Do we predict upcoming speech content in naturalistic environments?

Heyselaar, Evelien; Peeters, David; Hagoort, Peter

[Link to publication in Tilburg University Research Portal](#)

Do we predict upcoming speech content in naturalistic environments?

Evelien Heyselaar , David Peeters & Peter Hagoort

Routledge
Taylor & Francis Group

REGULAR ARTICLE

ᵃ OPEN ACCESS   Check for updates

Do we predict upcoming speech content in naturalistic environments?

Evelien Heyselaar ⓘ [a,b], David Peeters[a,c,d] and Peter Hagoort[a,c]

ᵃMax Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; ᵇBehavioural Science Institute, Radboud University, Nijmegen, The Netherlands; ᶜDonders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands; ᵈDepartment of Communication and Cognition, TiCC, Tilburg University, Tilburg, The Netherlands

ABSTRACT

The ability to predict upcoming actions is a hallmark of cognition. It remains unclear, however, whether the predictive behaviour observed in controlled lab environments generalises to rich, everyday settings. In four virtual reality experiments, we tested whether a well-established marker of linguistic prediction (anticipatory eye movements) replicated when increasing the naturalness of the paradigm by means of immersing participants in naturalistic scenes (Experiment 1), increasing the number of distractor objects (Experiment 2), modifying the proportion of predictable noun-referents (Experiment 3), and manipulating the location of referents relative to the joint attentional space (Experiment 4). Robust anticipatory eye movements were observed for Experiments 1–3. The anticipatory effect disappeared, however, in Experiment 4. Our findings suggest that predictive processing occurs in everyday communication if the referents are situated in the joint attentional space. Methodologically, our study confirms that ecological validity and experimental control may go hand-in-hand in the study of human predictive behaviour.

Introduction

In the last few decades, there has been an increased interest in the role of prediction in language comprehension. The idea that people predict (i.e. context-based pre-activation of upcoming linguistic input) was deemed controversial at first (e.g. Fodor, 1983). However, present-day theories of language comprehension have embraced linguistic prediction as the main reason why language processing tends to be so effortless, accurate, and efficient (see Clark, 2013; Friston, 2010 for an overview). Current theories of prediction involve the creation of an internally generated model of anticipated upcoming information, similar to the efference copy proposed to drive prediction in the motor movement field (see Wolpert & Flanagan, 2001). The actually encountered linguistic information is then compared against this forward model of anticipated linguistic information (Pickering & Garrod, 2007) and any prediction error is used as a learning mechanism that influences future predictions (Dell & Chang, 2014). Such theories are typically inspired by data collected via EEG and the visual world paradigm, the latter of which we will focus on in this study.

The visual world paradigm (VWP) builds on the observation that when participants are presented with spoken language whilst viewing a visual scene, their eye movements are very closely synchronised to a range of different linguistic events in the speech stream (Cooper, 1974; Huettig et al., 2011b). Altmann and Kamide (Altmann & Kamide, 1999) exploited this behaviour to illustrate that listeners anticipate upcoming linguistic information during online language comprehension. In their seminal study, participants were presented with a visual scene depicting, for example, a boy, a cake, and some toys. While participants heard sentences such as "the boy will move the cake" or "the boy will eat the cake", the authors observed that participants would fixate on the cake significantly earlier after hearing the verb form "eat" (but before "cake" was uttered) compared to after hearing the verb form "move". Hence in an anticipatory way they would move their eyes towards the object corresponding to an assumedly predicted upcoming word. The VWP has since proven to be an excellent method to provide direct evidence of what type of information is anticipated (Coco & Keller, 2015; Hintz et al., 2017; Kamide et al., 2003; Knoeferle & Crocker, 2006, inter alia). Other commonly used methods, such as EEG, have also provided evidence

in favour of linguistic prediction, although it is debatable whether several observed effects truly reflect prediction rather than integration of encountered input with the preceding context (Kochari & Flecken, 2018; Kuperberg & Jaeger, 2017; Kutas et al., 2011; Nieuwland et al., 2018; Pickering & Gambi, 2018; van den Brink et al., 2000). The elegance of the VWP is its ability to measure the direct interaction of language and the visual world, not surprisingly making it a commonly used paradigm.

Although the look-and-listen variant of the VWP is commonly considered as a relatively naturalistic way to measure the interaction between language and visual attention, it is not without its limitations. Participants are typically seated in front of a computer screen and presented with 2D objects, which, in the majority of such experiments, are four simple line drawings presented in a 2 × 2 grid on a white background. It is an open question whether the findings obtained in such settings generalise to everyday situations. For instance, the simple cartoon-like images often have no thematic connection to a broader visual context, and therefore the visual system can almost *only* be guided by linguistic input, potentially making the observed results more relevant to theories of visual search than predictive behaviour (Henderson & Ferreira, 2004). Additionally, a relatively simple visual display may allow experimental participants to preview all the objects and possible targets, subvocalize them, and thus pre-generate the linguistic labels that may appear in the subsequently encountered speech (Andersson et al., 2011), again eliciting behaviour that may resemble predictive processing, but may not be driven by it.

There has also been research using more complex, photographic scenes (Coco et al., 2016; Staub et al., 2012), which has replicated the anticipatory eye-movement behaviour, suggesting that the behaviour observed using cartoon-like images was indeed driven by linguistic predictive processing. These studies used stimuli ranging from photographs of an agent and four objects (Staub et al., 2012) to more complex, cluttered scenes (Coco et al., 2016; Coco & Keller, 2015). The ecological advantage of using such richer scenes is that they include a broader thematic context, which can be considered more reflective of natural everyday situations. In naturalistic scenes, a theme is often clearly evident (i.e. "this is a kitchen"), allowing listeners to anticipate which types of objects will be mentioned and where they can be found. Moreover, such a setup also allows for presenting objects in a realistic spatial perspective, unlike traditional look-and-listen VWP studies in which all objects (cf. a tea cup *vs*. a dog) typically had a similar size on a computer screen.

In a naturalistic conversation, interlocutors furthermore converge on topics that may actually restrict the referential domain. This was illustrated by Brown-Schmidt and Tanenhaus (2008) who used a semi-permanent grid of 57 randomly placed objects, while naïve participants conducted a conversation about these objects. Via eye-tracking, it was shown how proximity, relevance, and recency of referents were helpful factors in restricting the relevant referential domain (Brown-Schmidt & Tanenhaus, 2008). Using more than a single sentence per scene may thus also help to create a more ecologically-valid paradigm to measure anticipatory eye-movement behaviour, but typical look-and-listen VWP experiments have commonly been restricted to the use of a single critical sentence per scene.

A final potential limitation of previous studies using the look-and-listen VWP, in terms of their ecological validity, is that experimental sentences were typically played from a disembodied voice, i.e. in the absence of a visible speaker. Despite recent pleas for the use of (visually as well as socially) richer scenes in experimental research (e.g. Hari et al., 2015; Knoeferle, 2015; Pan & Hamilton, 2018; Willems, 2015), look-and-listen anticipatory eye-movement studies typically lack a visible speaker who produces a communicatively motivated spoken message for the participant addressee.

To address these methodological concerns, we conducted the current experiments in virtual reality (VR). Using VR allowed us to immerse participants in rich, visual scenes in which the presented objects were thematically embedded. Spoken sentence stimuli were communicatively motivated as spoken by a virtual agent who maintained eye contact with the participant. Unlike experimental setups using 2D videos, immersive VR places the participant "in the stimulus", as in everyday situations (cf. Parsons, 2015; Peeters, 2019).

In four experiments, we tested whether anticipatory eye movements are observed when increasing the naturalness of the paradigm by means of: (i) immersing participants in naturalistic everyday scenes, (ii) increasing the number of distractor objects present, (iii) modifying the proportion of predictable noun-referents in the experiment, and (iv) manipulating the location of referents inside or outside the joint attentional space shared by speaker (virtual agent) and addressee (participant). We will further discuss the theoretical rationale behind each of these experiments below.

Experiment 1: immersion in virtual reality

We conducted this study in Virtual Reality (VR) to ensure that participants would feel immersed in the experimental environment, while retaining the required levels of experimental control for reliable data collection (Peeters, 2019). Previous studies have built VR versions

of common psycholinguistic tasks and have shown comparable behaviour with the traditional version (Heyselaar et al., 2015; Peeters & Dijkstra, 2018; Tromp et al., 2018). Recently, Eichert et al. (2018) moreover showed robust anticipatory eye-movement behaviour in a VR version of the classic Altmann and Kamide (1999) look-and-listen VWP task, suggesting that using 3D objects versus 2D pictures in itself does not change participants' anticipatory eye-movement behaviour. The current experiment will go several steps further in making use of the unique affordances of VR and increase the naturalness of the VWP in ways that are hard or impossible to imagine in traditional versions.

As discussed above, a central component of everyday communication is that it typically takes place in a broader, thematically consistent, visual context. Therefore, the backbone of the current experimental set-up is the immersion of participants in realistic everyday scenes such as a living room, an office, a neighbourhood, etc. As a first step towards mimicking real-world face-to-face interaction, participants will be taken on a tour by a virtual agent, who will deliver the critical sentences as she tells the participant about aspects of her life in various relevant visual environments. Contrary to classic VWP experiments, we will present four (rather than one) critical sentences per scene, increasing the odds that participants remain unaware of the goal of the study. In previous experiments, participants would typically receive one critical sentence per scene, and then immediately be presented with a novel scene. Even in studies with multiple utterances per scene (i.e. Andersson et al. (2011) who had three utterances per scene), only one utterance was the critical sentence that referred to an object present in the scene. In Experiment 1, all four utterances refer to an object present in the scene. To minimise any benefits of guessing, we have increased the number of objects from the traditional four to the current six.

In sum, Experiment 1 allowed us to test whether anticipatory eye movements are observed in situations that can be considered more reflective of everyday communication compared to traditional paradigms. The three main changes compared to earlier studies are (i) placing the participant in the role of addressee in the presence of a visible speaker who produces communicatively motivated messages, (ii) at the same time placing the participant in rich visual environments that are thematically organised, and (iii) having multiple critical utterances per scene. The subsequent experiments in this study will manipulate further aspects of this set-up, such as the number of distractor objects or the predictability of the sentences, to build towards a more accurate reflection of real-world situations.

Experiment 2: more potential referents

Previous studies have shown converging evidence that increased visual complexity affects anticipatory eye-movement behaviour. For example, Sorensen and Bailey (2007) observed a significant decrease in the strength of the typically observed anticipatory effect when presenting participants with more than 4 items, and anticipatory eye movements were non-existent in a context with 16 items. Additionally, studies using complex, photographic scenes have also shown reduced language-driven eye-movement activity (Andersson et al., 2011; Coco et al., 2016; Coco & Keller, 2015).

There are concerns that a simple display may allow participants to preview and pre-generate linguistic labels before hearing the linguistic input, and hence perform anticipatory eye-movements that are not supported by prediction mechanisms (Andersson et al., 2011). However, studies have shown that increasing the preview time of the objects does not affect the strength of the anticipatory effect (Sorensen & Bailey, 2007). This suggests that the limitations observed in anticipatory eye movement behaviour may have been due to the number of items the participant could choose from. However, if the preview time is less than 200 ms, visual attention shifts are co-determined by the time-course of retrieval of phonological, shape, and semantic knowledge, an aspect we are not focusing on in this study (Huettig & McQueen, 2007).

