify the target if he or she was using all of the information available. However, POD is also sometimes used to refer to the point in time where looks to the target actually begin to differ from looks to competitors. The competitor terminology is not typically used in production studies, but the logic is similar, with researchers examining the relationship between looks to a region of interest (e.g., a potential agent or patient) and

aspects of the utterance, e.g., when a picture is mentioned, and in what grammatical or thematic role (e.g. subject or object and agent or patient, respectively).

Apparatus

The biggest decision one faces when setting up a lab is what type of eye-tracker to choose. Here we describe the two most commonly used systems.

In determining which system is most suitable for a given type of experimental paradigm or experiment, factors to be taken into account include: properties of the experiment (the nature of the task, e.g. the form of interaction with the visual world); requirements for temporal and spatial sensitivity (an eye-tracker with a high temporal sampling frequency may be desired when subtle differences in the timing of effects are of interest, while a system with low spatial resolution may be used when the number of regions of interest in the display is small, and these regions are spatially distinct); the population(s) that will be tested; whether automatic coding of the data is desired; and affordability.

The simplest, least expensive, and most portable system is a video camera, which records an image of the participant's eyes. The camera can be mounted above or below a computer screen, or positioned in the center of a platform with real objects (Snedeker & Trueswell, 2004). Eye movements are coded through frame-by-frame examination of the video recording. Temporal resolution is limited by the video equipment, which usually records at 30 or 60 Hz. The objects in the visual display need to be located such that fixations to each of the objects result in clearly distinct images of the eye. An important limitation is that participants are required to keep their eyes positioned in front of the camera.

Many eye-tracking systems use optical sensors to infer gaze location by measuring the orientation of the eye in its orbit. An image of one or both eyes is recorded by one or two eye cameras, which are either head-mounted or remote. The image is processed by dedicated hardware and gaze location is established on the basis of the image of the pupil, or by computing the vector between the center of the dark pupil and the corneal reflection. The latter is obtained by exposing the eyes to invisible near-infrared light originating from an illuminator. Importantly, gaze location is contingent on both eye orientation and the orientation of the head relative to the visual display. Most optical systems compensate for head movements (e.g. remote systems track the shape of a small sticker attached to the participant's forehead to record head position and orientation).

Optical eye trackers typically generate output in the form of a stream of XY coordinates reflecting the participant's gaze location. If this output is in the form of screen coordinates, coding of eye movements to regions of interest in the visual world can be automatized. Some optical systems use an additional scene camera and produce video output in which the participant's gaze location is superimposed on a video recording of the visual workspace. Head-mounted systems typically operate with a higher sampling rate and spatial resolution than remote eye trackers. However, temporal and spatial resolution can be improved by using some form of head stabilization, e.g. a chin rest.

Common Variations across Experiments

Language

The language can differ along any number of dimensions, from manipulations of fine-grained acoustic-phonetic features (duration, VOT, formant structure, fundamental frequency, etc.) to properties of words (syntactic category, semantic features, frequency of occurrence, etc.) to linguistic structure (syntactic structure, information structure, semantic and pragmatic properties such as implicating and questioning, etc.).

The source of the speech is important. The language often comes from a disembodied voice, which provides a narrative (e.g., *The doctor will hand the scalpel to the nurse*) or an instruction (e.g., *Put the large candle above the fork*). The default assumption is that the speaker and the listener have access to the same information in the visual world. In more interactive tasks, naïve participants and/or confederates generate the utterances of interest.

Visual world

The characteristics of the workspace play an important role in determining the questions that can be asked in a VW experiment. The most frequently used set up is a screen display depicting an array of pictures, a schematic scene, or a real-world scene. The workspace can also contain real objects arranged on a tabletop or a more complex apparatus. When real-world objects are used in conjunction with instructions to manipulate them, one can ask research questions such as how affordances of objects interact with the language, which might be less natural with screen displays. These questions could be asked in a more controlled environment by using virtual reality, which

would allow for a wide range of interesting manipulations, including sophisticated saccade-contingent changes to the virtual environment.

More complex workspaces are useful for asking questions about perspective taking and for generating a variety of utterance types. For example, control of what information is *shared* and what information is *privileged* between participants can be achieved by constructing an appropriate physical apparatus, e.g., one with cubbyholes that are open or occluded such that only one interlocutor can see one or more of the objects.

Task

There are two common variants of VW experiments. *Task or action-based* studies borrow from the vision-in-natural-tasks literature. Participants interact with real-world objects or, more typically, interact with pictures in a screen-based workspace to perform a motor task, typically clicking and dragging pictures to follow explicit instructions (*Put the clown above the star*), clicking on a picture when its name is mentioned, or manipulating real objects (e.g., *Pick up the apple. Now put it in the box*). Explicit goal-directed motor tasks encourage the participant to rapidly identify and fixate the target object of the linguistic expression. Participants typically generate a saccade to the referent (or maintain an earlier fixation), and keep fixating it until the mouse or hand approaches the goal (visually-guided reaching). The choice indicates the final interpretation, which can be used for response-contingent analyses (e.g., analyzing trials with looks to the voiced competitor *beach* when the participant choses the voiceless target *peach* upon hearing a token with a particular VOT). The earliest language-mediated fixations occur 200-250 ms after the relevant acoustic landmark that could establish a POD (Salverda, Kleinschmidt, & Tanenhaus, 2014). Throughout a trial, a high proportion of the fixations are controlled

by the goal, including fixations to objects that are relevant to establishing reference as the language unfolds (Salverda, Brown, & Tanenhaus, 2011; for a discussion of an alternative, activation-based hypothesis, see Altmann & Kamide, 2007).

Look-and-listen studies (sometimes misleadingly called *passive listening studies*) do not require participants to perform an explicit task other than to look at the computer screen. Because the interpretation of the language is co-determined by information in the scene, participants' attention is drawn to referents, including pictures that the listener anticipates will be mentioned or pictures associated with implied events (e.g., an action that will take place in the future). In a variation introduced by Altmann (2004), a blank screen replaces the schematic scene at some point in the narrative.

There is a paucity of work that directly compares "task-based" and "look-and-listen" studies that are designed to address the same question, which makes claims about the strengths and weaknesses of each approach somewhat speculative.

