LANGUAGE, COGNITION AND NEUROSCIENCE
2022, VOL. 37, NO. 4, 481–508
https://doi.org/10.1080/23273798.2021.1994621

Routledge
Taylor & Francis Group

REGULAR ARTICLE

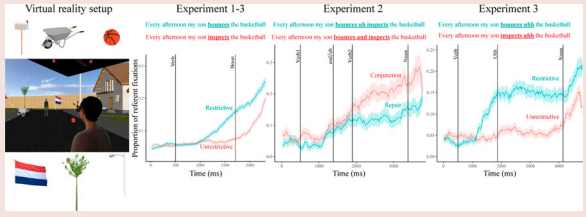
Prediction of upcoming speech under fluent and disfluent conditions: eye tracking evidence from immersive virtual reality

Eleanor Huizeling [a], David Peeters[a,b,c] and Peter Hagoort[a,c]

[a]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [b]Department of Communication and Cognition, TiCC, Tilburg University, Tilburg, The Netherlands; [c]Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

ABSTRACT

Traditional experiments indicate that prediction is important for efficient speech processing. In three virtual reality visual world paradigm experiments, we tested whether such findings hold in naturalistic settings (Experiment 1) and provided novel insights into whether disfluencies in speech (repairs/hesitations) inform one's predictions in rich environments (Experiments 2–3). Experiment 1 supports that listeners predict upcoming speech in naturalistic environments, with higher proportions of anticipatory target fixations in predictable compared to unpredictable trials. In Experiments 2–3, disfluencies reduced anticipatory fixations towards predicted referents, compared to *conjunction* (Experiment 2) and *fluent* (Experiment 3) sentences. Unexpectedly, Experiment 2 provided no evidence that participants made new predictions from a repaired verb. Experiment 3 provided novel findings that fixations towards the speaker increase upon hearing a *hesitation*, supporting current theories of how hesitations influence sentence processing. Together, these findings unpack listeners' use of visual (objects/speaker) and auditory (speech/disfluencies) information when predicting upcoming words.



ARTICLE HISTORY
Received 14 April 2021
Accepted 10 October 2021

KEYWORDS
Prediction; Disfluencies; Visual world paradigm; Virtual reality; Eye tracking

1. Introduction

A longstanding question remains as to how spoken language is processed so efficiently and so rapidly. An increasing body of literature supports that this fast processing is assisted by the online prediction of upcoming linguistic input (Kuperberg, 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2007). Although it is clear that listeners are indeed able to predict upcoming speech, the majority of the supporting evidence derives from relatively artificial laboratory experiments that could introduce atypical processing strategies. Furthermore, natural communication occurs in rich and dynamic environments, and contains many subtleties that could be used to inform the prediction of upcoming speech. For example, natural speech contains frequent disfluencies (e.g. hesitations and repairs) and is often produced by a visible interlocutor. The current work aimed to provide novel insights regarding the degree to which subtle cues in speech (specifically hesitation and repair disfluencies) inform one's predictions in naturalistic, everyday environments in which listeners are directly addressed by a visible speaker.

Until recently, much of the speech prediction literature has been related to investigating whether listeners predict, rather than defining the extent that subtle cues in speech are utilised to inform predictions. For example, the visual world paradigm (VWP), in which participants view images and listen to sentences while their eye gaze is recorded, has shown that, when the upcoming content of a sentence is highly constrained, participants make anticipatory eye movements towards a referent before hearing its associated noun (Altmann & Kamide, 1999). This pattern of results has been widely replicated, and suggests that indeed listeners may predict upcoming

CONTACT Eleanor Huizeling eleanor.huizeling@mpi.nl Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen 6525XD, The Netherlands
Supplemental data for this article can be accessed doi:10.1080/23273798.2021.1994621

linguistic input (Altmann & Kamide, 1999, 2007; Coco et al., 2016; Kamide et al., 2003; Kukona et al., 2014; Rommers et al., 2013). But to what extent do listeners predict upcoming words under natural circumstances?

In contrast to typical, carefully controlled, experimental stimuli, natural speech contains approximately six disfluencies (e.g. hesitations/repairs/repetitions/silent pauses) in every one hundred words (Bortfeld et al., 2001). Rather than disfluencies hindering speech processing, as one may intuitively think, information spoken after a disfluency has been shown to be better recalled (Collard et al., 2008; Corley et al., 2007; Fraundorf & Watson, 2011; MacGregor et al., 2010) and is responded to faster when participants must respond to a target word (Fox Tree & Schrock, 1999) or to instructions (Corley & Hartsuiker, 2011). It is debated as to whether such benefits are simply due to the temporal delay providing more processing time (Corley & Hartsuiker, 2011; Wester et al., 2015), or whether disfluencies in speech enhance lexical processing, for example through increased attention (Collard et al., 2008), improved chunking of information (Fraundorf & Watson, 2011, 2014), or by providing information about the upcoming content. The latter proposition suggests that listeners learn from the non-arbitrary distribution of disfluencies in speech, such as the tendency of disfluencies to occur when the speaker is under higher cognitive load or high uncertainty, speaking about topics with more expressive flexibility, or when trying to retrieve low-frequency words (Bortfeld et al., 2001; Brennan & Williams, 1995; Fraundorf & Watson, 2014; Schachter et al., 1991; Smith & Clark, 1993). Listeners are assumed to apply knowledge about when disfluencies typically occur to anticipate upcoming speech and prepare for comprehension to be demanding (Bosker, 2014; Bosker et al., 2019). As such, under natural circumstances, listeners could use subtle disfluencies in perceived speech to inform their predictions. In the following paragraphs we continue to discuss the current theories of how specifically *repair* and *hesitation* disfluencies influence sentence processing.

1.1. Repair disfluencies

Evidence from the VWP supports that *repair* disfluencies can influence ongoing sentence processing beyond the effects of increased processing time. Corley (2010) demonstrated that, in predictable sentences such as "the boy will eat the cake", after the highly constraining verb ("eat") was marked as incorrect and amended to a non-constraining verb (e.g. "eat uh move"), the proportion of anticipatory fixations towards the predicted referent (i.e. the cake) decreased. Listeners, therefore, seem to be able to rapidly update their predictions about upcoming speech upon hearing a *repair* disfluency.