Indeed, there is already evidence suggesting that anticipatory eye movements are not dependent on a concurrent visual scene, but rather on the mental record of that scene (Altmann, 2004). In experiments using the so-called "blank screen paradigm", participants hear the critical sentence only after the VWP scene is removed. Yet participants still show anticipatory eye-movements to the location of the referent, although all they see is a blank screen (Altmann, 2004). A likely candidate to maintain this visual record is the working memory system.

There is a growing consensus for the role of working memory in prediction (see for review Huettig et al., 2011a). In the VWP, anticipatory looks to the referent object can occur as early as 200 ms after the verb is heard, a timeframe that suggests that participants already had the potential objects activated to some extent. Huettig and colleagues propose that objects in the display are first encoded to a visuospatial type of working memory (cf. Alvarez & Cavanagh, 2004;

Baddeley, 1998; Pylyshyn, 1989), which triggers perceptual hypotheses in long-term memory. These hypotheses then trigger a cascade of activations of associated semantic and phonological codes, all within a few hundreds of milliseconds (cf. Huettig & McQueen, 2007). This results in a nexus of associated knowledge, which is bound to an object's location within working memory. Hence object selection and planning a saccade to the location of that object is faster due to the already activated representations within working memory, and participants do not need to see the object to be able to make a saccade to its location, as observed in the blank screen paradigm.

Although a recent study showed a correlation between a working memory construct and predictive looks towards 4 objects (Huettig & Janse, 2016), we are not aware of a study showing a more direct link between working memory and predictive looks. Working memory has a limited capacity; therefore, if working memory indeed plays a role in prediction, one would assume that by increasing the number of potential object referents in a visual scene, the anticipatory eye movement behaviour will decrease as participants can no longer accurately maintain the objects' representations online. This prediction is in line with the work of Sorensen and Bailey (2007), who indeed have shown a decrease in anticipatory eye-movement behaviour as the number of objects in the scene increases. Additionally, this behaviour should be modulated by the participant's individual working memory capacity. Therefore, in addition to the main aim of replicating anticipatory eye-movements in a VWP with increasing items, a correlation between the participants working memory capacity and their performance will be explored, to determine whether any decrease in anticipatory eye-movement behaviour is indeed due to the increased number of items the participants need to encode.

If the VWP is indeed an ecologically valid methodology to study the interaction of language and the visual world, then one would predict that anticipatory eye movements also occur in visually rich environments resembling the real world. However, as working memory capacity is limited, even though participants could use strategies such as "chunking" to reduce the load on working memory in thematic scenes, we still expect a decrease in anticipatory eye-movement behaviour when more items are present (Experiment 2) compared to Experiment 1.

Experiment 3: less predictable input

Increased realism is not limited to visual complexity. As displacement is an important and common feature of

present day human communication (Hockett, 1960), not every sentence in a conversation necessarily refers to an object in the interlocutor's immediate environment. Therefore, in Experiment 3, we will include filler sentences that refer to objects not present in the scene. The distribution is such that per scene, only 50% of the sentences refer to any of the objects present, and only 25% of the total sentences will utilise verbs that allow the noun to be predicted on the basis of the visual context. This manipulation therefore also tests whether participants would adapt their predictive behaviour when they realise that the majority of the referential nouns cannot be predicted and therefore it would be relatively inefficient to try.

Although there are many different proposed mechanisms underlying prediction (cf. Altmann & Mirković, 2009; Chang et al., 2006; Dell & Chang, 2014; Kahneman, 2011; Kuperberg, 2007; Pickering & Garrod, 2007, 2013), the majority propose that prediction makes use of previous experience. Events tend to recur and show regularities and therefore are likely to be an important organising principle of past experience. As described in Dell and Chang (2014, p. 4):

> the central component of the model tries to predict the next heard word from the word that preceded it and a representation of prior linguistic context. It then compares the predicted next word with the actual next word. The resulting prediction error is used to change the model's internal representations, thus enabling the model to acquire the knowledge that helps it make these predictions.

Errors in prediction in general are a valuable source of information about whether an organism's representation of the environment is effective, and are the main mechanism underlying reinforcement learning. With respect to *linguistic* prediction, recent preliminary evidence indeed suggests that predictive behaviour can be influenced by immediate past experience (i.e. even within a single experimental session). For instance, Experiment 2 in Brothers et al. (2017) showed an elimination of word predictability during a self-paced reading task when predictable cues were no longer valid. This suggests that linguistic prediction may not be an automatic process, but can be strategically manipulated as a function of distributional variation in recent linguistic input (for further discussion, see Pickering & Gambi, 2018).

In terms of our experimental goal, if a single VWP session is influential enough to discourage participants from producing anticipatory eye-movements, this would suggest that anticipatory eye-movements may not be very prevalent in ecologically valid interactions in which speakers also not always necessarily refer to

entities in their direct environment in each utterance they produce.

Experiment 4: less obvious attentional focus

Unlike typical studies using the look-and-listen VWP, the present experiments include an immediate source for the sentences participants perceive. For each scene a virtual agent will be present and will speak the sentences to the participant. In our experiments, the participant faces the virtual speaker, to a certain extent mimicking naturally occurring communication in which interlocutors often form a conversational dyad. We know that, in everyday communication, interlocutors transform physical space into meaningful space (Kendon, 1977; Scheflen & Ashcraft, 1976). They typically use their bodies to separate their *joint attentional space* of engagement from the larger outside world (Kendon, 1990a, 1992). Certain objects speakers refer to may be present within this joint attentional space, whereas others may be located outside of it in the participant's visual periphery (Peeters et al., 2015), and interlocutors typically keep track of whether they are attending to something in common (Tomasello, 1995).

In Experiment 4, we exploit the unique affordances of immersive virtual reality and place object-referents outside the joint attentional space shared between participant and speaker. It is an open question whether the canonical pattern of anticipatory eye movements replicates when referents are placed slightly outside central vision in a rich and interactive everyday environment. Would the typical pattern of anticipatory eye movements have been observed if the critical stimuli were presented distributed over an entire 3D visual scene, rather than in central focus of attention in front of a participant on a computer monitor? After all, in naturally occurring communication we also talk not only about entities that are located directly inside the conversational dyad between speaker and addressee. Experiment 4 will test whether participants will still consider objects placed outside the joint attentional space as a potential target for the sentences uttered by the virtual agent.

Overall aim

In sum, the VWP is an important methodology used to investigate linguistic prediction. Although it aims to measure ecologically-relevant behaviour, it comes with several limitations that may have encouraged behaviour that the average person may not produce in the real world. Therefore, in this study we will create a more realistic VWP by placing participants in 3D worlds with thematic objects and an actual, virtual speaker. By increasing the number of objects and manipulating how often participants hear a sentence with a predictable noun-referent, we not only measure anticipatory eye-movements in real-world contexts, but we are also able to empirically test whether elements such as working memory and past experience do indeed play an important role in linguistic prediction in everyday settings.

Experiment 1: improving the visual world paradigm

This study and experiments 1, 2, and 3 were pre-registered via the Open Science Framework and can be found under the title: "Language-driven anticipatory eye-movements in naturalistic settings". All the data, stimuli, and analysis scripts are available on the Open Science Framework under the same title.[1] Experiment 4 and Overall Results were not pre-registered and therefore fully exploratory. The chosen sample size per experiment was a priori determined to be identical to Eichert et al. (2018).

Materials and methods

Participants
Twenty native speakers of Dutch (13 female, M_{age}: 22.8 years, SD_{age}: 3.50 years) were recruited from the Max Planck Institute for Psycholinguistics database. The data of 24 participants was recorded, but one participant was discarded due to insufficient accuracy of the eye-tracking data and three stated during the debrief stage that they did not understand the virtual agent properly (clarity rating < 3 out of 5). The participants gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Materials
Virtual agent. The virtual agent was adapted from a stock avatar produced by WorldViz (Santa Barbara, CA; "casual03_f_highpoly"). The virtual agent's appearance suggested that she was a Caucasian female in her mid-twenties, which matched the age and ethnicity of the native Dutch speaker who recorded her speech. All the virtual agent's speech was pre-recorded.

Scenes. Eight scenes were designed to represent places in the virtual agent's life (her office, her neighbourhood, her living room, etc.; see Appendix I for a full list of scenes). The scenes were designed to appear as realistic as possible (Figure 1) but initially without any objects. Scenes that came with furniture (such as the table in Figure 1) were designed such that these items would
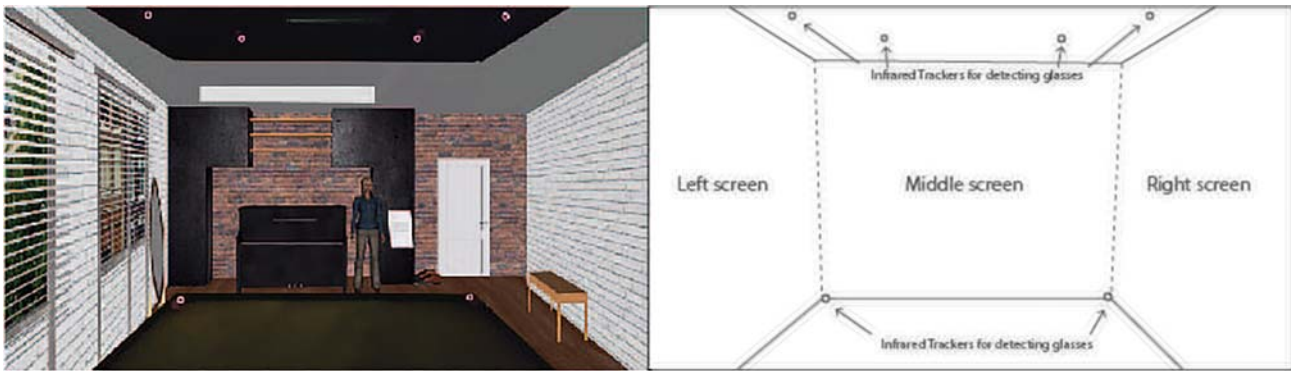
Figure 1. Example scene. The figure illustrates the living room scene, one of eight scenes used in this study. Participants, wearing 3D-glasses, stood in the middle of the room during the task and were allowed to move their heads to view the entire scene. The location of the two screens to the left and right gives the participants a feeling of being immersed in the scene. The virtual agent was always present on the middle screen, to ensure participants were able to locate her easily and would feel addressed when she spoke to them. The 6 objects present (mirror, piano, letter, shoes, door, and table in this specific scene) were placed in naturalistic locations and scaled to realistic proportions relative to the room. Participants heard four sentences while viewing this scene, two of which were restrictive (e.g. *Tonight, I should remember to mail the letter*) and two of which were unrestrictive (e.g. *Tonight, I should remember to move the letter*). In this case, a pre-test indicated that participants found the letter the only mail-able object in the scene, yet they found the letter, mirror, piano, shoes, and table moveable.

not be predictable given the sentences. Objects were then placed in realistic locations in each scene. For example, the car in the neighbourhood scene was placed in the driveway, the tree was placed on the grassy lawn, and the basketballs were placed on the sidewalk. The aim was to place the objects in such a way that they were not overly salient, however, objects always appeared in the middle screen between the virtual agent and the participant so that the participant did not have to search for them. The virtual agent appeared in each scene in the middle screen such that participants would feel addressed when she spoke to them.