General Considerations Affecting Design and Interpretation

Many first-time users want to know what steps to follow to design and analyze VW experiments. We find that an analogy to cooking is helpful. Everyone cooks to some degree, but expertise varies. Some people rarely cook and know almost nothing about cooking techniques. If you are one of those, you can feed yourself, but you cannot create anything new. And if you get adventurous and try a recipe, it's unlikely to turn out well; even the most detailed recipe requires knowledge of some basic cooking techniques. In contrast, master chefs have expertise with preparing a wide range of dishes in multiple genres of cooking; they are also aware of the molecular processes involved in cooking and the latest technology. Whereas master chefs rarely make mistakes when preparing

established dishes, their novel creations are not always successful. However, when a dish fails, they have good intuitions about what went wrong and how to correct it. One need not be a master chef to use the VW paradigm. But being the equivalent of someone who rarely cooks and occasionally tries to follow a recipe is likely to be problematic.

Every VW experiment combines aspects of both spoken language and vision. Successful use of the paradigm therefore requires some basic knowledge about, and sensitivity to, properties of both systems. This is challenging because few psycholinguists are knowledgeable about vision. Moreover, many psycholinguists who study higher-level processes (e.g., syntactic processing, interpretation, inference and implicature) have limited experience with the speech signal. Conversely, many who are knowledgeable about the speech signal have only a cursory knowledge of how it is impacted by higher-level factors. In what follows, we present some of the factors in speech and in vision in natural tasks that strongly impact the design, analysis, and interpretation of VW studies.

Speech and spoken language

Speech is a temporal, rapidly changing signal. Acoustic cues are transient, and there are no acoustic signatures that correspond to linguistic categories. Relevant cues to a category, or even a phonetic feature such as voicing, are determined by multiple cues, many of which arrive asynchronously and are impacted by both high and low level linguistic subsystems. Linking eye movements to relevant linguistic information in the speech signal is therefore critically dependent on having some understanding of where, when, and why information in the speech signal provides information about linguistic structure.

Time-locking eye movements to an acoustic landmark typically requires determining the onset of a speech sound or spoken word. This task is straightforward when a target word is presented in isolation, for instance, the word *beaker* starts with the release of the plosive /b/. However, most studies use spoken sentences where the target word is embedded in continuous speech, for instance, *Click on the beaker.* Words in continuous speech can have very different characteristics than words spoken alone. Determining when a target word starts in continuous speech can be complicated and we therefore recommend consulting with a phonetician. For example, in *Click on the beaker*, the release of the plosive /b/ does *not* correspond to the onset of *beaker*. The closure preceding the release is an integral part of the articulation of plosives in continuous speech, and the onset of the closure therefore constitutes the onset of *beaker*.

Coarticulation, the temporal and spatial overlap in the articulation of two or more speech sounds, is a ubiquitous property of speech. At any moment in time, then, the speech signal provides information about multiple speech sounds, though the strength of coarticulation varies depending on many factors. This has consequences for the time-locking between speech and eye movements, especially under conditions where it is essential to estimate the earliest information in the speech signal that might influence a language-mediated eye movement. Careful examination with a speech editor (using a spectrogram) or evaluation of the stimuli using incremental auditory presentation can improve the quality of the segmentation of a linguistic event (such as a speech sound). The influence of coarticulation can be reduced by using cross-spliced materials when possible and otherwise by carefully choosing the stimuli.

Speech is determined by constraints at multiple levels. The same acoustic cues that provide information about phonemic segments may also generate expectations about syntax, information structure, and pragmatics. Many aspects of these higher-level processes are manifested by prosody and intonation, which affect acoustic cues (such as duration) that are also used in processing phonemes and spoken words. Thus, higher-level information may be available earlier than one might otherwise think. Therefore, it is important to consider the locus and extent of various cues to aspects of linguistic structure in the speech tokens used in a VW study. Moreover, manipulation of speech cues may impact interpretation at multiple, and perhaps mutually constraining, levels of linguistic representation.

Eye movements in natural tasks

While the classic literature on visual search with simple displays, and more recently, scenes, is informative for VW researchers, a newer literature on vision in natural tasks is arguably more relevant (Salverda, Brown, & Tanenhaus, 2011). Traditional visual-search studies focused on the role of low-level perceptual features (e.g., color, orientation, and shape) in pre-attentive visual processing and in the subsequent allocation of visual attention. These studies used simple, static, and largely unstructured displays, on the assumption that these elementary perceptual features would have similar effects on visual attention in complex real-life scenes. Given this assumption, basic stimulus features should be key predictors of the deployment of visual attention. Indeed, in the absence of a task, global estimates of visual salience derived by integrating multiple feature values at each location within a screen correlate with gaze patterns during viewing of a scene (Parkhurst, Law, & Niebur, 2002).

Feature-based salience, however, is a poor predictor of gaze patterns when a participant is engaged in a well-defined task (Tatler, Hayhoe, Land, & Ballard, 2011). In studies of everyday visuomotor behaviors, such as preparing tea, making sandwiches, and driving, the vast majority of fixations, typically 90% or more, can clearly be attributed to task-based goals. Participants have a strong tendency to fixate objects immediately before they become relevant to the execution of a task subgoal (e.g. fixating an object immediately prior to reaching for it). Moreover, participants direct their fixations to those parts of an object that are behaviorally most relevant (e.g., the spout of a tea kettle during the pouring of hot water).

In addition to influencing the location and timing of fixations, cognitive goals play a key role in determining the information encoded during fixations and the retrieval, during a fixation, of information that is stored in memory. Importantly, aspects of the task that a participant performs, including those that change dynamically, can strongly influence the time and resources available for accessing information, and thus the information that is encoded during a fixation. For instance, as task complexity increases in a block-sorting task, participants begin to rely less on working memory and more on the external environment (Droll & Hayhoe, 2007).