In a different line of VWP experiments, it has been argued that listeners can use the content of a repair disfluency to predict the upcoming amended speech. Such a proposition is consistent with the suggestion that one uses semantic priors to interpret the intended meaning of error-filled, semantically implausible sentences (Gibson et al., 2013; Levy, 2008). Specifically, there is evidence that information from an erroneous noun can inform listeners' predictions about the target noun. For example, listeners' fixations have been shown to shift to semantically or phonologically related images relative to the erroneous noun (Karimi et al., 2019; Lowder & Ferreira, 2016). It is thought that the error is interpreted as an intrusion from a semantically or phonologically related item and the prediction is updated accordingly. Thus, rather than the properties of a verb imposing constraints on the upcoming linguistic content, as in Corley (2010), constraints derive from the expected cause of the erroneous noun. The authors demonstrated that the effect went beyond any lexical priming observed with a silent pause or the conjunction *and also* (Karimi et al., 2019; Lowder & Ferreira, 2016). For example, Lowder and Ferreira (2016) demonstrated that, after hearing a repair disfluency "salt, uh I mean … ", eye movements towards the semantic competitor *pepper* increased before the onset of the repaired noun, more so than towards a distractor object (e.g. *milk*). Crucially, the increase in semantic competitor fixations was greater after hearing a repair than after hearing a coordination condition "and also … ", where semantic competitor fixations were expected to be related to priming alone. The increased proportion of looks towards the semantic competitor in the repair condition were thought to reflect the listener interpreting the error as an intrusion from the competitor. In contrast, looks towards the critical competitor in the coordination condition were thought to be driven by lexical priming alone. However, an alternative explanation could be that, upon hearing the onset of a repair, listeners actively inhibited their prediction and attention was automatically drawn to the next most activated lexical item, i.e. the erroneous noun's semantic and phonological competitors, through semantic priming (Collins & Loftus, 1975). The effect of semantic priming might be expected to be stronger after inhibiting the first noun in the repair condition compared to without such inhibition in the coordination condition. Although the source of semantic priming (i.e. the noun) may be inhibited, residual activation of competing lexical items could remain. The extent to which the content of an error can inform future predictions beyond the influence of lexical priming, therefore, remains unclear.

1.2. Hesitation disfluencies

Further evidence from the VWP supports the idea that more subtle disfluencies such as *hesitations* and silent pauses can also inform predictions. Upon hearing a hesitation, participants have been shown to be more likely to predict a less frequent or discourse new noun, rather than a highly frequent or discourse given noun (Arnold et al., 2003, 2004; Arnold & Tanenhaus, 2011). It is thought that the listener interprets the hesitation as a cue that the speaker is having difficulty in retrieving a word and utilises this cue to inform their predictions about what the speaker is trying to say. A difficulty with retrieval could imply that the upcoming utterance is unlikely to be the most predictable outcome, but instead may be a less frequent or discourse new item, or a phonological or semantic competitor. Indeed, speakers tend to produce more disfluencies when they are uncertain in their response, which in turn results in a reduction of the listener's confidence in what the speaker is saying (Brennan & Williams, 1995; Lowder & Ferreira, 2019; Smith & Clark, 1993). In contrast, hesitations do not seem to affect the prediction of upcoming speech when there is an alternative explanation for the hesitation, for example, if the participant is a non-native speaker (Bosker et al., 2014) or has object agnosia (Arnold et al., 2007).

An alternative theoretical outlook proposes that, rather than informing predictions, hesitations cue the listener's attention. Disfluencies often occur before the utterance of complex information (Fraundorf & Watson, 2014) and may signal to listeners that they should pay close attention to upcoming input to facilitate comprehension. Bosker (2014) highlighted that enhanced attention to upcoming speech could play a role in two ways, either as an automatic cognitive process as the delay triggers attention, or as an informed process with strong priors (i.e. that they need to pay attention). In support of an enhanced attention account, participants have been shown to have an improved memory for information spoken after a disfluency (Collard et al., 2008; Corley et al., 2007; Fraundorf & Watson, 2011; MacGregor et al., 2010) and were faster at responding on a picture recognition task when the spoken picture name was preceded by a disfluency (Corley & Hartsuiker, 2011).

1.3. Natural language processing

During natural language processing, one rarely hears sentences in the absence of visual input, whether in the form of visual cues from the speaker, or contextual information from the environment. VWP research has been pioneering in going beyond the study of sentence processing in the absence of visual information, to investigate how visual and linguistic information interact (Allopenna et al., 1998; Altmann & Kamide, 1999, 2007; Huettig et al., 2011; Kamide et al., 2003; Magnuson, 2019; Teruya & Kapatsinski, 2019; Weber et al., 2006). However, in contrast to the rich natural world, the VWP often involves the participant viewing four simple images. Although some VWP experiments have used richer stimuli (Andersson et al., 2011; Ferreira et al., 2013), and anticipatory fixations have also been observed when presenting more complex displays, such as simple scenes (Coco et al., 2016), photographs (Staub et al., 2012) or larger array sizes (Sorensen & Bailey, 2007), such two-dimensional (2D) displays lack important elements of the rich natural world. For example, the real world is three-dimensional (3D), visually complex, and dynamic. Crucially, one has a feeling of presence within the environment – a feeling of a shared space with the people and objects in the environment. Furthermore, increasing the visual complexity of displays resulted in a much lower proportion of anticipatory fixations reaching the target object (Coco et al., 2016; Sorensen & Bailey, 2007) compared to previous work, which raises questions as to whether predictive eye movements would disappear completely when increasing the visual complexity to real-life settings.

In addition to the 2D visual displays, in a typical VWP participants are only presented with one sentence per visual display, which may lead to participants attempting to deduce the experimental manipulation. Of a particular concern is "good behaviour" in participants, where participants alter their behaviour to react in the way they believe the experimenter expects them to. It is a current topic of debate as to whether the VWP may similarly encourage altered processing strategies, both in instruction based and look-and-listen VWPs (Magnuson, 2019). Additionally, the listener typically can only hear and not see the speaker, which removes many of the fundamental properties of spoken communication between two or more interlocutors (Redcay & Schilbach, 2019; Schilbach et al., 2013). The aforementioned methodological limitations raise concerns about how well we can generalise findings from the VWP to behaviour in everyday environments.