Objects and sentences. Thirty-two sentence pairs were created, of which one sentence was *restrictive* (the verb imposed constraints on its arguments such that only one of the visually presented objects was a plausible completion of the sentence) and one was *unrestrictive* (no such constraints were imposed; the sentence could be completed with at least three of the objects present in the scene). The sentence pairs therefore only differed in their verb. For example, a sentence pair would consist of *na werktijd drinkt soms iemand een kopje koffie* ("after work, sometimes someone drinks a cup of coffee") versus *na werktijd haalt soms iemand een kopje koffie* ("after work, sometimes someone gets a cup of coffee"; see Appendix I for a full list of sentences and their English translations). The verbs were chosen such that their word length and frequency were not significantly different between conditions (length: Mann–Whitney $U = 416$, $p = .189$; frequency: Mann–Whitney $U = 475$, $p = .619$).

All the objects present in the experiment were selected from a standardised database of 3D objects (Peeters, 2018) to ensure that all objects were easily identifiable. The experiment contained eight scenes. Each scene included four sentences; six objects were present in each scene. This ensured that even with the fourth sentence, there were still three objects that had not yet been mentioned, ensuring that participants could not accurately guess the target object for the final sentence.

Thirty-eight participants (who were not invited for the main experiment) completed an online Cloze-like task to ensure that the target object was the most likely completion for the restrictive sentences (M: 92.67%, SD: 18.98%) compared to the unrestrictive sentences (M: 19.51%, SD: 21.21%). Participants were given the incomplete sentence and asked to choose the most likely completion from a list of the objects in the scene.

Sentences were recorded in a sound-proof booth, sampling at 44.1 kHz (stereo, 16 bin sampling resolution). All files were equalised for maximal amplitude. Sentences were annotated using Praat (Boersma & Weenink, 2009) by placing digital markers at onsets and offsets of critical words: Verb onset, verb offset, noun onset, noun offset, and end of sentence. The mean duration of the sentences was 2,474 ms. During recording of the sentences, we ensured an average of 571 ms (SD: 116 ms) between the end of the verb and the start of the noun (time of interest [TOI]), as previous research has shown that at least 500 ms is necessary to successfully allow prediction effects (Salthouse et al., 1999). Typically, verb and noun were separated by at

least two words (e.g. an article and an adverb). We observed no significant difference in the length of the TOI between the two conditions ($t(62) = -0.51$, $p = .612$).

Apparatus

CAVE system

The experiment was run in a CAVE Virtual Reality set-up (see Figure 1), the layout of which has been described before in detail (Eichert et al., 2018, see their Figure 4). The CAVE system consisted of three screens (255 cm x 330 cm, VISCON GmbH, Neukirchen-Vluyn, Germany) that were arranged at right angles. Two projectors (F50, Barco N.V., Kortrijk, Belgium) illuminated each screen indirectly through a mirror behind the screen. The two projectors showed two vertically displaced images which were overlapping in the middle of the screen. Thus, the complete display on each screen was only visible as combined overlay of the two projections. For optical tracking, infrared motion capture cameras (Bonita 10, Vicon Motion Systems Ltd, UK) and Tracker 3 software (Vicon Motion Systems Ltd, UK) were used. The experiment was programmed and run using 3D application software (Vizard, Floating Client 5.4, World-Viz LLC, Santa Barbara, CA), which makes use of the programming language Python. Sound was presented through two speakers (Logitech, US) that were located at the bottom edges of the middle screen.

Eye-tracking

Eye-tracking was performed using special glasses (SMI Eye-Tracking Glasses 2 Wireless, SensoMotoric Instruments GmbH, Teltow, Germany) that combine the recording of eye gaze with the 3D presentation of VR. The recording interface used was a tablet that was connected to the glasses by cable. The recorder communicated with the externally controlled tracking system via a wireless local area network, which enabled live data streaming.

The glasses were equipped with a camera for binocular 60 Hz recordings and automatic parallax compensation. The shutter-device and the recording interface were placed in a shoulder bag worn by the participants. This enabled the participants to move freely through the CAVE if they so chose. In reality, the participants stayed standing in the centre of the room, roughly 180 cm away from the central screen. Gaze tracking accuracy was estimated by the manufacturer to be 0.5° over all distances. We found the latency of the eye-tracking signal to be 60 ms ± 10 ms. This latency was corrected for in the statistical analyses (see below).

By combining eye-tracking and optical head-tracking, we were able to identify the exact location of participants' eye gaze in three spatial dimensions, allowing them to move their heads during the experiment. Optical head-tracking was accomplished by placing light reflectors on both sides of the glasses. Three spherical reflectors were connected on a plastic rack and two of such racks with a mirrored version of the given geometry were manually attached to both sides of the glasses using magnetic force. The reflectors functioned as passive markers which were detected by the infrared tracking system in the CAVE. The tracking system was trained to the specific geometric structure of the three markers and detected the position of the glasses with an accuracy of 0.5 mm.

Regions of interest

In order to determine target fixations, we defined individual 3D regions of interest (ROIs) around each object in the virtual environment. The x (width) and y (height) dimensions of the ROI were adopted from the frontal plane of the object's individual bounding box, facing the participant. We adjusted the size of this plane to ensure a minimal size of the ROI. The minimal width was set to 0.8 and the minimal height to 0.5. For the presented layout of objects, the adjusted x and y dimensions were sufficient to characterise the ROIs. Despite the 3D view, the plane covered the whole object sufficiently to capture all fixations. The z dimension (depth) of the ROI was therefore set to a relatively small value of 0.1. An increased z value of the ROIs would not have been more informative about the gaze behaviour, but would have led to overlapping ROIs in some cases. The eye-tracking software automatically detected when the eye gaze was directed to one of the ROIs and coded the information online in the data stream. Some previous studies have used contours of the objects to define ROIs, but rectangles have been shown to produce qualitatively similar results (Altmann, 2011; Eichert et al., 2018). In addition to the six objects in each scene, an ROI was also coded for the virtual agent.

Design and procedure

Participants were instructed to stand in the middle of the CAVE system, roughly 180 cm away from the middle screen. They put on the VR glasses, which were softly fastened using a strap on their head to ensure stability. Prior to the start of the experiment, two calibration steps were performed. For the first calibration step, calibration was done using the SMI software "One-step Calibration" programme. The second calibration step is as described by Eichert et al. (2018): Participants were asked to look at three displayed spheres successively. The experimenter

selected the corresponding sphere. The computer software computed a single dimensionless error measure of the eye-tracker combining the deviance in all three coordinates. The computer-based calibration was repeated until a minimal error value (<4) and thus maximal accuracy was reached. Deviance was checked during the break and re-calibrated using the three-sphere procedure if the error value was greater than 4. This was only necessary for one participant.

Prior to the start of the experiment, participants were informed that they would be given a tour of a virtual agent's life and that the goal of the experiment was to form an opinion of the virtual agent. After the virtual reality portion, they were told they would be given a questionnaire asking for their opinion of the virtual agent. This ensured that the participants paid attention to the virtual agent and drew potential attention away from the objects. During the debrief stage, none of the participants had guessed at the purpose of the experiment, although one participant thought that they had to memorise which objects were present in each scene.

Participants were presented with two experimental blocks of four scenes each. The first block contained the office, forest, café, and canteen scene; the second block contained the living room, bathroom, attic, and neighbourhood scene (see Appendix I). All scenes were randomised within each block for each participant, although the living room scene was always the first scene presented in the second block for all participants (see below). Each scene had a preview time of 1s before the virtual agent gave a short introduction ($M = 2.02s$), after which there was a 2.5s wait time before the first sentence was played. This gave participants an average of 4.5s preview time of each scene. For the living room scene, the virtual agent's introductory text was "welcome to my house" and hence it was always the first scene of that block. The task took around 7 min to complete.

We created two lists of 32 restrictive sentences and 32 unrestrictive sentences taken from each sentence pair. No list contained both the restrictive and unrestrictive versions of the same sentence pair. Participants were assigned to a list based on their participant number (odd participants were assigned to list 1; even participants were assigned to list 2). Sentences were presented randomly within each scene for each participant. As the last sentence presented in each scene meant that the participants had had a maximal viewing time of the scene and its objects compared to the first sentence presented, by randomising the sentences, this balanced out any beneficial effects across the experiment.

Participants were given a self-timed break after the fourth scene. During this time participant's calibration was checked and re-calibrated if necessary. Calibration

was also checked at the end of the experiment. After the experiment, participants were given a debrief questionnaire in which they were asked to rate the clarity of the virtual agent's speech as well as indicate which objects they heard the virtual agent refer to. This list contained all the objects present in the experiment, of which only 66.67% were actually named by the virtual agent. Accuracy on this questionnaire was taken as an indication of how well the participants paid attention to what the virtual agent was saying.

Statistical analyses

Data was acquired at a sampling frequency of 60 Hz. We corrected for the 60 ms latency shift caused by the eye-tracking system by time-locking the data to 60 ms (~4 frames) after each sentence onset. A fixation was defined as a look to the same ROI that lasted at least 100 ms. This correction on the experimental data led to an exclusion of 6.93% of all frames logged as object fixations, and 2.36% of all frames logged as virtual agent fixations. Fixation data was then aggregated into time bins of 50 ms (i.e. three data frames).

We followed the steps outlined in Porretta et al. (2017) for analysing visual world paradigm data with general additive mixed models (GAMM). This differs from the approach we preregistered, as we reported that we would use generalised linear mixed effects models (GLMER). Unlike ANOVAs or GLMER, GAMM does not assume linearity (although it can find a linear form if supported by the data). Instead, GAMM strikes a balance between model fit and the smoothness of the curve using either error-based or likelihood-based methods in order to avoid over- or under-fitting. Thus, the data guide the functional form (Hastie & Tibshirani, 1990). The *p*-value provided therefore indicates whether or not the curve is significantly different from zero (a flat line). Additionally, GAMM also allows the inclusion of random effects to capture the dependencies between repeated measures. As discussed in Porretta et al. (2017), VWP experiments produce time-series data in which the sequential measurements tend to be correlated. This "autocorrelation" violates the assumptions of many statistical tests (Baayen et al., 2016), and is a problem also not addressed in growth curve analysis, where it results in an increased risk of overconfidence (for a further comparison of growth curve analysis and GAMMs, we refer the reader to Porretta et al. (2017), p. 271). Additionally, as VWP time series data are rarely linear, this poses challenges for statistical methods that assume a linear relationship (such as GLMER). Statistical "work-arounds" often involve data simplification, resulting in the potential loss of information, or an incorrect

reflection of the true underlying trends. As the paper by Porretta and colleagues outlines how GAMM analysis can be applied specifically to VWP data, we opted to diverge from our pre-registration. The data were never analysed using GLMER.

The analysis was conducted using the *mgcv* package (version 1.8-22; Wood, 2017) and *itsadug* package (version 2.3; van Rij et al., 2017) in R (version 3.4.2; R Core Development Team, 2011). As the dependent variable we entered the empirical logits of the proportion of target fixations per time bin. Instead of random effects or random slopes, we used random smooths as they adjust the trend of a numeric predictor in a non-linear way. We built the model as per the procedure outlined in Porretta et al. (2017).