The most general implication for VW studies is that where and when participants will look will be strongly determined by both explicit and implicit task goals. For example, one might be interested in using the proportion of looks to a previously mentioned picture as an indication that it is being considered as a potential referent for a referring expression. However, a participant who already knows the location and the properties of that object might not look at the picture even though it is being considered as a possible

referent—even if the picture is interpreted as the most likely referent of a referring expression (Yee & Heller, 2012). This does not mean that the VW paradigm is poorly suited to studying pronoun resolution; indeed, some of the most elegant and influential VW studies have done so. But it does mean that one has to be careful about interpreting the absence of looks to an object or picture. More generally this highlights the importance of not confusing your dependent measure with an underlying process. While this might seem obvious, it commonly occurs, especially when one assumes that there are "signature" data patterns that are diagnostic of a particular cognitive process (Tanenhaus, 2004). Finally, in the absence of a specific goal structure, it can be problematic to "back engineer" explanations based on fixation patterns.

Nature of Stimuli

Visual world

Each trial in a VW study begins with the presentation of a display that includes the target and typically one or more competitors (see Figure 1). Unrelated distractors provide a baseline for the assessment of speech-driven effects in the eye movements, which are revealed by differences in fixations to the target, competitor, and distractors. In order to avoid baseline differences that complicate interpretation and increase noise in the data, distractor objects should not have any direct or indirect relationship to the relevant information that might be activated (even temporarily) by the linguistic stimulus along phonological, semantic, and visual dimensions. Distractors with visual properties that might attract the participant's attention irrespective of the language should also be avoided.

The structure of the visual world varies across experiments, from a grid with objects to less structured visual scenes and workspaces. To facilitate coding of eye movements, objects should be situated some distance from each other. Systematic patterns in exploratory fixations (e.g., the tendency to fixate the top left picture in a search array early in a trial; Dahan, Tanenhaus, & Salverda, 2007) can be counteracted by randomizing or counterbalancing object positions. (Note that this is less of a concern when the experimental design employs a within-item manipulation of the linguistic input.) Unless there are other compelling reasons, we recommend against instructing participants to fixate a specific location at the start of a trial (e.g. by using a fixation cross). Maintaining fixation is resource-intensive. Moreover, asking participants to control their initial fixation can reduce the number of eye movements, with some participants maintaining fixation until before they begin to perform an action.

In production studies the characteristics of the display are often manipulated to examine how fixations to different objects affect lexical choice and grammatical encoding. Participants' attention is sometimes manipulated by a transient visual stimulus in a specified location.

Some studies use a preview phase, where objects are presented one at a time along with their intended name. Such familiarization is useful when constraints on item selection result in pictures that may not be readily associated with the intended name.

Linguistic stimuli

On each trial, a spoken instruction or sentence refers to one or more objects in the visual world. Utterances are designed such that there are clear predictions about how the combination of visual and linguistic information would yield different patterns of

fixations as the language unfolds, given a particular set of hypotheses. The time course of information integration can be examined in carefully chosen designs that use minimal differences in the timing and/or availability of linguistic information between experimental conditions (see the Example Study section.)

Timing

Comprehension studies typically use pre-recorded speech that is segmented and labeled with a speech editor. Time codes corresponding to the onset and offset of acoustic landmarks (e.g. onset/offset of the target word) are provided to the experiment software, so that eye-movement data can be aligned relative to particular linguistic material. Appropriate segmentation of the speech stimuli has direct consequences for the interpretation of eye movements during the unfolding of the linguistic stimulus (see also the section General Considerations Affecting Design and Interpretation). Systematic language-mediated fixations earlier than 200 ms after an acoustic landmark are likely due to biasing coarticulatory information before the marked event (Salverda, Kleinschmidt, & Tanenhaus, 2014; see also the section Nature of Stimuli). In production studies, the experimenter typically records the participant's utterances and then uses speech editing software to identify landmarks that are time-locked to the onset of the display or to looks to a particular location on the screen.

In most VW studies, the presentation of the linguistic stimulus follows the display with a brief delay of about a second, to allow participants to identify the objects in the display without giving them much opportunity to engage in strategic behavior. The complexity of the display is a factor in determining the appropriate duration of preview.

Data Collection and Analysis

The primary VW eye-movement data are a stream of gaze locations recorded at the sampling rate of the eye-tracker. These data are superimposed on a video recording of the visual world and/or stored in a digital file as XY coordinates. The latter type of output includes time-stamped messages that provide essential information about the trial, including the identity and position of the objects and the timing of acoustic landmarks in the speech stream (e.g. target word onset/offset). A digital sequence of XY coordinates can be parsed into a sequence of fixations, saccades, and blinks using dedicated software.

Coding

In order to assess what the participant was looking at throughout a trial, the experimenter defines regions of interest (ROIs) in the visual world, each of which is associated with one or more objects. We recommend extending regions of interest beyond the edges of objects (e.g., to the cell of a grid within which a picture appears) because visual attention is focused on a region, not a point, in space, and because gaze location as estimated by the eye-tracker is subject to error. A coder or automated coding procedure then scores each fixation as directed at one of the ROIs, or as not directed at any ROI. Saccades can be scored too, even though the visual system receives minimal input during a saccade—a phenomenon known as saccadic suppression. Because a saccade is triggered by a shift in visual-spatial attention to a new location, that location can be considered the locus of attention during a saccade. Similarly, a sequence of saccades and fixations to one ROI can be scored as one long fixation to that region, and blinks can be scored as continuing fixations if the same object is fixated prior to and following the blink. Eye movements can be scored until the end of the trial or until the point in time when the participant

performs an action indicating that they arrived at a definitive interpretation of the spoken input (e.g., the moment that a participant clicks on the target object, or the onset of the preceding mouse movement).

Visualization

A widely used method for summarizing results of VW studies plots the proportion of fixations to different objects throughout a trial (see Figure 3; see the Example Study section for another illustration). A proportion-of-fixations plot represents, at each moment in time throughout a time window, the proportion of trials with a look to each type of picture, averaged across participants (or items). Over the course of a trial, fixation proportions change in response to the processing of linguistic information and the integration of this information with information in the visual world. For instance, a rise in fixation proportions to an object reflects increased evidence for a particular linguistic interpretation associated with that object.