The importance of studying language processing embedded within naturalistic contexts is becoming increasingly salient (Hasson et al., 2018). Advances in technology mean that we can now reach a balance between experimental control and ecological validity by bringing psychological and psycholinguistic experiments into an enriched context in virtual reality (VR; Eichert et al., 2018; Heyselaar et al., 2020; Pan &

Hamilton, 2018; Parsons, 2015; Peeters, 2019; Tromp et al., 2018). Recent work from our VR laboratory supports that prediction does indeed occur in naturalistic environments (Eichert et al., 2018; Heyselaar et al., 2020). The authors moved beyond the traditional VWP by making it more naturalistic in a number of ways. Firstly, referents were embedded into 3D scenes in VR, rather than being displayed on a 2D computer screen (Eichert et al., 2018; Heyselaar et al., 2020). Secondly, sentences were spoken to the participant by a virtual agent (Heyselaar et al., 2020), which avoided sentences being spoken by a disembodied voice and reintroduced the crucial communicative component of spoken language (Redcay & Schilbach, 2019; Schilbach et al., 2013). Finally, more than one sentence was spoken per scene (Heyselaar et al., 2020), which reduced the salience of the experimental manipulation. The authors found that, in the predictable sentence condition, anticipatory eye movements were made towards objects that were embedded in these rich virtual environments (Eichert et al., 2018; Heyselaar et al., 2020), even when increasing the number of distractor objects in the scene and reducing the proportion of predictable sentences (Heyselaar et al., 2020). Findings, therefore, corroborate the previous literature and demonstrate that we not only *can* predict but *do* predict in naturalistic settings.

1.4. The current work

It is clear from the evidence outlined so far that disfluencies in speech can influence the listener's predictions, possibly as they are indicative of high processing demands in the speaker while preparing the production of upcoming speech. However, there are two outstanding theoretical questions. Firstly, it remains unclear whether listeners simply inhibit their prediction upon hearing a repair disfluency, or whether they additionally use information from the error to inform further predictions. Previous investigations do not clearly determine whether shifts in gaze towards semantic and phonological competitors are driven by lexical priming or informed priors (Karimi et al., 2019). Secondly, although it remains under debate as to whether shifts in eye gaze upon hearing a hesitation are due to strong priors regarding the content of post-hesitation speech, or enhanced attention alone, competing theoretical accounts seem to agree that there would be a shift in eye gaze towards the speaker upon hearing a hesitation. Under a prediction account, a hesitation signals that the upcoming speech is difficult for the speaker to produce. Considering that the semantic features of items in the speaker's field-of-view should already be activated, it could be argued that these items are easier for the speaker to produce (compared to items outside of the speaker's field of view), and are therefore unlikely to be the (post-hesitation) target referent. In such a scenario, the listener's predictive eye gaze would be expected to move towards the speaker to wait for the sentence to become disambiguated. Under an attentional account, hesitations are thought to enhance attention through either an automatic shift in attention or in preparation to hear complex input (Bosker, 2014). Both attentional mechanisms would be expected to result in a shift in eye gaze towards the speaker as the listener directs their attention to the upcoming speech. However, to our knowledge, looks towards the speaker upon hearing a hesitation disfluency have not previously been measured.

In addition to the outstanding theoretical questions, the same intrinsic restrictions as typical VWPs (as outlined above) are present in the aforementioned literature. To reiterate, presenting simple visual displays and a single sentence per display (spoken by a disembodied voice) could have led to altered processing strategies compared to real life environments. Given the increasing relevance of studying naturalistic language processing, it is important to investigate whether disfluencies in speech similarly affect predictive eye movements when embedded in more naturalistic environments.

In a series of three VR experiments, the current work aimed to (a) confirm the extent to which listeners predict upcoming speech in naturalistic environments (Experiment 1), and (b) investigate to what degree disfluencies (repairs and hesitations) influence predictions of upcoming speech, thereby informing different theoretical accounts (Experiments 2 and 3). In all three experiments, participants listened to predictable and unpredictable sentences that were spoken by a virtual agent during a virtual tour of eight scenes (e.g. an office, a living room, a canteen) while their eye gaze was being recorded. Firstly, we aimed to replicate the findings of Heyselaar et al. (2020) with new sentence stimuli and object combinations, together with an increased number of sentences per condition (Experiment 1). In doing so we tested the generalisability and reproducibility of Heyselaar et al.'s (2020) findings of increased anticipatory fixations towards a referent in highly constrained sentences compared to those with low sentence constraints. Secondly, we aimed to test how repairs influence predictive eye movements compared to sentences containing a conjunction (Experiment 2). Finally, in Experiment 3 we investigated the influence of hesitations on predictive eye movements.

The current work achieved three notable methodological improvements compared to previous work, which allowed us to provide novel contributions to current theories of the role of disfluencies in the prediction of upcoming speech. Firstly, VR provided a platform in which we could embed spoken linguistic stimuli in a naturalistic

context while retaining experimental control (Pan & Hamilton, 2018; Parsons, 2015; Peeters, 2019). Critically, this allowed us to investigate fixations not only towards objects within a rich and dynamic environment, but also towards the speaker, i.e. the virtual agent. Doing so enabled (a) the investigation of fixations towards critical distractor objects and the virtual agent upon hearing a *repair* disfluency, thereby providing information about whether the listener predicts a new item based on the repair, or whether their attention moves towards the speaker, and (b) the novel investigation of fixations towards the virtual agent upon hearing a *hesitation* disfluency, thereby testing whether it is possible to observe a fundamental property of current accounts of the effect of hesitations on the listener's sentence processing, i.e. a shift in eye gaze towards the speaker. Secondly, we increased both the number of trials per condition and the number of participants compared to previous research (Corley, 2010; Lowder & Ferreira, 2016). Finally, we analysed the data with Generalised Additive Mixed Models (GAMMs), which allowed us to detect non-linear patterns in the data and approximate the latency at which differences between conditions occur (Porretta et al., 2018).