The model included random smooth interactions for *Time* by *Subject*, factor smooth interactions for *Time* by *Sentence*, as well as a smooth for *Time* by *Condition* (restrictive versus unrestrictive; sum contrast coded). We included *Condition* as a parametric component, which is necessary to estimate the time curve for each level of *Condition*. We also included weighted linear regression over empirical logits as weights in the model (Barr, 2008). After fitting the model, we determined an appropriate value for the AR1 parameter using the start_value_rho to account for autocorrelation in the residuals (i.e. error). We used the function plot_diff to approximate the time intervals of significant differences between conditions based on the model predictions.

Results

Participants were able to accurately identify which objects the virtual agent had named and which she had not named 90.25% of the time (SD: 8.19%) after the experiment. Therefore, we are confident that all participants listened to the virtual agent throughout the experiment.

Inspection of the grand mean

To determine whether participants fixated on the target object at all during each trial, regardless of condition, we plotted the proportion of target fixations collapsed over all participants, trials, and conditions (Figure 2A). For this figure, each trial is time-locked to verb onset to give an accurate indication of eye movement behaviour in the moments after the verb is comprehended. Visual assessment of the grand mean shows a robust increase in the proportion of fixation to the target object after the noun was mentioned.

Effect of condition

For the main statistical analysis, we defined a critical time window where we expected the experimental manipulation to have an effect on the proportion of target fixations. We chose the onset of the critical window as 200 ms after verb onset, assuming that it takes approximately 200 ms to plan and initiate a saccadic movement (Matin et al., 1993). As offset of the critical time window we chose the average onset of the noun (900 ms after verb onset), in line with previous studies (Altmann & Kamide, 1999; Eichert et al., 2018).

We performed a generalised additive mixed model (GAMM) analysis as outlined in Porretta et al. (2017). The model included factor smooth interactions for *Time* by *Subject*, factor smooth interactions for *Time* by *Sentence*, as well as a smooth for *Time* by *Condition* (restrictive versus unrestrictive). We included *Condition* as a
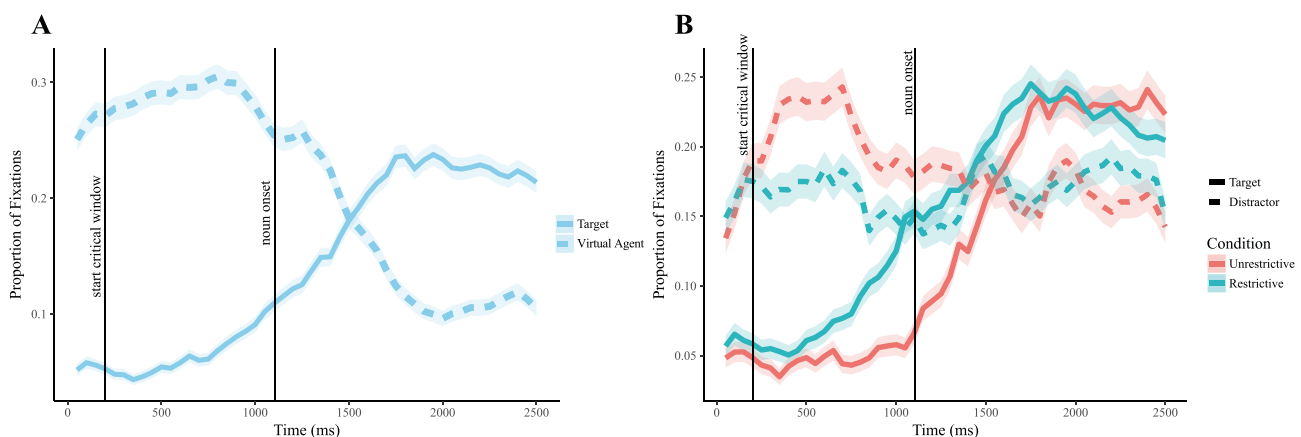
Figure 2. Mean proportions of fixations. A. To the target object and virtual agent. B. To the target and distractor objects shown per condition. Vertical lines indicate critical time points. 0 ms indicates verb onset, the label "start of critical window" is the start of the critical window (200 ms after verb onset). The main statistical analysis was performed on the interval between the start of the critical window and noun onset. Error clouds indicate standard error.

Table 1. Summary of the generalised additive mixed model for changes in target fixations over time, per condition (restrictive versus unrestrictive sentences) for Experiment 1.

Parametric coefficients:					
	Estimate	SE	t-value	p-value	
Intercept	−1.64	0.05	−30.98	<.001	***
Condition	−0.10	0.02	−4.21	<.001	***
Smooth terms					
	edf	Ref.df	F-value	p-value	
Smooth for Time – Unrestrictive	1	1	1.11	.293	
Smooth for Time – Restrictive	1.18	1.28	14.05	<.001	***
Random effect for Subjects	58.80	179	3.18	<.001	***
Random effect for Sentences	170.66	575	2.22	<.001	***

*** < .001.
Effective degrees of freedom (edf), reference degrees of freedom (Ref.df).

parametric component. After fitting the model, we determined an appropriate value for the AR1 parameter, in this case $\rho = -0.10$, to account for autocorrelation in the residuals (i.e. error). Table 1 provides a model summary.

The model output for a GAMM consists of two sections. The Parametric Coefficients report the non-smoothed (i.e. linear) estimates. The Smooth Terms report the smoothed factors, as defined in the model. If the Estimated Degrees of Freedom are equal to 1, that means the correlation was linear. Anything greater than 1 indicates a non-linear relationship.

The model revealed that the parametric coefficient Condition was significant, suggesting that a linear model would also have revealed a significant difference between the restrictive and unrestrictive conditions. The smooth curve for the restrictive condition as a function of time (Smooth for Time – Restrictive) was significantly different from zero (i.e. the curve changed significantly over time), whereas this was not the case for the unrestrictive condition (Smooth for Time – Unrestrictive, $p = .293$). This suggests that there is a significant increase in target fixations over time (within the critical window) for the restrictive, but not the unrestrictive, condition.

In addition to an inspection of the model summary, the *itsadug* package also allows for a visual comparison of the model's estimates (via the *plot_diff* function) to test for significance. It does this by the visual plotting of the estimated difference between two conditions (in this case, restrictive versus non-restrictive) from a GAMM. In addition to a visual plot of the differences, the function also gives as output the time window in which the two factors are significantly different from each other, allowing us to narrow down, within the critical window, to when the two conditions significantly deviate. The difference between the restrictive and unrestrictive condition was significant between 398 and 900 ms after the start of the critical window, estimated based on the model. Fixation proportions time-locked to verb onset are illustrated in Figure 2B.

We performed the same analysis on the mean distractor fixations. The model revealed the same effects, except that now there was a significant effect for the unrestrictive condition ($p = .011$) and not the restrictive condition ($p = .329$). The difference between the restrictive and unrestrictive condition became significant between 314 and 900 ms after the start of the critical window.

These results are consistent with the hypothesis that participants directed their gaze towards the target object before noun onset in the restrictive condition, but not in the unrestrictive condition. Complementary to the target fixations, fixations to the distractor objects revealed that participants fixated more on distractor objects during the unrestrictive condition compared to the restrictive condition. There was no effect of condition on the proportion of fixations on the virtual agent ($p > .203$).

Experiment 2: more potential referents

Experiment 1 showed the standard anticipatory eye movement effects seen in the literature (Altmann & Kamide, 1999; Eichert et al., 2018), even in the more realistic setting our virtual reality system provided. As a next step, we enhanced the complexity of our scenes by increasing the number of objects in each scene from 6 to 10. For each sentence, the participants will therefore have to select from 10 potential objects within 500 ms. We additionally measured the participant's working memory capacity using a sequential comparison task with the aim of correlating it to their anticipatory eye-movement behaviour.

Materials and methods

Participants
Twenty native speakers of Dutch (17 female, M_{age}: 21.5 years, SD_{age}: 1.76 years) were recruited from the Max Planck Institute for Psycholinguistics database. These participants had not participated in the previous experiment. The data of 27 participants was recorded, but six participants were discarded due to insufficient accuracy of the eye-tracking data and one stated during the debrief stage that they did not understand the virtual agent properly (clarity rating < 3 out of 5). The participants gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Materials and design
The same materials and apparatus were used as described for Experiment 1. We selected four extra

objects per scene that fit the theme of the scene (for example, a calculator in the office scene – see Appendix I for a full list of added objects). The objects were not predictable given the restrictive verbs used in that scene, however, they were allowed to be candidates for completion in the unrestrictive conditions (i.e. "my colleagues hate it when someone throws away a – "). The objects were placed in realistic locations within each existing scene.

Visual working memory task

We used a saccadic adaptation of the sequential comparison task (Heyselaar et al., 2011; Luck & Vogel, 1997) to assess visual working memory capacity. We chose this task as it arguably reflects the working memory used to complete the anticipatory language task in a reliable way: Participants view objects to be remembered and make a saccadic eye movement to the target object. The visual working memory task was performed after the participants completed the recall questionnaire and was conducted in the CAVE system. Although the items were not rendered as 3D, the CAVE enabled us to use the eye-tracking system to record their eye movements and fixations. The visual working memory task took place on the middle screen only, with the entire array visible without the participant needing to move their head.

Our task is based on the one described by Heyselaar et al. (2011). Stimulus arrays consisted of sets of two to five coloured squares presented around a central fixation spot (Figure 3). For each set size, the spatial configuration of the squares remained identical across trials. For set size two, squares were on the right and left sides of the fixation spot. For set size three to five, squares were arranged equidistantly from each other with one square located directly above the fixation spot.

The colour of each square was chosen randomly from a pre-determined library of six colours highly discriminable from each other. We used the Adobe Color Wheel (www.color.adobe.com) to choose six analogous colours. A given colour could only appear once in each array.

Figure 3 depicts the order of events in one trial. Each trial began with the presentation of a white fixation spot at the centre of the middle screen. Participants were required to fixate this spot for a jittered period of 500–800 ms. While they maintained fixation, a memory array composed of a randomly determined set of two to five squares was presented for 100 ms. Offset of the memory array was followed by a 900 ms retention interval, in which the display screen was blank with the exception of the central fixation spot. At the end of the retention interval, a test array was presented
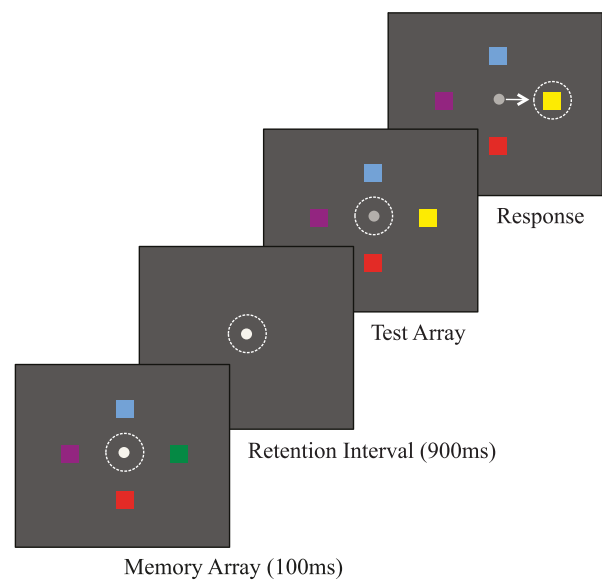
Figure 3. Depiction of a correctly performed trial in the sequential comparison task. Dotted lines and arrow represent current eye position and the saccade response. Participants were required to maintain fixation on the central fixation spot until the spot turned grey, signalling that they were allowed to move their eyes. Adapted from Heyselaar et al. (2011).

consisting of the same number and spatial configuration of the squares as in the memory array, but with the colour of one square changed. Concurrent with this, the fixation spot was dimmed and participants were required to make a saccade to the location of the changed square within 2 s. An inter-trial interval of a jittered 1000–1500 ms followed before the next trial started.