Insert Figure 3 about here

Proportion-of-fixation plots are useful because they provide a comprehensive (though by no means exhaustive) representation of the eye-movement record. Changes in the distribution of gaze to different types of pictures in the display over time reveal important aspects of the eye-movement data. They are also useful for some first-pass checks: Are objects fixated to the degree expected? Are only a small proportion of looks not directed at any of the ROIs? Do looks converge on the target picture? Are there baseline differences in fixation proportions? More generally, if the results of statistical analyses are inconsistent with what can be seen in proportion-of-fixation plots then something has gone awry. Moreover, as discussed below, it is inappropriate to first look at proportion-

of-fixation plots and then define an analysis region based on where one sees the biggest effects.

Proportion-of-fixation plots are constructed by taking a specific time window and computing, for each moment in time (limited by the sampling rate), the proportion of all relevant trials on which each of the objects is fixated. Figure 3 presents data from one participant in Experiment 1 of a study by Salverda, Kleinschmidt and Tanenhaus (2104) where the participant saw a display with a target picture and three distractors and followed a simple spoken instruction to click on the target. Figure 3a presents, for each trial, looks to the target during a time interval of one sec beginning at target-word onset. Proportion-of-fixations to the target are presented in Figure 3b. For instance, at 200 ms, the target was fixated on 7 out of 29 trials, resulting in a fixation proportion of 7/29 = 0.24. After the data have been aggregated across participants, it can be useful for purposes of data inspection or presentation to bin fixation proportions (e.g., using 20-ms bins for data recorded at 250 Hz; see Figure 4 in the Example Study section for an example). Such "down-sampling" reduces the influence of incidental moment-by-moment variation in the proportion of fixations observed.

Proportion-of-fixations plots usually present data aligned to a relevant linguistic event, which typically requires temporal realignment of the data across trials. For instance, in Figure 3, zero ms corresponds to wherever the target word started for each of the trials. For the evaluation of data in proportion-of-fixations plots it is important to take into account that information in the speech signal influences eye-movements with a delay of approximately 200-250 ms (Salverda et al., 2014).

An important issue arises when the amount of eye-movement data in a time window of interest varies across trials. For instance, if a participant's response terminates the trial, there is no eye-movement data from that moment onwards. When fixation proportions are computed for such data, early fixation proportions reflect data from all trials, whereas later fixation proportions reflect only the subset of trials on which the participant has not made or initiated a response. A frequently used solution is to extend the final fixation of each trial as an ongoing look in accordance with the participant's response, for example, a look to the picture that was selected. The rationale is that this "artificial" look reflects the participant's final interpretation of the speech signal. Extending the final fixation ensures that each trial contributes the same amount of information to the statistical analysis of fixation proportions across time.

Statistical analyses

VW eye-movement data can be analyzed with a range of statistical analyses on dependent measures that provide information about the speed and ease of target identification and the degree to which the participant considers competing interpretations. The most basic types of analyses examine the timing or occurrence of saccades to the target and competitor(s), such as the time it takes to generate a saccade to the target (on trials on which it was not already fixated), or the likelihood of making a saccade to the target or competitor during a time window. Analyses of mean fixation proportions across time windows can yield a more focused and detailed measure of the degree to which a picture is looked at over a temporal region. (Note that fixation proportions are bounded between 0 and 1 and thus violate data distribution assumptions of many statistical tests and models. In such cases, an appropriate data transformation, such as log odds or empirical

log odds, is required; see Barr, 2008, and Jaeger, 2008.) An important limitation of mean fixation proportions is that they do not capture trends in changes in fixation proportions across the window for which they are computed. Some analysis methods model the proportion-of-fixations curves directly (e.g. growth-curve analysis, Mirman, Dixon, & Magnuson, 2008, and Mirman, 2014; generalized additive mixed models, Nixon et al., 2016; bootstrapped difference of timeseries, Oleson, Cavanaugh, McMurray, & Brown, in press). Vandeberg, Bouwmeester, Bocanegra, and Zwaan (2013) introduced a different type of analysis, which predicts the likelihood of eye-movement transitions from one type of picture to another as a function of time.

In most studies, researchers are interested in eye movements in response to the presentation of relevant linguistic information in the speech stream, which translate to temporal windows that are time-locked to particular linguistic events (e.g., a window that captures eye movements during the presentation of the target word). For example, if one is interested in looks that could be triggered by "put" in *Put the large apple* before effects of "large", then the region might be the onset of "put" plus 200 ms to the onset of "large" plus 200 ms. If there was a theoretical reason to focus on the region before "apple", then the region that began with the onset of "put" would end 200 ms after the onset of "apple". Note that these regions must be calculated for each item.

Researchers often want to compare two or more conditions over an extended time interval, starting with the onset of a word. Here one can use any size window. However the choice of window size should be motivated and chosen before analysis. Any change in window size should be acknowledged as being a post-hoc choice and the windows that did not show significant effects should be reported. Selectively reporting statistically

significant results for post-hoc time windows is a form of "*p*-hacking" (cherry-picking the analyses you report to obtain a statistically significant result), which sharply increases the odds that results will not replicate. Perhaps the most dangerous form of *p*-hacking arises when one first inspects a proportion-of-fixations plot and then chooses the most promising windows.

If there are more looks to a related object (the target or competitor) relative to an unrelated object, this suggests that the listener perceived evidence for the linguistic information uniquely associated with the related object. In production studies, looks are taken as evidence that the participant attended to, and therefore likely encoded, that object. When contrasting looks to multiple objects within the same display, it may be necessary to compute a single measure in the form of a ratio for some types of statistical analyses which require independent measures. For example, the following ratio evaluates if the mean proportion of fixations to the competitor is higher than that to a distractor (in which case the result is larger than .5):

$$\frac{proportion\ of\ fixations\ to\ competitor}{proportion\ of\ fixations\ to\ competitor\ +\ distractor}$$

Variation in the degree of evidence in favor of a particular linguistic interpretation as a function of experimental condition can be assessed by comparing looks to the same target or competitor object across conditions. For instance, in the Example Study section, we discuss a VW study by Dahan and Tanenhaus (2004), who predicted (and found) a statistically significant difference in cohort competition between two experimental conditions.