2. Experiment 1: prediction in naturalistic environments

The aim of Experiment 1 was to confirm the generalisability of the findings of Heyselaar et al. (2020) by replicating their experiment with newly developed stimuli and by increasing the number of trials per condition. Participants listened to sentences while immersed in VR scenes in which six objects were embedded (e.g. letterbox, flag, tree, lamppost, basketball, and wheelbarrow). For an example scene, see Figure 1. Participants' eye gaze towards the objects was recorded. As confirmed by a pre-test, sentences were either predictable or unpredictable based on verb constraints, where the verb in the sentence was either related to a single item in the scene (Restrictive), as in sentence 1a below, making the sentence predictable, or the verb in the sentence could be related to multiple items in the scene (Unrestrictive), as in 1b below, making the sentence unpredictable. Sentences were presented in Dutch and have been translated into English here, underneath the example.

- 1a. De gemeente vergadert over het *kappen* van de boom.
  The council has a meeting about *cutting down* the tree.
- 1b. De gemeente vergadert over het *verzetten* van de boom.
  The council has a meeting about *moving* the tree.

It was hypothesised that, in a critical time-window preceding noun onset, there would be an increase in the proportion of fixations towards the target object in the Restrictive but not the Unrestrictive condition.

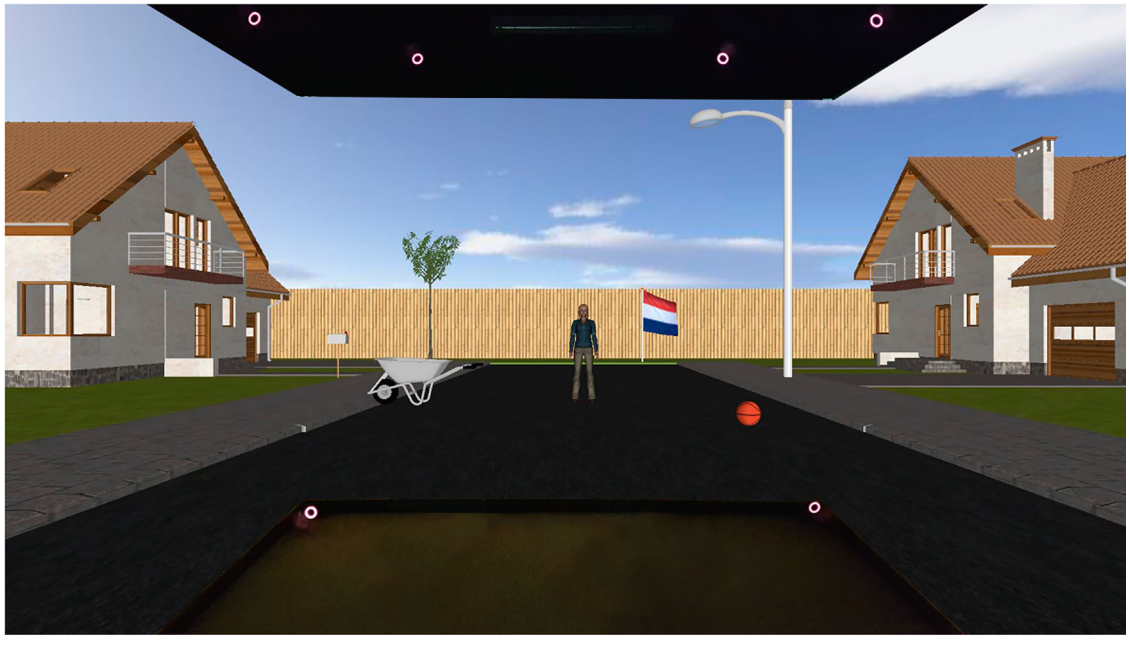


Figure 1. An example of the CAVE set up. Six of the ten infrared motion tracking cameras are visible, four at the top of the three projector screens and two in the bottom corners. The scene displays the virtual agent in the centre of the street scene, along with the six target objects (lamppost, basketball, flag, tree, letterbox, and wheelbarrow).

2.1. Materials and methods

2.1.1. Participants

Thirty-five native Dutch speakers completed the experiment, of which 32 were included in the analysis (24 female, 8 male, mean age = 24.63, age range = 19-45, SD = 5.29), in exchange for a standard fee. One participant did not complete the experiment due to cyber sickness and two participants were excluded, the first due to poor eye tracker calibration and the second due to a technical fault. Participants with photosensitive epilepsy, uncorrected visual or hearing impairments, language impairments or dyslexia were excluded from participation. All participants provided informed consent before taking part. The research was approved by Radboud University's Faculty of Social Sciences and complied with the Declaration of Helsinki.

2.1.2. CAVE system

The experiment was conducted in a cave automatic virtual environment (CAVE) system (Cruz-Neira et al., 1992). The set-up has previously been described in detail by Eichert et al. (2018) and is pictured in Figure 1. The CAVE system included three 255 × 330 cm screens (VISCON GmbH, Neurkirchen-Vluyn, Germany) arranged at right angles. Two vertically displaced, overlapping displays were indirectly back-projected onto each screen via a mirror by two projectors (F50, Barco N.V., Kortrijk, Belgium).

The experiment was programmed and run in Python through 3D VR software Vizard, Floating Client 5.4, WorldViz LLC, Santa Barbara, CA. Audio was presented through four speakers (Logitech, US) that were located in the bottom corners of the CAVE, in addition to one centred at the bottom of the middle screen.

2.1.3. Eye– and head-tracking

Participants' eye gaze was recorded throughout the experiment with specialised glasses designed to both display the presented scenes in 3D and track eye gaze within the 3D scene (SMI Eye tracking Glasses 2 Wireless, SensoMotoric Instruments GmbH, Teltow, Germany). The calibration and recording interface for the eye tracking glasses were presented on a tablet, which wirelessly transmitted data to the tracking software (described below). A camera on the glasses measured 60 Hz binocular recordings with automatic parallax compensation. Gaze tracking accuracy is reported to be 0.5 degrees for each dimension by the manufacturer. The eye tracker latency was 60 ms +/−10 ms, which was corrected for in the analysis.