Participants completed 80 trials, 20 for each set size. For each trial, there was always one square that was changed. The first square fixated was taken as the participant's response. Therefore, participants could not fixate all squares within the 2s and still be marked as correct (unless the first square fixated was the changed square). This task took around 10 min to complete.

Statistical analysis

The same statistical analysis was used as described in Experiment 1. We removed 9.47% of all frames logged as object fixations and 4.70% of all frames logged as virtual agent fixations.

Results

Participants were able to accurately identify which objects the virtual agent had named and which she had not 90.53% of the time (SD: 6.96%) after the experiment. Therefore, we were confident that all participants listened to the virtual agent throughout the experiment.

A

B
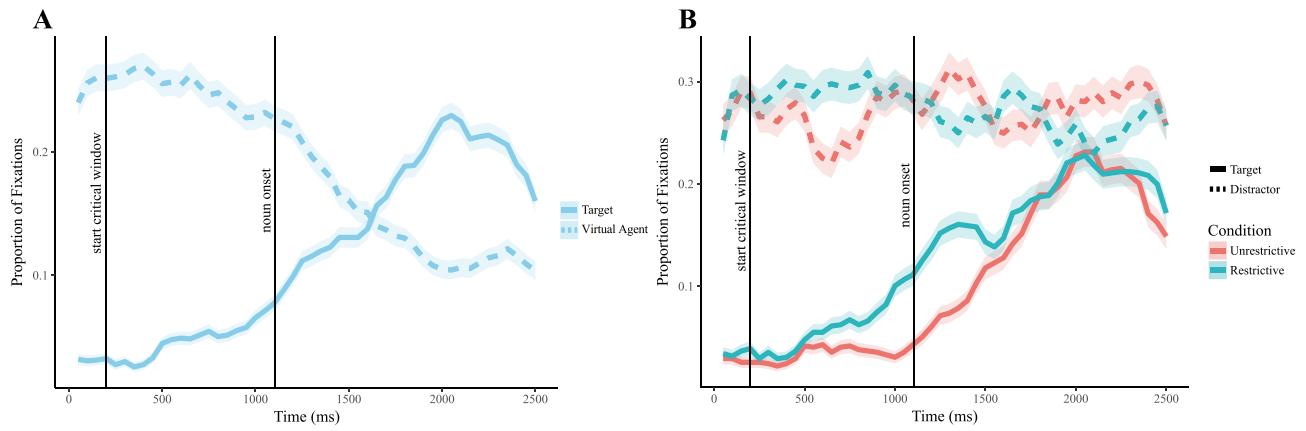
Figure 4. Mean proportions of fixations. A. To the target object and virtual agent. B. To the target and distractor objects shown per condition. Vertical lines indicate critical time points. 0 ms indicates verb onset, the label "start of critical window" is the start of the critical window (200 ms after verb onset). The main statistical analysis was performed on the interval between the start of the critical window and noun onset. Error clouds indicate standard error.

Figure 4A illustrates the grand mean for this experiment. We again observed a robust increase in the proportion of looks to the target object after it was named. Figure 4B illustrates the proportion of looks per condition over time. After fitting the model, we determined an appropriate value for the AR1 parameter, in this case $\rho = -0.03$, to account for autocorrelation in the residuals (i.e. error). Table 2 reports the summary output of the GAMM analysis.

We again observed an increase in the proportion of looks to the target object as a function of time for the restrictive ($p = .015$) but not the unrestrictive ($p = .936$) condition. The difference between the two conditions became significant between 688 and 900 ms after verb onset, based on the model. No effect of condition on the proportion of distractor fixations was observed ($p > .181$).

We thus again observed anticipatory eye-movement behaviour for the restrictive condition versus the unrestrictive condition, in spite of an increase in the number of potential target objects. This suggests that, even in scenes

enriched with more objects, participants anticipated which object the virtual agent would name on the basis of restrictive information encountered at the verb.

An additional analysis was conducted that tested for the role of individual differences in working memory capacity in driving participants' anticipatory eye movements in Experiment 2. The observed median working memory capacity in our participants was 2.67 items ($M = 2.66$), in line with previous visual working memory capacity studies using the sequential comparison task (Luck & Vogel, 1997; Vogel et al., 2006; Vogel & Machizawa, 2004, inter alia). We next conducted a GAMM analysis to determine whether working memory capacity could influence the anticipatory eye-movement behaviour of the participant. The model included a factor smooth for *Subject*, a factor smooth for *Sentence*, as well as a smooth for *Working Memory Capacity* by *Condition* (restrictive versus unrestrictive). We included *Condition* as a parametric component. Table 3 reports the summary output of the GAMM analysis. A significant effect of working memory on anticipatory eye

Table 2. Summary of the generalised additive mixed model for changes in target fixations over time, per condition (restrictive versus unrestrictive sentences) for Experiment 2 (More referents).

Parametric coefficients:					
	Estimate	SE	*t*-value	*p*-value	
Intercept	−1.73	0.04	−39.93	<.001	***
Condition	−0.07	0.04	−1.66	.097	
Smooth terms					
	edf	Ref.df	*F*-value	*p*-value	
Smooth for Time – Unrestrictive	0.24	0.38	0.01	.936	
Smooth for Time – Restrictive	1	1	5.90	.015	*
Random effect for Subjects	20.81	25.33	4.81	<.001	***
Random effect for Sentences	149.88	574	2.80	<.001	***

*** < .001

Effective degrees of freedom (edf), reference degrees of freedom (Ref.df).

Table 3. Summary of the generalised additive mixed model for changes in target fixations per working memory capacity, per condition (restrictive versus unrestrictive sentences).

Parametric coefficients:					
	Estimate	SE	*t*-value	*p*-value	
Intercept	−1.68	0.06	−29.83	<.001	***
Condition	−0.07	0.04	−1.60	.109	
Smooth terms					
	edf	Ref.df	*F*-value	*p*-value	
Smooth for WM – Unrestrictive	1	1	0.03	.867	
Smooth for WM – Restrictive	7.13	8.08	4.34	<.001	***
Random effect for Subjects	13.55	15.00	25.26	.002	***
Random effect for Sentences	58.68	62.00	22.27	<.001	***

*** < .001.

Effective degrees of freedom (edf), reference degrees of freedom (Ref.df).
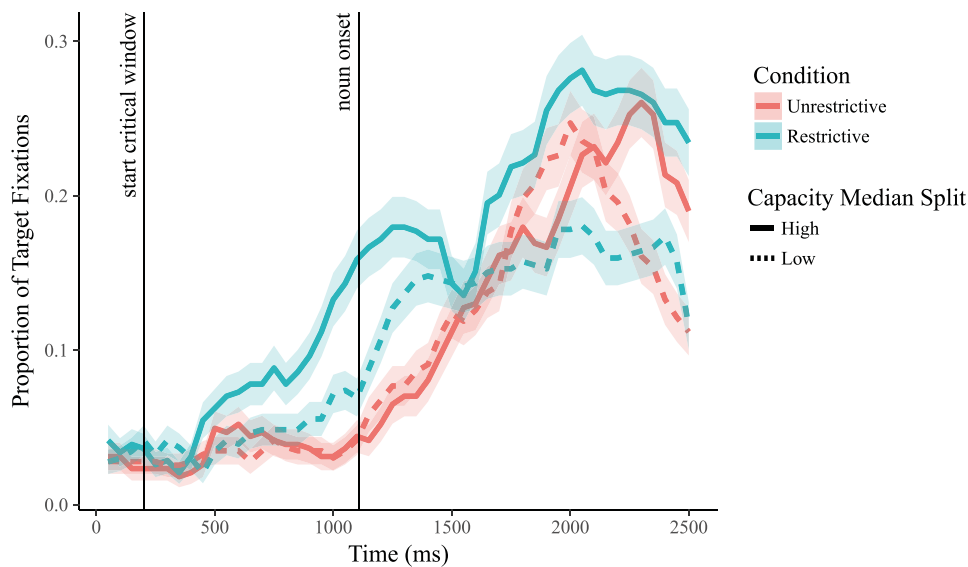
Figure 5. Mean proportions of target fixations for participants with low (<2.67 items) and high working memory capacity (>2.67 items), per condition. The analysis treated working memory capacity as a continuous variable and illustrates a significant effect of working memory on anticipatory eye movements for the restrictive condition ($p < .001$). Error clouds represent standard error.

movements for the restrictive condition ($p < .001$) was observed.

For illustrative purposes, we have categorised participants as having low working memory (<2.67) or high working memory (> 2.67). Figure 5 illustrates the fixation patterns of these two groups, per condition.

Modelling visual working memory capacity as a continuous variable in the GAMM model, the results suggest that participants with a higher working memory capacity showed anticipatory eye-movement behaviour earlier and more robustly compared to their peers with a lower working memory capacity.

Experiment 3: manipulating referent predictability

Experiments 1 and 2 have shown that participants show anticipatory eye movement behaviour during restrictive sentences even when faced with rich everyday scenes including 10 potential referent objects. This could be because every sentence spoken by the virtual agent concerned an object in the scene, a pattern that participants could have realised early in the experiment. Therefore, in Experiment 3 we introduced eight filler sentences per scene: Sentences that did not concern objects present in the scene. These filler sentences were similar to the restrictive/unrestrictive sentences in that they did concern an object (e.g. "People bring their own briefcase to work") and therefore participants were not able to detect whether a sentence spoken by the virtual agent was a filler or not until the object was named. Verbs were again

controlled to ensure that they were not predictive of objects already present in the scene. In sum, in this experiment only 50% of all sentences spoken concerned an object that the participants could fixate, and in only 25% of all sentences spoken, a unique target object could be anticipated given the verb.