It is important to note that current analyses do not map onto a generative model of the primary data that are evaluated in VW studies, which come from saccadic eye movements to real or depicted objects. These saccades are *events* and they are state-dependent. At the very least, where to and when a saccade is executed is affected by the spatial relationship among objects (e.g., distance and what trajectory, e.g. vertical, horizontal, or oblique, is required to shift gaze to a new location). However, current methods analyze where people are looking and not the events that underlie looks. We believe that advances in the analysis of VW data will come from the application of generative statistical models that predict events at the trial level, as a function of linguistic input, time, and the eye-movement record up to that point in time (i.e., the sequence of saccades, fixations, and their duration). While no such analyses currently exist, if and when they are developed, common practice may change.

Example Study

In this section we discuss an experiment that combines aspects of sentence processing and word recognition. Dahan and Tanenhaus (2004) conducted a VW study in Dutch to examine the effect of verb-based semantic constraints on lexical competition. Listeners heard spoken sentences that mentioned one of four depicted objects (the target) in the context of a semantic constraint that was introduced either before or after the target word. Their task was to click on the target object. Dahan and Tanenhaus took advantage of the fact that in Dutch, a verb can precede or follow its subject. When the verb precedes the noun, as in *Nog nooit klom een bok zo hoog* (Never before climbed a goat so high), it creates a constraining context that is consistent with the target *bok* (goat) but inconsistent with the cohort competitor *bot* (bone). When the verb follows the noun, *Nog nooit is een*

bok zo hoog geklommen (Never before has a goat climbed so high), the context preceding the target noun is neutral with respect to the target and the cohort competitor. (For ease of exposition we will use the English target "goat" and substitute the word "goal" as a cohort competitor, because the English words "goat" and "bone" do not overlap at onset.)

The experimental manipulation involved a repeated-measures design, in which each participant was exposed to multiple trials in each experimental condition. Issues that could arise from repeated presentation of pictures or target words, in particular across conditions, were avoided by presenting each item once and splitting the items across experimental conditions. For each participant, each item occurred in only one of the experimental conditions (neutral verb or constraining verb), and the assignment of items to conditions was counterbalanced across participants. Filler trials were designed to counteract contingencies in the experimental trials and included sentences with a verb that was semantically consistent with two of the pictures in the display (e.g. melt; icecream/butter). In a subset of the fillers, the two distractors were phonologically similar, to discourage participants from developing the expectation that pictures with phonologically similar names were likely targets. The order of trials was randomized. (Note that with some setups, it can be helpful to have practice trials at the start of the experiment to familiarize the participant with the experimental task and procedure.)

Figure 1 (shown at the beginning of this chapter) presents an example of a visual display including a target (goat), a cohort competitor (goal), an unrelated distractor (mirror), and a semantic competitor (spider). The latter was included to provide a baseline to separate effects of processing the target from effects that are due only to the verb. Figure 4 presents the proportion of fixations to the target, cohort competitor, and

distractor. In the neutral-verb condition, competitor fixation proportions increased from about 100 to 400 ms after the onset of the target word, and then dropped until they merged with distractor fixations. (The early looks might reflect coarticulation and/or information from the preceding verb.) This suggests that the cohort competitor was temporarily considered for recognition during the presentation of the target word. In the constraining-verb condition, a strikingly different pattern was obtained: Competitor fixation proportions did not increase significantly above their baseline level. This suggests that listeners made immediate use of verb-semantic constraints made available by the verb *climb* to eliminate the cohort competitor *goal* from the set of candidate words upon hearing the target word *goat*.

Insert Figure 4 about here

Advantages and Common Applications

Unlike other on-line psycholinguistic paradigms, the VWP is intrinsically *referential*: Language-mediated eye-movements to objects and locations in the visual workspace occur because processing the language makes the object or region of the workspace potentially relevant.

A particular advantage of the VWP is its *versatility*. The VWP can be used in a wide-range of natural (goal-based) tasks, with minimal restrictions. It can be used with a wide range of populations, including infants (using a variant of the preferential looking paradigm, see Chapter 2), elderly adults, bilinguals, and patients (e.g., aphasics). It has proved particularly useful in studying sentence processing in pre-literate children. It can also be used to study most topics in language comprehension (and to a lesser extent,

language production) at multiple levels, ranging from phonetic to pragmatic processing. We briefly outline some of the most common applications.

The VWP is widely used as a real-time measure in speech perception and spoken word recognition in continuous speech because it is extremely *sensitive* to fine-grained manipulations of the speech signal, including small variations in sub-phonemic acoustic/phonetic variables, for example 5 ms within-category differences in VOT (McMurray, Tanenhaus, & Aslin, 2002). We note that, while they are related, *sensitivity* and *sampling rate* are not equivalent. A dependent measure can have a high sampling rate, yet not be sensitive to a within-category 5 ms manipulation in VOT.

The VWP is used to study a wide spectrum of questions in sentence processing at multiple linguistic levels. In comprehension it is widely used in investigations of prosody and intonation, parsing, reference and discourse, and issues in experimental semantics and pragmatics. It is also well suited for studying the interaction of constraints across different linguistic levels, including asynchronous information. In language production, the VWP has been used to study lexical planning, grammatical encoding, and the interface between message planning, message updating, and utterance formulation.

The VWP is frequently used to study interactive task-based dialogue in conjunction with goal-based tasks such as the Edinburgh MAP task and *targeted language games*—a term introduced by Brown-Schmidt and Tanenhaus (2008). The MAP task is a collaborative task in which speakers sit opposite one another, with each having their own map. The instructor, who has a route, directs the follower to reproduce the route. Targeted language games are a type of interactive referential communication task constructed so that the conditions that one might design as experimental trials in a

factorial experiment emerge spontaneously and with sufficient frequency to conduct informative analyses.

Disadvantages, Limitations, and Concerns

There are some intrinsic limitations to the VWP in both the form and types of questions that can naturally be asked with VW designs, and in the types of inferences that can be drawn from VW data. Some of these limitations are obvious and have to do with domains of applicability and inquiry. For example, the VWP cannot be used for the study of (a) language that is not at least partially about the visual world; (b) language that is about events and entities that cannot easily be depicted, and (c) written language. Other limitations are more nuanced.