Participants' head movements were tracked via six spherical passive reflective markers attached symmetrically to the outer side of each lens of the glasses. The reflective markers were detected by ten infrared cameras (Bonita 10, Vicon Motion Systems Ltd, UK) in the CAVE system (six distributed above the screens and four distributed below the screens, see Figure 1) and recorded to an accuracy of 0.5 mm with Tracker 3 software (Vicon Motion Systems Ltd, UK). Eye– and head-tracking data were combined online to continuously determine the location of participants' eye gaze in 3D space along three axes.

2.1.4. Visual stimuli

Target objects and corresponding sentences were embedded in VR scenarios. In each scenario, a virtual agent discussed her relation to the scene, with lip synchronisation and gaze towards the participant. The position of the virtual agent within each scene was chosen to make her easy to locate by the participant.

There were eight scenes that were each presented twice, with different target objects and sentences with each presentation. Six target objects were in each scene, four of which were mentioned by the virtual agent. The full list of scenes and target objects can be found in the supplementary material. Many of the target objects were taken from a standardised set of 3D objects developed in our VR laboratory (Peeters, 2018). Participants' eye gaze towards the target objects, as well as to the virtual agent, were recorded by placing (invisible to the participants) a cuboid region-of-interest (ROI) around the entirety of each object. Each fixation within a ROI was recorded online. By means of counterbalancing the sentences, target objects were counterbalanced across conditions, so that each target object was mentioned an equal number of times in each condition. Object positions remained constant in each presentation of the scene and were therefore matched across conditions. It was therefore not necessary to control for the size or position of objects within the scene. Instead, objects were placed and sized to look as natural as possible. Variability in the proportion of fixations towards objects of different sizes and locations was controlled for in the analysis, with random smooths for item (see Data analysis section).

2.1.5. Sentence stimuli

Spoken stimuli that were produced by the agent in VR were recorded by a trained native speaker of Dutch that matched the agent in apparent age and ethnicity. To retain the natural auditory characteristics of speech, each version of the sentence was recorded individually, rather than produced through cross-splicing. Fixations during the VWP have previously been demonstrated to be sensitive to subtle phonetic details like co-articulation (Dahan et al., 2001) and vowel length (Salverda et al.,

2003). Cross-splicing may have introduced misleading phonological cues (Steinberg et al., 2012) that could have led to unnatural behaviour. Furthermore, the small variations in timing within sentence sets should not have varied systematically across conditions.

Sentences consisted of 128 sentence pairs that contained either a subject-verb-object or verb-subject-object clause. Sentences within each pair were identical apart from the critical verb, one of which was Restrictive (related to a single object in the scene), whereas the other was Unrestrictive (related to multiple objects in the scene), as confirmed by a pre-test with 36 participants who did not take part in the main experiment (proportion of participants to select the target object to complete Restrictive sentences = 0.91, SD = 0.10; Unrestrictive sentences = 0.26, SD = 0.20). Sentence pairs were separated into two lists that participants were randomly assigned to, so that no participant heard both sentences from a pair. Only 50% of the 128 sentences referred to an object that was present in the scene. The remaining 50% of sentences were used as filler trials. There were therefore 32 trials per condition. Sentences were adapted from stimuli used in Heyselaar et al. (2020). Sentence onsets were recorded as the moment the audio was detected from the speaker. There was a mean duration of 596 ms (SD = 335 ms) and 613 ms (SD = 349 ms) between verb offset and noun onset for Restrictive and Unrestrictive sentences respectively. The full list of sentence stimuli and filler sentences can be found in the supplementary material.

2.1.6. Procedure

Participants stood in the VR environment 150 cm away from the central screen. They were instructed that they would be given a virtual tour of the agent's life and that their only task was to listen to her. Before beginning the experiment two calibration steps were performed. Firstly, the eye tracker was calibrated using SMI's "One-point Calibration" software. Secondly, calibration was performed with a programme developed in-house, which has previously been described by Eichert et al. (2018). Participants were asked to focus their eye gaze successively on three spheres that were embedded in the 3D scene. Each sphere differed in location along all three spatial dimensions (X/Y/Z axes). The experimenter used the keyboard to indicate which sphere the participant was looking at, which corrected the calibration. This was repeated until the deviance in all three coordinates was below a minimal threshold value (<4), consistent with Heyselaar et al. (2020).

Trials were presented in four blocks of four scenes each. Between each block the precision of the 3D eye tracking calibration was tested, and half-way through the experiment (after eight scenes) the eye tracking was re-calibrated. If the calibration had deteriorated in between each block, the eye tracker was additionally calibrated in the first and third breaks. Scenes were presented in the same order for each participant. However, the order of presentation of each scene's two sets of objects was counterbalanced across participants. For example, for 50% of participants the first scene (the street scene) contained object set A (lamppost, basketball, flag, tree, letterbox, and wheelbarrow) and for 50% of participants the first scene contained object set B (lolly, hula hoop, traffic barrier, umbrella, balls, and bucket).

After the experiment participants were asked to fill in two questionnaires. An "Object questionnaire" contained a list of all target objects that were presented throughout the experiment. Participants indicated whether or not they heard the virtual agent refer to each object and their accuracy was used to confirm their attentiveness. The second questionnaire ("Verb questionnaire") presented the list of verbs that participants had heard during the experiment. Participants were asked to indicate whether they knew what the verb meant.

2.1.7. Pre-processing

Data were recorded at a sampling frequency of 60 Hz. Trials were time-locked to 60 ms after sentence onset to account for the eye tracker latency. Eye gaze on an object was considered a fixation if it exceeded 100 ms. Samples in which coordinates across all three dimensions remained the same for more than two samples were removed to account for eye blinks or loss of eye tracker connectivity. If more than 25% of samples were removed during the critical time window of a trial, the trial was excluded. This resulted in a maximum exclusion of four trials (mean = 0.75, SD = 1.17) per participant. The critical window that was entered into the GAMMs was defined as 200 ms after verb onset until the mean noun onset, to account for the time it takes to programme an eye-movement (Rayner et al., 1983).