Materials and methods

Participants
Twenty native speakers of Dutch (12 female, M_{age}: 22.7 years, SD_{age}: 2.11 years) were recruited from the Max Planck Institute for Psycholinguistics database. These participants had not participated in the previous experiments. Data from one additional participant was discarded due to insufficient accuracy of the eye-tracking data. The participants gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Materials
The same materials were used as described for Experiment 1. We created eight extra filler sentences per scene (64 extra sentences in total). Frequency of the verbs between the three conditions (restrictive, unrestrictive, and filler) was not significantly different (F(2,127) = 1.861, $p = .160$) although length was (F(2,127) = 8.12, $p < .001$). *Post-hoc* comparison showed that the filler verbs were significantly longer ($M = 7.59$ characters, $SD = 2.32$, Tukey's HSD, $p < .033$) compared to the restrictive ($M = 5.91$ characters, $SD = 1.69$) and unrestrictive ($M = 6.47$ characters, $SD = 1.78$) conditions.
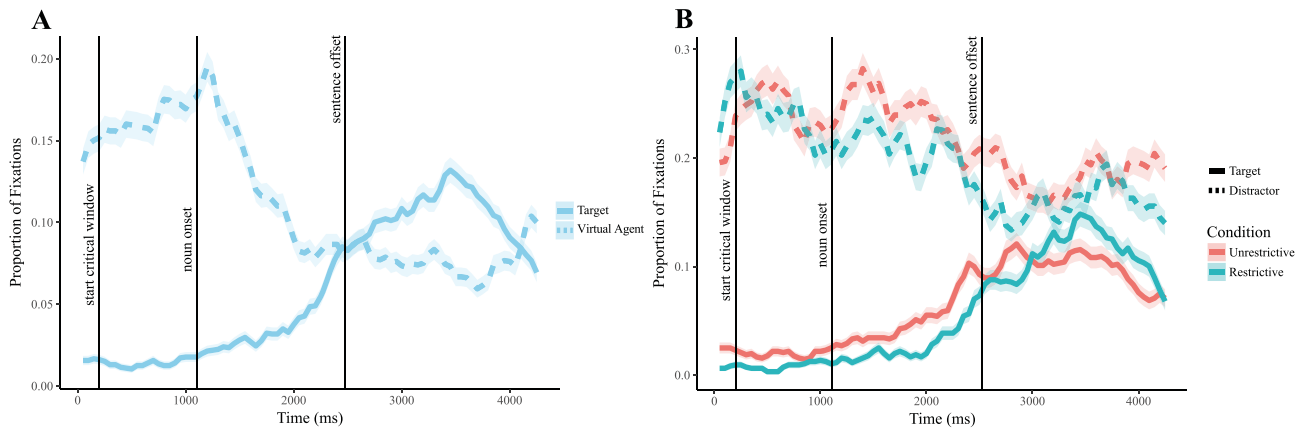
Figure 6. Mean proportions of fixations. A. To the target object and virtual agent. B. To the target and distractor objects shown per condition. Vertical lines indicate critical time points. 0 ms indicates verb onset, the label "start of critical window" is the start of the critical window (200 ms after verb onset). The main statistical analysis was performed on the interval between start of the critical window and noun onset. Error clouds indicate standard error.

Sentences from all three conditions were presented randomly in each scene. Due to the increase in sentences, the task took 15 min to complete.

Statistical analysis

The same statistical analysis was used as that described in Experiment 1. We removed 7.67% of all frames logged as object fixations and 2.32% of all frames logged as virtual agent fixations.

Results

Participants were able to accurately identify which objects the virtual agent had named and which she had not 82.68% of the time (SD: 8.67%) after the experiment. Therefore, we are confident that all participants listened to the virtual agent throughout the experiment.

Figure 6A illustrates the grand mean for this experiment. We again observed a robust increase in the proportion of looks to the target object after it is named. Figure 6B illustrates the proportion of looks per

Table 4. Summary of the generalised additive mixed model for changes in target fixations over time, per condition (restrictive versus unrestrictive sentences) for Experiment 3 (Less predictable input).

Parametric coefficients:					
	Estimate	SE	*t*-value	*p*-value	
Intercept	−1.77	0.04	−46.17	<.001	***
Condition	−0.05	0.03	−1.50	.133	
Smooth terms					
	edf	Ref.df	*F*-value	*p*-value	
Smooth for Time – Unrestrictive	1.96	2.33	1.00	.458	
Smooth for Time – Restrictive	1	1	19.84	<.001	***
Random effect for Subjects	41.25	179	1.30	<.001	***
Random effect for Sentences	157.12	574	2.40	<.001	***

*** < .001.
Effective degrees of freedom (edf), reference degrees of freedom (Ref.df).

condition. Table 4 reports the summary output of the GAMM analysis. For this analysis, the filler condition was not included as, by definition, there was no target object to fixate. After fitting the model, we determined an appropriate value for the AR1 parameter, in this case $\rho = -0.04$, to account for autocorrelation in the residuals (i.e. error).

We again observed an increase in the proportion of looks to the target object as a function of time for the restrictive ($p < .001$) but not the unrestrictive ($p = .458$) condition. The difference between the two conditions became significant between 710 and 900 ms after the start of the critical window. We observed no effect of condition on the proportion of distractor fixations for any of the three conditions (restrictive: $p = .551$; unrestrictive: $p = .646$; filler: $p = .716$).

Thus, we again observed significant anticipatory eye-movement behaviour for the restrictive condition, even though this behaviour was only efficient for 25% of the sentences heard. Appendix II presents a *post-hoc* analysis showing that the pattern of anticipatory eye movements changed over time during the course of the experiment, the most important finding being that no anticipatory eye-movements were observed during the last two scenes in the experiment. This suggests that previous experience did cause participants to stop producing anticipatory eye-movements, suggesting that participants stopped predicting the referent object within a single experimental session.

Experiment 4: objects outside the joint attentional space

This series of experiments is the first, to our knowledge, to include a dynamic visible source (i.e. an actual

speaker) for the sentences presented during look-and-listen VWP studies. The motivation behind including a virtual agent was part of the aim of Experiment 1: Making the VWP more realistic and therefore more ecologically valid. However, the inclusion of the virtual agent also presented an opportunity to investigate the role of joint attentional space in prediction studies, another component that, to our knowledge, has not been investigated previously. Therefore, this experiment was decided on *post hoc* and was not included in the pre-registration.

In order to manipulate joint attentional space, the same set-up as in Experiment 1 was used (6 objects, 4 sentences per scene), however the target objects were placed outside the joint attentional space between the virtual agent and the participant. As the virtual agent was always present in the middle screen, directly in front of the participant, *outside joint attentional space* was defined as the left or right screen (see Figure 1). Target objects were divided equally between these screens. The location of the distractor objects was unchanged compared to Experiment 1.

Materials and methods

Participants

Twenty native speakers of Dutch (17 female, M_{age}: 22.4 years, SD_{age}: 2.44 years) were recruited from the Max Planck Institute for Psycholinguistics database. These participants had not participated in the previous experiments. The data of 24 participants was recorded, but three participants were discarded due to insufficient accuracy of the eye-tracking data and one stated during the debrief stage that they did not understand the virtual agent properly (clarity rating < 3 out of 5).

The participants gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Materials

The same materials were used as described for Experiment 1. Only the location of the four target objects per scene was changed such that two were present on each of the peripheral screens.

Statistical analysis

The same statistical analysis was used as that described in Experiment 1. We removed 11.85% of all frames logged as object fixations and 5.18% of all frames logged as virtual agent fixations.

Results

Participants were able to accurately identify which objects the virtual agent had named and which she had not named 93.16% of the time (SD: 4.78%) after the experiment. Therefore, we are confident that all participants listened to the virtual agent throughout the experiment.

Figure 7A illustrates the grand mean for this experiment. We did not see the robust increase in the proportion of looks to the target object that we observed in the other experiments. In fact, the peak (0.13) occurred 976 ms after sentence offset. This suggests that some participants did search for the object, even after it was named; however, the majority did not. Only 37.42% of the target objects had been fixated by the participants before they were named by the virtual agent. However, even for these fixated objects (229 trials), we still observed no anticipatory behaviour (see
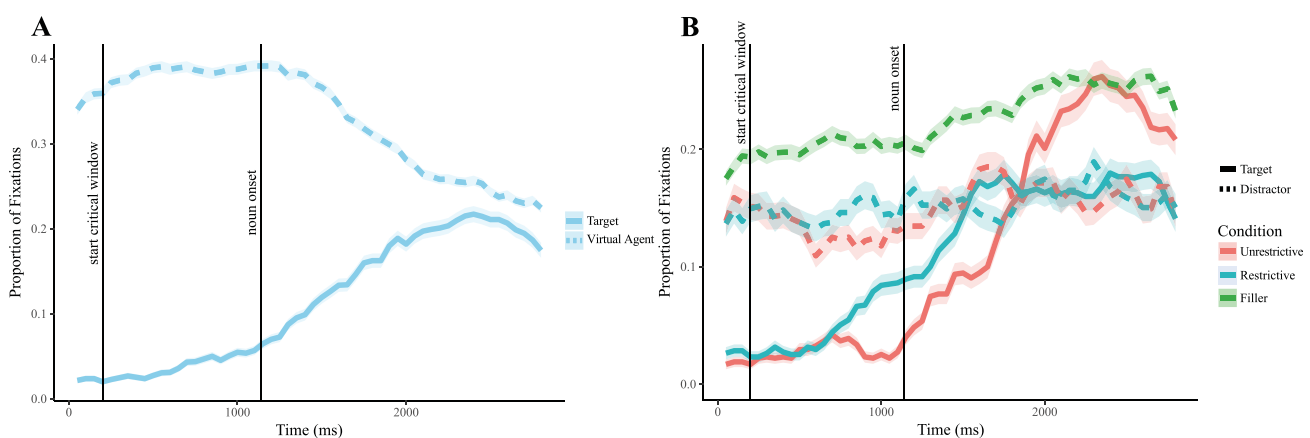
Figure 7. Mean proportions of fixations. A. To the target object and virtual agent. B. To the target and distractor objects shown per condition. Vertical lines indicate critical time points. 0 ms indicated verb onset, the label "start of critical window" is the start of the critical window (200 ms after verb onset). The main statistical analysis was performed on the interval between the start of the critical window and noun onset. Error clouds indicate standard error.

Table 5. Summary of the generalised additive mixed model for changes in target fixations over time, per condition (restrictive versus unrestrictive sentences) for Experiment 4 (Less attentional focus).

	Parametric coefficients:				
	Estimate	SE	*t*-value	*p*-value	
Intercept	−1.89	0.02	−89.79	<.001	***
Condition	0.02	0.02	1.38	.169	
Smooth terms					
	edf	Ref.df	*F*-value	*p*-value	
Smooth for Time – Unrestrictive	1	1	0.63	.429	
Smooth for Time – Restrictive	1	1	1.88	.170	
Random effect for Subjects	35.02	179	1.02	<.001	***
Random effect for Sentences	96.19	574	1.55	<.001	***

*** < .001.
Effective degrees of freedom (edf), reference degrees of freedom (Ref.df).

below). As the target objects were not present in the joint attentional space between the participant and the virtual agent (whereas they were in Experiment 1), they may have been encoded differently (or even not at all) and hence not considered as a potential target in the upcoming sentence.

Figure 7B illustrates the proportion of looks per condition. Table 5 reports the summary output of the GAMM analysis.

As illustrated in Figure 7B, there were no anticipatory looks to the target object during the critical window.

The results suggest that only objects located in the joint attentional space between the virtual agent and the participant are considered potential referents in the sentence. This conclusion is supported not only by the lack of anticipatory looks within the critical window, but also by a lack of increased target fixations after the sentence was spoken (there was only a 13% increase, compared to the >20% in the other three studies; see *Overall Results*). This suggests that hearing a restrictive verb does not initiate a visual search from the participant to look for an object that could fit that verb if none exist within the joint attentional space directly between speaker and addressee.