Many questions in sentence processing focus on "processing difficulty". Because the VWP is a referential task, there is no transparent mapping between the time to fixate a potential referent and a theoretical construct hypothesized to underlie processing difficulty. For example, an experimenter could manipulate "surprisal" and see whether it affects the likelihood of fixation to a mentioned target, the duration of fixations, and the time from an acoustic landmark (e.g., word onset) to when a saccade is launched. However, there is no clear linking hypothesis that would map surprisal onto any of these measures.

VW studies can be used to address questions about when different types of information are used and integrated. However, one cannot attribute a fixation to a particular process (word recognition, parsing, inference, etc.), nor infer a processing stage (e.g., pre-or post-bottleneck) from the timing of a saccade.

Perhaps the broadest concern about the VWP is that because the visual world creates a restricted set of possible referents, it might introduce task-specific strategies that bypass "normal" language processing. This issue has been directly addressed in studies of spoken-word recognition. Three important results are incompatible with the concern that normal processing is bypassed. First, there are effects of lexical frequency (Dahan, Magnuson, & Tanenhaus, 2001). Second, there are neighborhood effects: Words which are similar to many other words (neighbors) are harder to process than words with fewer neighbors (Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Third, target fixations are sensitive to frequency and neighborhood effects in so-called "hidden competitor" designs in which all of the non-target pictures are unrelated distractors and none of the words and pictures are repeated (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Magnuson et al., 2007).

A related concern is that because most language use is not about concrete co-present referents, conclusions drawn from VW studies will not generalize to less constrained situations. To the best of our knowledge, there is little or no evidence suggesting that this might be the case**.** Rather, insights from studying language processing in constrained situations using the VWP seem to scale up to language that is not about a restricted visual context (for discussion see Tanenhaus & Brown-Schmidt, 2008).

Conclusion

The Visual World Paradigm provides a sensitive, time-locked response measure that can be used to investigate a wide range of psycholinguistic questions in language production and language comprehension, ranging from speech perception to collaborative-task-

oriented dialogue. The VWP can be used with participants of all ages, including special populations.

In VW studies, eye-movements to objects or pictures in a visual workspace are monitored as the participant produces and/or comprehends spoken language that is about the co-present "visual world". As visual attention shifts to an object in the workspace, there is a high probability that a saccadic eye movement will rapidly follow to bring the attended area into foveal-vision. Where a participant is looking, and in particular when and to where saccadic eye movements are launched in relationship to the speech, can therefore provide insights into real-time language processing. The VWP combines spoken language and visual search. Therefore, users need to take into account how different aspects of language impact the speech signal. They also need to be cognizant of results about the relationship between eye-movements and visual attention from the relatively new literature on vision in natural tasks.

Acknowledgments

We thank Delphine Dahan, Bob McMurray, and John Trueswell for helpful comments.

Key terms

Competitor: Object in the visual workspace which is related to the target along some specified dimension.

Distractor: Object in the visual workspace which is unrelated to the target.

Look-and-listen VWP: The participant is not given an explicit task.

Point-of-disambiguation: Point in time at which speech and visual context uniquely specify the target; also: point in time at which the proportion-of-fixations curves diverge in favor of the target.

Proportion of fixations: Proportion of trials on which the participant looks at a particular type of picture.

Task-based VWP: The participant performs a well-defined action in VW.

Target: Object in the visual work space which is the referent of the linguistic expression.

Visual world paradigm (VWP): Experimental paradigm which monitors eye movements to objects in a visual workspace as participants listen to, or produce, spoken language about elements of the workspace.

References

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

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition, 93*, B79–87.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–64.

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.

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

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive science, 32*, 643–684.

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

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology, 42*, 317–367.

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

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498–513.

Dahan, D., Tanenhaus, M. K., & Salverda, A. P. (2007). The influence of visual processing on phonetically driven saccades in the "visual world" paradigm. In R.P.G. van Gompel, R.H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 471–486). Oxford: Elsevier.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1352−1365.

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

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

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

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 133–56.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*, B33–42.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology, 60,* 661–671.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*, B25–B33.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Chapman and Hall / CRC.

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.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language, 90*, 103−125.

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (in press). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107−123.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica, 137*, 172–80.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language, 71*, 145–163.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*, 238–299.

Tanenhaus, M. K. (2004). On-line sentence processing: past, present, and future. In M. Carreiras and C. Clifton, Jr. (Eds.), *On-line sentence processing: ERPS, eye movements and beyond* (pp. 371–392). New York: Psychology Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217)*, 1632−1634.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*, 1−23.

Trueswell, J. C., Sekerina, I., Hill., N. M., & Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition, 73*, 89−134.

Yee, E., & Heller, D. (2012). Looking more when you know less: Goal-dependent eye movements during reference resolution. Poster presented at the Annual Meeting of the Psychonomic Society, Minneapolis, MN.

Further reading and resources

For a historical review of foundational VW studies:

Spivey, M. J., & Huette, S. (in press). Toward a situated view of language. In P. Pyykkönen-Klauck, P. Knoeferle, & M. Crocker (Eds.), *Visually Situated Language Comprehension*. Amsterdam: John Benjamins Publishing.

For a more comprehensive review:

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.

As a methodological tool for interactive conversation:

Tanenhaus, M. K., & Trueswell, J. C. (2005). Eye movements as a tool for bridging the language-as-product and language-as-action traditions. In J. C. Trueswell & M. K.

Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 3−37). Cambridge, MA: MIT Press.

Vision and eye movements in natural tasks:

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Visual Neuroscience, 26*, 51–62.

R packages for processing and visualizing visual-world data:

Dink, J. W., & Ferguson, B. F. (2015). eyetrackingR: An R library for eye-tracking data analysis (R package version 0.1.6). Retrieved from http://www.eyetrackingr.com.

Porretta V., Kyröläinen A., van Rij, J., & Järvikivi, J. (2016). VWPre: Tools for preprocessing visual world data (R package version 0.5.0). Retrieved from https://cran.rstudio.com/web/packages/VWPre/

Figure captions

Figure 1. Example of a screen-based visual world paradigm experimental setup.

Figure 2: Example visual display modeled after Altmann and Kamide (1999).