2.1.8. Data analysis

To conform with the analysis of Heyselaar et al. (2020) and to recent recommendations for analysing VWP eye tracking data (Porretta et al., 2018), data were analysed with generalised additive mixed-effects models (GAMM), in R (version 3.5.2; R Core Team, 2018) with packages *mgcv* (version 1.8-26; Wood, 2017) and *itsadug* (version 2.3; Van Rij et al., 2017). GAMMs are a non-linear mixed-effects regression approach, with the ability to model curvy effects by fitting smooth terms to the data. The advantages of using GAMM models are outlined in detail elsewhere (Porretta et al., 2018; van Rij et al., 2019; Wieling, 2018; Winter & Wieling,

2016). In brief, the advantage of using GAMMs for the current data are twofold. Firstly, we avoid making assumptions about the linearity of the data, as GAMMs can model both linear and non-linear effects. Secondly, GAMMs allow for non-linear interactions between continuous variables, allowing us to keep time as a continuous measure and investigate whether target fixations change dependent on time and condition.

Binomial responses of target "hit" vs "no-hit" were entered into the GAMM as the dependent variable, applying a logit link function. Analysing aggregate looks towards objects is the typical approach taken for VWP data (opposed to fixation durations, for example). The model included *Constraint* (Restrictive/Unrestrictive) as a parametric component, factor smooth interactions of *Time* × *Constraint*, *Time* × *Sentence,* and *Time* × *Subject*. Constraint was coded with a deviation contrast-coding scheme (with *contr.sum)*, which compares the mean proportion of fixations for each level of Constraint with the overall mean across levels. To avoid overfitting, the parameter k (which limits the order of base functions used to fit the model) was limited to five. Maximum Likelihood estimation was selected as the smoothing parameter estimation method to facilitate model comparison procedures (Wieling, 2018).

The model was interpreted with the *summary*, *gam.check*, and *plot_smooth* functions. Time periods of significant differences were evaluated with the *plot_diff* function, which uses confidence intervals to estimate time periods of significant differences.

2.2. Results

All participants achieved over 65% accuracy on the Object questionnaire (mean = 83%, SD = 0.09%), suggesting that all participants were attentive. If participants reported that they did not know the meaning of a verb in the Verb questionnaire, the trial containing the corresponding sentence was excluded from the analysis for that participant (mean = 2.73%, SD = 1.91%). No further analysis was conducted on the questionnaire data.

The proportion of target and distractor fixations for Restrictive and Unrestrictive conditions are presented in Figure 2. We hypothesised that there would be an increase in the proportion of target fixations in the Restrictive condition but not the Unrestrictive condition in the time window between verb and noun onset, due to verb constraints rendering the sentence predictable. The model summary is presented in Table 1. The model confirmed that there was a significant parametric effect of condition. The significant smooth for time for the Restrictive condition demonstrates that the change in proportion of target fixations was significantly greater than zero. In contrast, the smooth for time in the Unrestrictive condition did not significantly differ from zero.

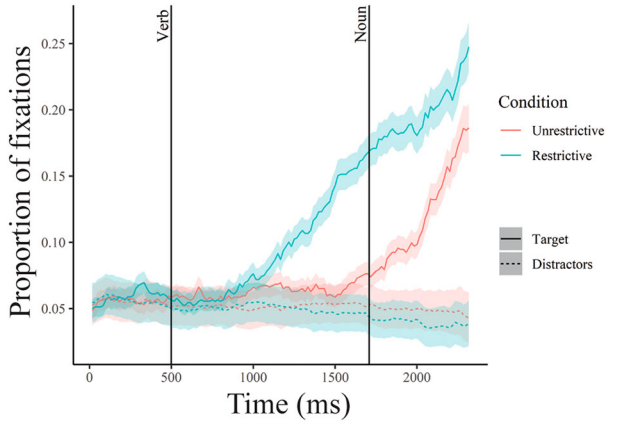


Figure 2. Proportion of target fixations. Vertical lines represent critical time points (mean verb onset and mean noun onset). Shaded ribbons display the standard error of the mean. Zero indicates 500 ms prior to verb onset.

Figure 3 displays the difference between the model-estimated smooths splines of the Restrictive and Unrestrictive conditions. Figure 3 confirms that there was a greater proportion of target fixations in the Restrictive compared to Unrestrictive condition during a time window of 368 ms after verb onset until noun onset. The model additionally estimated the proportion of target fixations to be lower in the Restrictive compared to Unrestrictive condition during a time window of 200–230 ms after verb onset.

2.3. Interim Discussion

The aim of Experiment 1 was to confirm the generalisability of the findings of Heyselaar et al. (2020) by replicating their experiment with newly developed stimuli and by increasing the number of trials per condition.

Table 1. Model summary for target fixations in Restrictive vs Unrestrictive conditions after verb onset

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−3.10	0.18	−17.21	<.001
Condition	0.26	0.01	21.00	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: Restrictive	1.01	1.01	42.16	<.001
Smooth for Time: Unrestrictive	1.00	1.01	0.85	.359
Random effect for Subjects	133.99	269.00	2143.73	<.001
Random effect for Sentence	309.983	575.00	5209.54	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Deviation contrast-coding: Restrictive (1); Unrestrictive (−1).

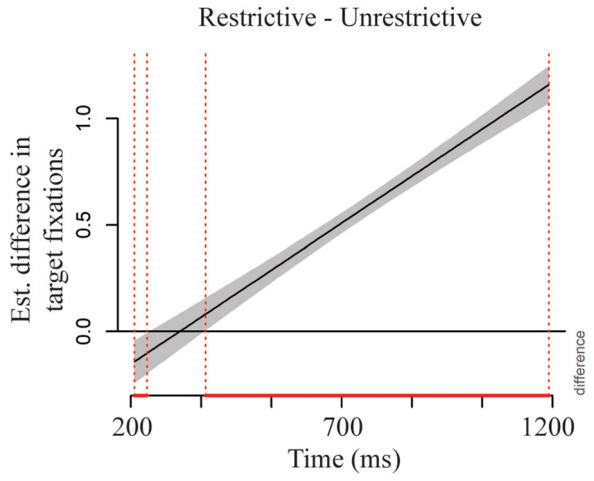


Figure 3. The difference between the model-estimated smooths splines of the Restrictive and Unrestrictive conditions in the 1000 ms preceding noun onset. Red dashed lines mark windows of significant differences. Time is relative to verb onset. Tick marks represented steps of 166.67 ms.