Overall results

For an overall comparison across experiments, Figure 8 illustrates the looks to the target object in the restrictive condition only, for each of the four experiments. We conducted a GAMM analysis to determine whether the observed anticipatory eye-movements in Experiments 2–4 were significantly different from those observed in Experiment 1, during the critical window. For this analysis, we created a difference smooth for Experiment 1 compared to each experiment individually (i.e. Experiment 1–2, Experiment 1–3, and Experiment 1–4).
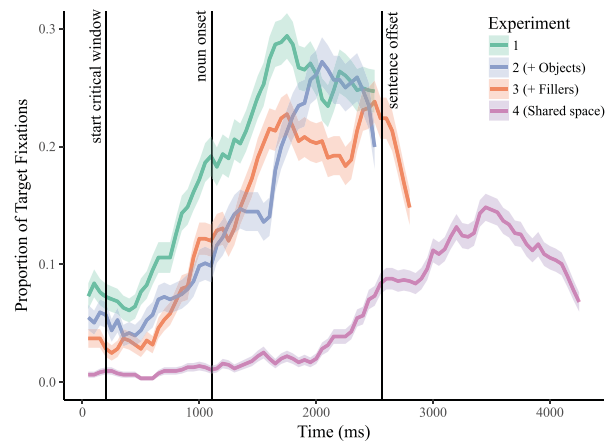
Figure 8. Mean proportions of fixations per experiment to the target object during the restrictive condition. Vertical lines indicate critical time points. 0 ms indicates verb onset, the label "start of critical window" is the start of the critical window (200 ms after verb onset). Experiment 1 induced significantly more fixations to the target object for both plots (*p* < .007). Error clouds represent standard error.

Therefore, the model not only analyses the difference between the curves, but also whether the steepness of the different curves is statistically the same (van Rij, 2015). As this involved separate models for each comparison, we have included the model outputs in the supplementary materials.

The results confirmed what is illustrated in Figure 8: The overall looks to the target object in the restrictive condition, during the critical window, were significantly higher for Experiment 1 compared to the other experiments (although this difference is marginal for Experiment 2 as *p* = .069). The models also allowed us to investigate differences in the steepness of the curves displayed in Figure 8. The models showed a significant difference (*p* = 0.48) for Experiment 1 compared to Experiment 2 (More Objects). In other words, participants fixated on the target objects less quickly (i.e. a less steep curve) for Experiment 2 compared to Experiment 1. There was no significant difference in the steepness of the curve for Experiment 1 compared to Experiment 3 (More Fillers; *p* = .135). In other words, even though there were more overall looks to the target object in Exp. 1 compared to Exp. 3, the speed at which the participants fixated on these objects was not significantly different between the two experiments. For Experiment 1 compared to Experiment 4 (Outside Shared Space), there was a significant difference in both the overall comparison of looks to target objects, as well as the speed at which this was done, providing more statistical evidence for a lack of anticipatory eye-movements in our fourth experiment.

Discussion

Prediction is commonly considered a central component of cognition. When processing incoming language input, the fact that we may predict upcoming words is generally used to explain why conversation in general and turn-taking in particular are often such efficient communicative activities. In four virtual reality experiments, we tested whether a well-established marker of linguistic prediction (i.e. anticipatory eye movements as observed in the visual world paradigm) replicated when increasing the naturalness of the paradigm by means of (i) immersing participants in naturalistic everyday scenes, (ii) increasing the number of potential referents present, (iii) modifying the proportion of predictable noun-referents in the experiment, and (iv) manipulating the location of referents inside and outside of the interlocutors joint attentional space. After all, previous experimental studies have mainly shown that listeners *can* predict, not necessarily that they *do* predict in naturalistic everyday settings.

In the current study we used anticipatory eye-movements as a measure of prediction (Altmann & Kamide, 1999). If participants predict the upcoming referent in naturalistic situations, we would expect robust anticipatory eye-movements towards the referent object after participants heard the restrictive verb (i.e. when the target object was identifiable based on the verb alone) compared to the unrestrictive verb (i.e. when the target object could not be identified based on verb information alone). Thus, if participants fixated the referent object significantly more and earlier after the verb was spoken but before the object was named in the restrictive condition, we would interpret that as evidence for predictive processing. This is exactly what we found in three of our four experiments. We were thus largely able to replicate the behaviour seen in traditional 2D (e.g. Altmann & Kamide, 1999) and 3D (Eichert et al., 2018) look-and-listen versions of the visual world paradigm.

Prediction in naturalistic environments

The main aim of the current study was to determine whether we predict in naturalistic everyday scenes by increasing the ecological validity of the visual world paradigm (VWP). In Experiment 1, we diverged from the traditional methodology by increasing the number of objects per scene (6 instead of 4), increasing the number of sentences per scene (4 instead of 1), and having a life-sized virtual agent deliver these sentences to the participants in a realistic 3D environment. Despite these changes, we were able to replicate anticipatory eye-movements in rich visual settings that included an actual, virtual speaker.

We do note, however, that the overall observed proportion of target fixations (~30%) in our study was lower compared to earlier studies (~90% in Altmann & Kamide, 1999) that used a computer monitor as their medium of stimulus display. They are, however, in line with an earlier study testing for anticipatory eye movements in virtual reality (~40%; Eichert et al., 2018). There are two complementary explanations for this difference in proportion of looks to the target. First, the mode of stimulus display (computer monitor versus CAVE) is different across studies. This means that in our study, visual objects were presented further away from the fovea in a visual context that was, purely in terms of display size, much larger than a simple computer monitor. Second, our stimulus environments (e.g. a forest, a living room) were visually significantly richer than those used in traditional studies (e.g. Altmann & Kamide, 1999). There is simply much more to be seen in our naturalistic setup compared to, for instance, the seminal study by Altmann and Kamide (1999). The fact that an increase in visual richness of a scene influences the overall proportion of looks to the target is confirmed by earlier work in which an increase in the set size of visible objects from 4 to 16 objects led to a decrease in the proportion of target fixations from 70% to ~40% (Sorensen & Bailey, 2007).

Nevertheless, as stated above, we were able to replicate anticipatory eye-movements in our more natural set-up, and thus for the remainder of the studies we continued to increase the number of objects (Experiment 2) and sentences (Experiment 3) to test whether participants still anticipated upcoming language input in these situations.

More potential referents

In Experiment 2, we increased the visual complexity of the scenes by increasing the number of objects from 6 to 10. Previous studies have tested the effect of an increased number of objects in 2D, traditional versions of the VWP (Andersson et al., 2011; Coco & Keller, 2015 versus Sorensen & Bailey, 2007). In these situations, however, participants were presented with cartoons or photorealistic 2D pictures and hence we questioned whether this was an ecologically valid representation of participant's behaviour when presented with more than the 4 items traditionally used in the look-and-listen VWP.

We nevertheless saw significant anticipatory eye movement behaviour, although it was significantly lower in Experiment 2 compared to Experiment 1 (*p*

= .006). This is a replication of the results seen in other VWP experiments (Andersson et al., 2011; Coco et al., 2016; Coco & Keller, 2015; Sorensen & Bailey, 2007) suggesting that the traditional VWP effects found do translate to more ecologically valid settings.

By increasing the number of objects to 10, we also taxed the visual working memory system, which may be the reason for the decrease in anticipatory eye-movement activity. Indeed, when we included the participant's working memory capacity estimate into the statistical model, we observed a significant mediation of working memory on anticipatory eye movement behaviour for the restrictive condition. This is in line with the claim that working memory is involved in mediating predictive language processes (see for review Huettig et al., 2011a). As working memory has a limited capacity, if it were involved in predictive processing, then we should see a decrease in predictive processing (in our case, anticipatory eye-movement behaviour) when the number of potential referents in a scene increases.

The proposal for a role of working memory in anticipating linguistic information is not new (Huettig et al., 2011a; Knoeferle & Crocker, 2007), although only few studies have attempted to provide empirical evidence to support this proposal. Huettig and Janse (2016) found a positive correlation between the ratio of target-distractor object looks and a working memory construct score, such that participants with a higher working memory construct score showed a stronger prediction tendency. A recent study provided more causal evidence by demonstrating that participants showed reduced anticipatory looks to the referent object if they were required to simultaneously remember five words (Ito et al., 2018). We build upon these earlier findings by providing a direct link between a participant's working memory capacity and their anticipatory eye-movement behaviour.

Another way to interpret our data is that participants with higher working memory capacity *predict* better, as their anticipatory eye movements occurred earlier and more frequently compared to their lower working memory capacity peers. However, this conclusion is hard to fully support given our data, and also calls into question the role working memory plays in predictive processes. Participants with higher working memory capacity may be able to encode and link the objects and their locations better, which does not necessarily suggest that they are better at predicting. It could be that all participants predicted equally well, but their working memory capacity limits how many potential objects they can retain to base their predictions on. Hence, although our study suggests that working

memory may play an important role in predictive processes, exactly what this role entails needs to be explored further.

Overall, the results of Experiment 2 showed that participants still show anticipatory eye-movement behaviour, arguably reflecting predictive language processing, even with an increasing number of items.

Less predictable input

In Experiment 3, we manipulated referent predictability by having only 25% of the sentences contain a verb that could be used to predict the specific upcoming referent. The remaining sentences were either unrestrictive (25%) or did not refer to an object present in the scene (50%). Several theories propose that prediction is supported by a statistical learning mechanism. The idea is that we anticipate upcoming linguistic information based on past experiences (Chang, 2002; Dell & Chang, 2014). If this were the case, then one would expect that participants stop exhibiting anticipatory eye-movements as this would be inefficient given the statistical probability of an object being either present in the scene or predictable given the verb. Although overall there was anticipatory eye-movement behaviour, when we conducted a *post hoc* analysis of the first two scenes in the experimental session compared to the last two scenes, we *did* observe a significant decrease in anticipatory eye-movement behaviour at the end of the experiment. Hence, if predictive behaviour turns out to be ineffective, for instance because only 25% of the target items are predictable, participants may stop predicting upcoming target referents over time. This has interesting implications for the role of predictive processing in everyday settings, as we are often confronted with language input that does not relate to objects that are immediately present. These results indicate that prediction might not occur under all circumstances in everyday conversation and confirm earlier suggestions that listeners may adapt their predictive behaviour as a function of distributional properties of recently received linguistic input (Pickering & Gambi, 2018; see also Havron et al., 2019; Yurovsky et al., 2017).

Less obvious attentional focus

The inclusion of a virtual speaker in our experiments introduced the concept of joint attentional space in research using the look-and-listen VWP. In all four experiments reported here, participants faced the virtual speaker, thereby to some extent mimicking everyday interaction in which interlocutors often form a conversational dyad. In naturally occurring communication,

interlocutors indeed typically use their bodies to separate their *joint attentional space* of engagement from the larger outside world (Kendon, 1977, 1990b, 1992; Scheflen & Ashcraft, 1976). Certain objects speakers refer to may be present inside this joint attentional space, whereas others may be located outside of it in visual periphery (Peeters et al., 2015), and interlocutors keep track of whether they are attending to something in common (Staudte et al., 2014; Tomasello, 1995). The results of Experiment 4 showed that the presence of anticipatory eye movements depends on whether target objects are located inside or outside the interlocutors' joint attentional space. Overall, no anticipatory eye movements were observed. Moreover, participants did not search for the objects after they were named, suggesting that hearing an object named that was not in the joint attentional space does not initiate a visual search to find the object.