Figure 3: A. Timing of target fixations for each trial, for one participant (from Salverda, Kleinschmidt, & Tanenhaus, 2014). B. Fixation proportions computed for same data.

Figure 4: Proportion of fixations over time (from target-word onset) to target (goat), cohort competitor (goal), and distractor in neutral verb and constraining-verb condition in Experiment 1 in Dahan and Tanenhaus (2004).

Figure 1

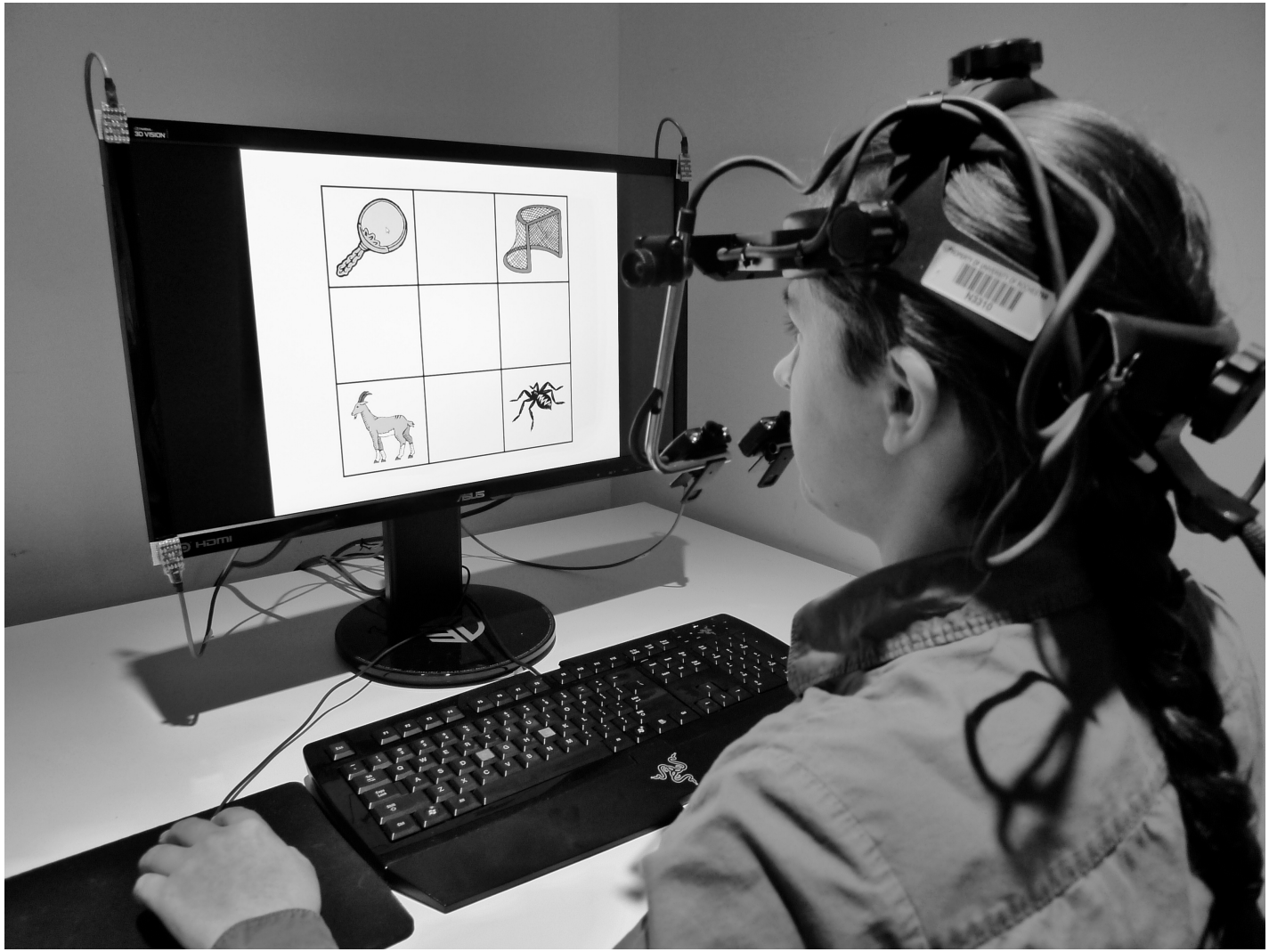


Figure 2

Figure 3

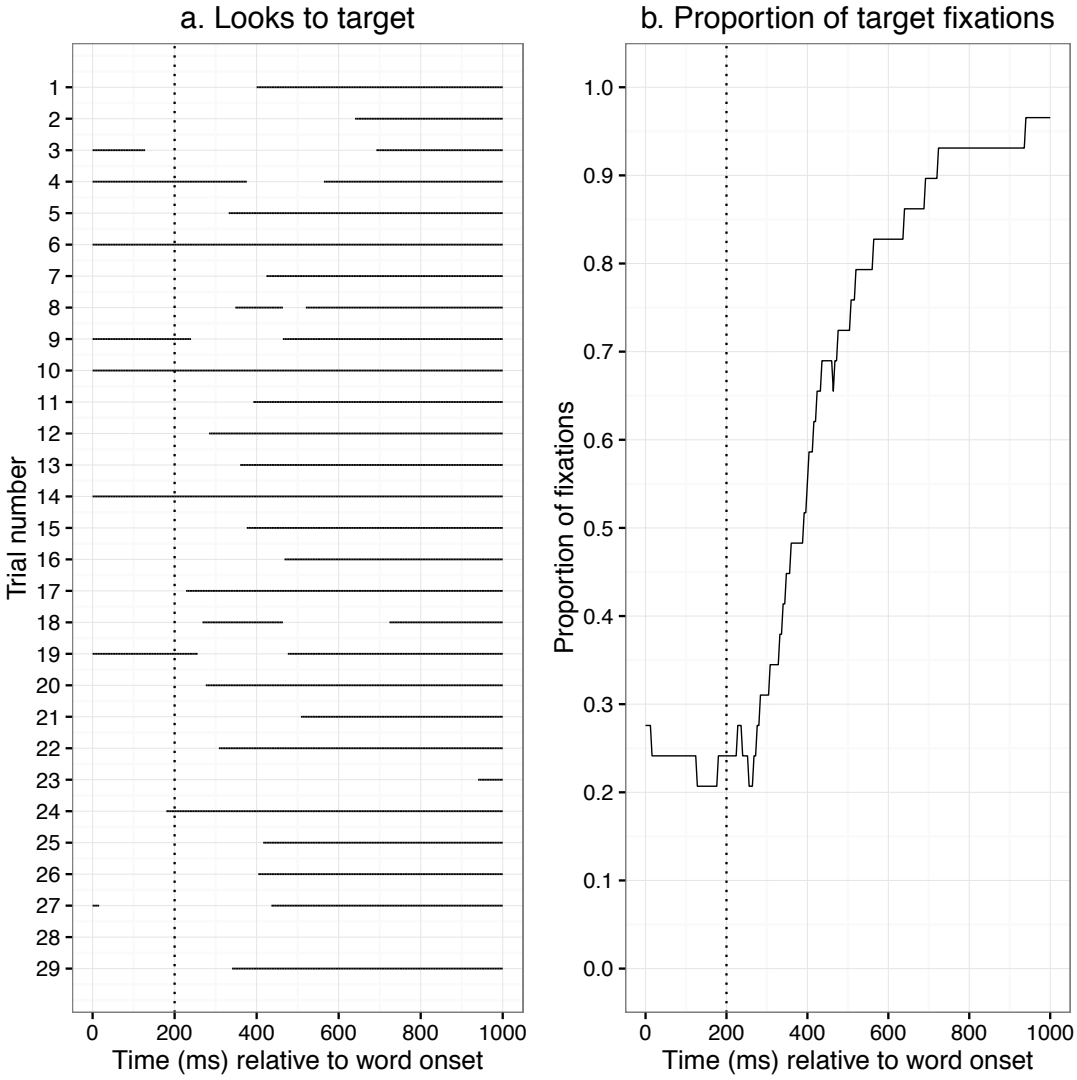