Supporting our hypothesis, in a time window preceding noun onset, there was a significantly greater proportion of target fixations in Restrictive sentences compared to Unrestrictive sentences from 368 ms after verb onset. In contrast to Restrictive sentences, there was no significant increase in the proportion of target fixations in Unrestrictive sentences prior to noun onset.

Although the model additionally estimated the proportion of target fixations to be lower in the Restrictive compared to Unrestrictive condition in a brief, early time window, this resulted from the model estimating the increase in the proportion of target fixations over time as linear, which can be seen both from an effective degrees of freedom (edf) close to 1 (see Table 1) and from Figure 3.

In summary, Experiment 1 replicated the findings of Heyselaar et al. (2020) with new stimuli, further demonstrating that listeners predict upcoming speech, not only in artificial experimental settings where visual input is limited, but also in naturalistic, visually rich environments, in which participants are directly addressed by a visible speaker who produces communicatively relevant speech. These findings provided us with a reliable basis to investigate the role of disfluencies in speech processing under visually rich circumstances.

3. Experiment 2. Repair vs Conjunction

The aim of Experiment 2 was to extend the findings of Experiment 1 and investigate the extent that subtle cues in speech, such as repair disfluencies, inform the listener's predictions. The inhibition of unlikely candidates for upcoming speech and enhancement of probable continuations has been suggested as one possible mechanism through which prediction could take place (Ryskin et al., 2020). As outlined in the Introduction, it remains unclear whether listeners simply inhibit their prediction upon hearing a repair disfluency, or whether they use information from the error to inform further predictions. Previous literature has demonstrated that the listener's gaze shifts towards semantic and phonological competitors of an erroneous noun, possibly suggesting that listeners interpret the error as an intrusion and adjust their prediction accordingly (Karimi et al., 2019). However, such findings have only been demonstrated with simple visual displays and lack an integral part of language in which a visible speaker is communicating a message to the listener. Karimi et al. (2019) provided some evidence in favour of a prediction, rather than a lexical priming account of looks towards the semantic competitor, by showing that looks towards the semantic competitor were greater in the repair condition compared to a coordination condition, "and also" (see also Lowder & Ferreira, 2016), as well as a silent pause condition. However, an alternative explanation is that the listener suppresses their prediction and attention is drawn to lexical competitors through automatic priming mechanisms (Collins & Loftus, 1975).

To investigate the extent that repair disfluencies inform the listener's predictions, Experiment 2 compared a *Repair* disfluency sentence (Sentence 2c below) with a *conjoined* verb sentence (*Conjunction;* Sentence 2d below), similar to Corley (2010). The Restrictive and Unrestrictive conditions from Experiment 1 (Sentences 2a and 2b below) were included as a means to check that participants were predicting in general.

- 2a. De gemeente vergadert over het *kappen* van de boom.
  The council has a meeting about *cutting down* the tree.
- 2b. De gemeente vergadert over het *verzetten* van de boom.
  The council has a meeting about *moving* the tree.
- 2c. De gemeente vergadert over het *kappen uh verzetten* van de boom.
  The council has a meeting about *cutting down uh moving* the tree.
- 2d. De gemeente vergadert over het *kappen en verzetten* van de boom.
  The council has a meeting about *cutting down and moving* the tree.

In line with the findings from Experiment 1, it was hypothesised that there would be an increase in the proportion of target fixations in the single verb Restrictive condition (Sentence 2a above) but not the single verb

Unrestrictive condition (Sentence 2b above) in a time window between verb and noun onset. Secondly, it was hypothesised that, in a critical time window between conjunction (and/uh) onset and noun onset, (a) there would be a lower proportion of target fixations in the Repair condition (Sentence 2c above) compared to the Conjunction condition (Sentence 2d above), and (b) that there would be a steeper increase in the proportion of target fixations over time in the Conjunction condition compared to Repair condition.

To distinguish between different theoretical accounts of how repair disfluencies influence the listener's prediction, we investigated looks towards both the virtual agent and distractor objects during the critical time window between conjunction onset and noun onset. If participants efficiently apply information from the repair to update their prediction, the proportion of fixations towards items compatible with the second, repaired (unrestrictive) verb (i.e. the *critical distractors*) would be expected to increase. In contrast, if participants merely inhibit their prediction and wait for the sentence to become disambiguated, the proportion of critical distractor fixations should remain constant over time. Instead, one would expect an increase in the proportion of virtual agent fixations.

In summary, we expected repair disfluencies (compared to conjunctions) to result in a reduced proportion of anticipatory target fixations, as listeners inhibit their prediction, and instead to either see an increased proportion of critical distractor fixations, as listeners attempt to make a new prediction, or alternatively, an increased proportion of virtual agent fixations, as listeners wait for the sentence to become disambiguated.

3.1. Methods

3.1.1. Participants

Thirty-six native Dutch speakers participated, of which 32 were included in the analysis (23 female, 9 male, mean age = 21.67, age range = 18-25, SD = 2.13), in exchange for a standard fee. One participant was excluded due to poor eye tracker calibration and three participants were excluded due to technical faults. The same ethical procedures and exclusion criteria were applied as outlined in Experiment 1 Methods (section 2.1).

3.1.2. Materials, procedure, and design

The same stimuli and procedure were used as in Experiment 1, with the added manipulation of Fluency (*Repair*/*Conjunction*). There were 16 trials in each condition. In both the *Repair* and *Conjunction* conditions the Restrictive verb was always followed by the Unrestrictive verb, either separated by a brief "uh" (mean duration of 398 ms, SD = 0.21) or an "and" conjunction – "en" in Dutch – (mean duration 163 ms, SD = 0.09). Sentence sets were separated into four lists that participants were randomly assigned to, so that participants only heard one sentence from each set. Only 50% of the sentences referred to an object present in the scene. The remaining 50% of sentences were filler trials. Consistent with Experiment 1, participants filled in an Object questionnaire and a Verb questionnaire after completing the experiment (see Experiment 1 Methods, section 2.1 for details).

3.1.3. Data analysis

Due to the difference in length of the single verb and conjoined sentences, two independent analyses were carried out to compare (1) the Restrictive and Unrestrictive single verb sentences, and (2) the repaired verb and conjoined verb sentences.