As the target objects were located on the side screens of the CAVE environment in Experiment 4 (see Figure 1), participants had to slightly turn their head to fixate these objects. Some of them did so already when the visual scene was presented but before any sentence was uttered. Even when focusing solely on these participants, we found no anticipatory behaviour to the objects when the sentence was spoken, suggesting they did not consider the objects outside the joint attentional space as potential targets. To address a *post hoc* theory, we conducted an exploratory analysis to determine whether participants may have noticed, during the course of the experiment, that the target objects were only located in the peripheral left and right side of their vision. If they did, this would mean that there should be an increase in anticipatory behaviour over the course of the experiment. We therefore compared looks to target objects in the first two scenes (8 trials in total per participant) to looks to target objects in the last two scenes (8 trials in total per participant). For both the first and the last two scenes, no anticipatory behaviour ($p = .741$ versus $p = .070$) was observed, although the GAMM results for the last two scenes do suggest a trend in the predicted direction.

Conclusion

Prediction is undoubtedly a central component of human cognition. Current theories of prediction involve the creation of an internally generated model of anticipated upcoming information. In the case of language processing, the actually encountered linguistic information is arguably compared against a forward model of anticipated linguistic information (Pickering & Gambi, 2018; Pickering & Garrod, 2007, 2013) and any

prediction error is used as a learning mechanism that influences future predictions (Dell & Chang, 2014). Theories of prediction in the domain of language have mainly been built on the basis of empirical data obtained from participants sitting in front of computer monitors looking at stimuli that are relatively poor 2D abstractions of everyday objects. In the current study, we investigated whether a robust marker of prediction – anticipatory eye movements as observed in the visual world paradigm – would be observed in a variety of rich, everyday environments that included a life-size speaker.

Do we predict upcoming linguistic content in rich, naturalistic environments? The evidence provided here is mixed. On the one hand, we observed robust anticipatory eye movements in naturalistic scenes, even when these scenes contained a relatively large number of objects and a relatively small number of sentences that allowed for a predicted noun-referent to be confirmed by the speaker's unfolding speech. On the other hand, however, our study sheds light on new influences on predictive processing in the VWP. For example, when the predictability of sentence endings was low, participants over the course of the experiment stopped anticipating which referent object was going to be named by the virtual agent. This finding confirms a recent suggestion that predictive behaviour may be circumvented when distributional properties of recent input deem it unhelpful (Brothers et al., 2017). Moreover, the well-established effect of prediction also disappeared when referent objects were placed outside the joint attentional space shared by speaker and participant. Together, these findings suggest limits to the generalizability of earlier experimental findings to naturalistic environments. As such, theories of the mind that involve prediction as a key feature of human cognition should take into consideration how the spatial, linguistic, and social context may modulate the waxing and waning of predictive behaviour.

In our study we focused on saccades to predetermined regions of interest. However, we tapped into linguistic predictions via their consequences for anticipatory eye-movements. In principle, predictions could be made even without consequences for eye-movements. Hence, we are not suggesting that participants would not predict in situations where referents are not available (e.g. while listening to the radio). Nevertheless, in many daily life situations linguistic utterances refer to the here and now of the environment in which the speaker and listener find themselves. In these cases, there seems to be a tight link between the utterances and the scenes that they refer to. Under those circumstances anticipatory eye-movements might be a

way to rapidly negotiate between language and the world. Methodologically, however, we hope that our study paves the way for future studies of human predictive behaviour in which ecological validity and experimental control go hand in hand.

Note

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Evelien Heyselaar ⬦ http://orcid.org/0000-0003-1138-1331

References

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The "blank screen paradigm". *Cognition*, *93*(2), B79–B87. https://doi.org/10.1016/j.cognition.2004.02.005

Altmann, G. T. M. (2011). The mediation of eye movements by spoken language. In   S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 979–1004). Oxford University Press.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Alvarez, G., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*(2), 208–216. https://doi.org/10.1016/j.actpsy.2011.01.007

Baayen, R., van Rij, J., Cecile, D., & Wood, S. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylan, & D. Geeraerts (Eds.), *Mixed effects regression models in Linguistics* (pp. 49–69). Springer.

Baddeley, A. (1998). Working memory. *Académie Des Sciences*, *321*, 167–173. https://doi.org/10.1016/S0764-4469(97)89817-4

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer* (5.1.05).

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, *32*(4), 643–684. https://doi.org/10.1080/03640210802066816

Chang, F. (2002). Symbolically speaking : a connectionist model of sentence production. *Cognitive Science*, *26*(26), 609–651. https://doi.org/10.1207/s15516709cog2605_3

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coco, M. I., & Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive Processing*, *16*(2), 131–150. https://doi.org/10.1007/s10339-014-0642-0

Coco, M. I., Keller, F., & Malcolm, G. L. (2016). Anticipation in real-world scenes: The role of visual context and visual memory. *Cognitive Science*, *40*(8), 1995–2024. https://doi.org/10.1111/cogs.12313

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

Eichert, N., Peeters, D., & Hagoort, P. (2018). Language-driven anticipatory eye movements in virtual reality. *Behavior Research Methods*, *50*(3), 1102–1115. https://doi.org/10.3758/s13428-017-0929-z

Fodor, J. A. (1983). *The modularity of mind*. MIT Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Hari, R., Henriksson, L., Malinen, S., & Parkkonen, L. (2015). Centrality of social interaction in human brain function. *Neuron*, *88*(1), 181–193. https://doi.org/10.1016/j.neuron.2015.09.022

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall/CRC Press.

Havron, N., de Carvalho, A., Fiévet, A. C., & Christophe, A. (2019). Three-to four-year-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child Development*, *90*(1), 82–90. https://doi.org/10.1111/cdev.13113

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

Heyselaar, E., Hagoort, P., & Segaert, K. (2015). In dialogue with an avatar, language production is identical compared to dialogue with a human partner. *Behavior Research Methods*, *1*, 15. https://doi.org/10.3758/s13428-015-0688-7

Heyselaar, E., Johnston, K., & Paré, M. (2011). A change detection approach to study visual working memory of the

macaque monkey. *Journal of Vision*, *11*(3), 1–10. https://doi.org/10.1167/11.3.11

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(9), 1352–1374. https://doi.org/10.1037/xlm0000388

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–96. https://doi.org/10.1038/scientificamerican0960-88

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. https://doi.org/10.1080/23273798.2015.1047459

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*(4), 460–482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011a). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013

Huettig, F., Rommers, J., & Meyer, A. S. (2011b). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, *21*(2), 251–264. https://doi.org/10.1017/S1366728917000050

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kendon, A. (1977). Some functions of gaze-direction in two-person conversation. *Studies in the Behaviour of Social Interaction*, 13–51.

Kendon, A. (1990a). *Conducting interaction. Patterns of behavior in focused encounters*. Cambridge University Press.

Kendon, A. (1990b). Spatial organization in social encounters. In A. Kendon (Ed.), *Studies in the behaviour of social interaction* (pp. 179–208). Peter de Ridder Press.

Kendon, A. (1992). The negotiation of context in face-to-face interaction. In A. Duranti, & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 323–334). Cambridge University Press.

Knoeferle, P. (2015). Language comprehension in rich non-linguistic contexts: Combining eye tracking and event-related brain potentials. In R. M. Willems (Ed.), *Cognitive neuroscience of natural language use* (pp. 77–100). Cambridge University Press.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science*, *30*(0), 481–529. https://doi.org/10.1207/s15516709cog0000_65

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*(4), 519–543. https://doi.org/10.1016/j.jml.2007.01.003

Kochari, A. R., & Flecken, M. (2018). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *PsyArXiv Preprints*, https://doi.org/10.17605/OSF.IO/9NPUE

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*(1), 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. R., & Jaeger, T. F. (2017). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what Lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the Brain: Using our past to generate a future* (pp. 190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380. https://doi.org/10.3758/BF03206780

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Wolfsthurn, V. G. Z., Bartolozzi, S., Kogan, F., Ito, V., Mézière, A., Barr, D., Rousselet, D. J., Ferguson, G. A., Busch-Moreno, H. J., Fu, S., Tuomainen, X., Kulakova, J., Husband, E., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, 1–24. https://doi.org/10.7554/eLife.33468

Pan, X., & Hamilton, A. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395–417. https://doi.org/10.1111/bjop.12290

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00660

Peeters, D. (2018). A standardized set of 3-D objects for virtual reality research and applications. *Behavior Research Methods*, *50*(3), 1047–1054. https://doi.org/10.3758/s13428-017-0925-3

Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic Bulletin and Review*, *26*, 894–900. https://doi.org/10.3758/s13423-019-01571-3

Peeters, D., & Dijkstra, T. (2018). Sustained inhibition of the native language in bilingual language production: A virtual reality approach. *Bilingualism: Language and Cognition*, *21*(5), 1035–1061. https://doi.org/10.1017/S1366728917000396

Peeters, D., Hagoort, P., & Özyürek, A. (2015). Electrophysiological evidence for the role of shared space in online comprehension of spatial demonstratives. *Cognition*, *136*, 64–84. https://doi.org/10.1016/j.cognition.2014.10.010

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). *An integrated theory of language production and comprehension*. 329–392. https://doi.org/10.1017/S0140525X12001495.

Porretta, V., Kyröläinen, A., van Rij, J., & Järvikivi, J. (2017). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. *International Conference on Intelligent Decision Technologies*, 268–277. https://doi.org/10.1007/978-3-319-92028-3

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. https://doi.org/10.1016/0010-0277(89)90014-0

R Core Development Team. (2011). *R: A language and environment for statistical computing*.

Salthouse, T. A., McGuthry, K. E., & Hambrick, D. Z. (1999). A Framework for analyzing and interpreting differential aging patterns: Application to three measures of implicit learning. *Aging, Neuropsychology, and Cognition*, *6*(1), 1–18. https://doi.org/10.1076/anec.6.1.1.789

Scheflen, A. E., & Ashcraft, N. (1976). *Human territories: How we behave in space-time*.

Sorensen, D. W., & Bailey, K. G. D. (2007). The world is too much: Effects of array size on the link between language comprehension and eye movements. *Visual Cognition*, *15*(1), 112–115. https://doi.org/10.1080/13506280600975486

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Visual Cognition*, *20*(8), 922–946. https://doi.org/10.1080/13506285.2012.715599

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*(1), 317–328. https://doi.org/10.1016/j.cognition.2014.06.003

Tomasello, M. (1995). Joint attention as social cognition. In D. Moore, & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Lawrence Erlbaum Associates.

Tromp, J., Peeters, D., Meyer, A. S., & Hagoort, P. (2018). The combined use of virtual reality and EEG to study language processing in naturalistic environments. *Behavior Research Methods*, *50*(2), 862–869. https://doi.org/10.3758/s13428-017-0911-9

van den Brink, D., Brown, C. M., & Hagoort, P. (2000). *Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects*. 967–985.

van Rij, J. (2015, March). *Overview GAMM analysis of time series data*. Https://Jacolienvanrij.Com/Tutorials/GAMM.Html

van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs* (R package version 2.3).

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436

Willems, R. M. (2015). *Cognitive neuroscience of natural language use*. Cambridge University Press.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. https://doi.org/10.1016/S0960-9822(01)00432-8

Wood, S. (2017). *Package "mgcv"*.

Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input in a noisy channel. *Psychological Science*, *28*(1), 132–140. https://doi.org/10.1177/0956797616668557