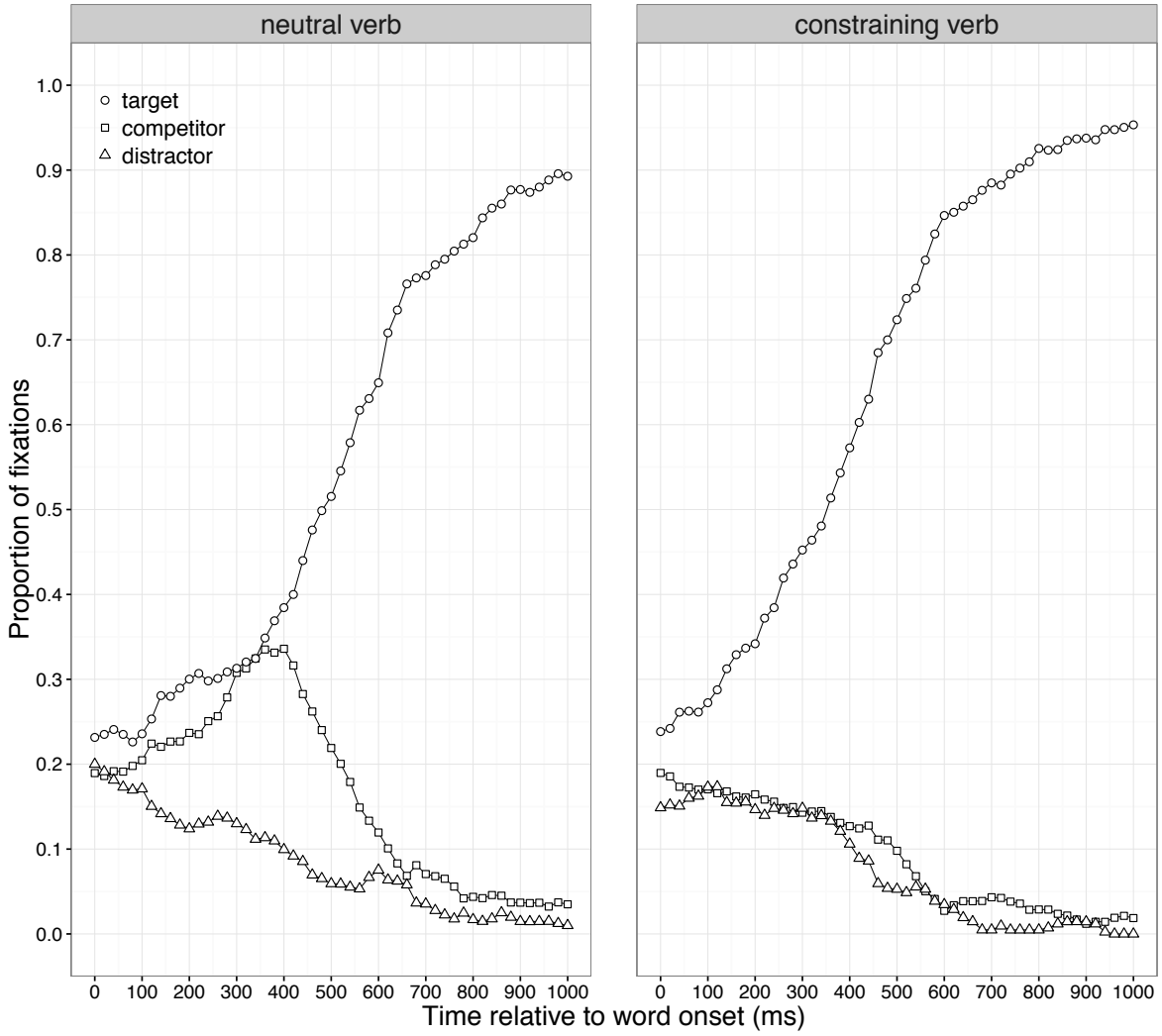

a. Looks to target
Trial number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
0
200
400
600
800
1000
Time (ms) relative to word onset
b. Proportion of target fixations
Proportion of fixations
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
200
400
600
800
1000
Time (ms) relative to word onset

Figure 4


neutral verb
constraining verb
target
competitor
distractor
Proportion of fixations
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
600
700
800
900
1000
Time relative to word onset (ms)