The analysis that compared the two single verb conditions was identical to the analysis of Experiment 1, and details can be found in Experiment 1 Data analysis (section 2.1). The same analysis procedure was again applied to compare the Repair and Conjunction conditions with the following differences. The Repair and Conjunction conditions were compared in a time window between mean conjunct onsets (and/uh) and mean noun onsets. The mean time between mean conjunct onset and mean noun onset was 1672 ms (SD = 254 ms) and 2368 ms (SD = 298 ms) for the Conjunction and Repair conditions respectively. A *Time* × *Trial Number* random smooth was added to the model to capture any changes to the pattern of fixations throughout the experiment. This was due to the possibility that participants may have learned throughout the experiment that the repair never led to a referent that was incompatible with the first (erroneous) verb, which meant they did not need to update their prediction. Including the random smooth of *Time* × *Trial Number* improved the model fit compared to a model without such a random smooth (AIC difference of 268.31, $p < .001$). The parameter k was set at the default value of 10. To test the different hypotheses outlined in the introduction to Experiment 2 (section 3), separate models were fitted for target fixations, virtual agent fixations, and critical distractor fixations. As unrestrictive verbs were designed to be compatible with at least two distractor objects in the associated scene, critical distractors were defined as the two objects most selected to fit the unrestrictive sentence in a pre-test. The proportion of participants in the pre-test who selected critical distractors and random distractors as possible continuations for each unrestrictive sentence, separated by

counterbalancing lists, is presented in Figure S2 in the supplementary material.

3.2. Results

All participants achieved over 63% accuracy on the Object questionnaire (mean = 82%, SD = 0.09%), suggesting that all participants were attentive. If participants reported that they did not know the meaning of a verb in the Verb questionnaire, the trial containing the corresponding sentence was excluded from the analysis (mean = 3.32%, SD = 2.05%). No further analysis was conducted on the questionnaire data.

3.2.1. Single critical verb

Figure 4 displays the proportion of fixations towards the target, virtual agent, critical distractors, and random distractors in the Restrictive and Unrestrictive single critical verb sentences. Consistent with Experiment 1, the model (see summary in Table 2) confirmed that there was a significant parametric effect of condition and a significant smooth for time in the Restrictive condition but not the Unrestrictive condition, supporting the hypothesis that there would be a significant change in the proportion of target fixations over time in the Restrictive but not the Unrestrictive condition. The model-estimated difference curve in Figure 5 suggests that there was a greater proportion of target fixations in the Unrestrictive compared to Restrictive condition 200–338 ms following verb onset, and in the Restrictive compared to Unrestrictive condition during a time window of 508 ms after verb onset until noun onset.

Table 2. Model summary for target fixations in Restrictive and Unrestrictive and conditions after verb onset

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−3.33	0.25	−13.14	<.001
Condition	0.15	0.02	9.12	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: Restrictive	1.01	1.01	18.13	<.001
Smooth for Time: Unrestrictive	1.00	1.01	0.38	.54
Random effect for Subjects	141.55	287.00	1644.99	<.001
Random effect for Sentence	279.46	575.00	3452.89	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Deviation contrast-coding: Restrictive (1); Unrestrictive (−1).

3.2.2. Paired critical verbs

Figure 6 displays the proportion of fixations towards the target, critical distractors, random distractors, and virtual agent in the paired critical verb sentences for the first half of trials (panel A) and the second half of trials (panel B) separately. Although trial number was entered into the model as a continuous random effect variable, here we have plotted early and late trials separately for visualisation purposes. An exploratory analysis comparing the change in the proportion of target fixations in early and late trials can be found in Experiment 2 Supplementary Material.

3.2.2.1. Target fixations. It was hypothesised that there would be a greater increase in the proportion of target fixations in the Conjunction compared to Repair condition in a time window between conjunction (and/uh)

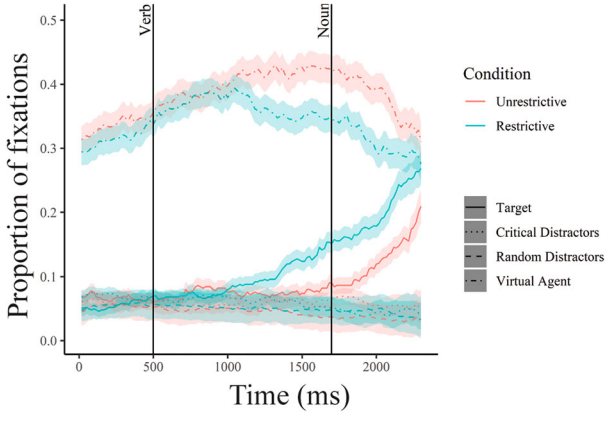


Figure 4. Proportion of fixations towards the target (solid line), critical distractors (dotted line), random distractors (dashed line), and virtual agent (dot-dashed line) in single critical verb sentences. Vertical lines represent critical time points (mean verb onset and mean noun onset). Shaded ribbons represent standard error of the mean. Zero indicates 500 ms prior to verb onset.

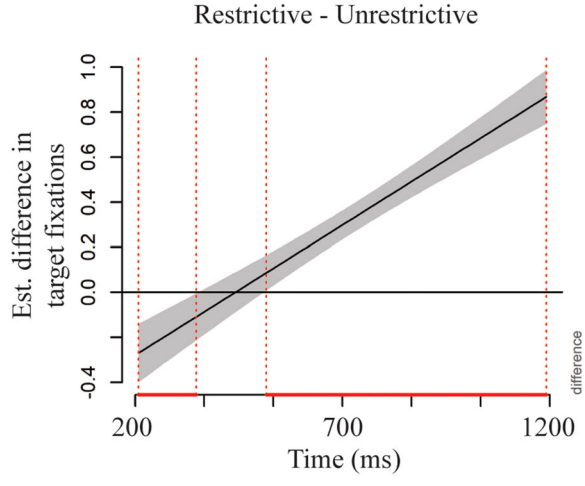


Figure 5. The difference between the model-estimated smooths splines of the Restrictive and Unrestrictive Fluent conditions in the 1000 ms preceding noun onset. Red dashed lines mark windows of significant differences. Time is relative to verb onset. Tick marks represent steps of 166.67 ms.

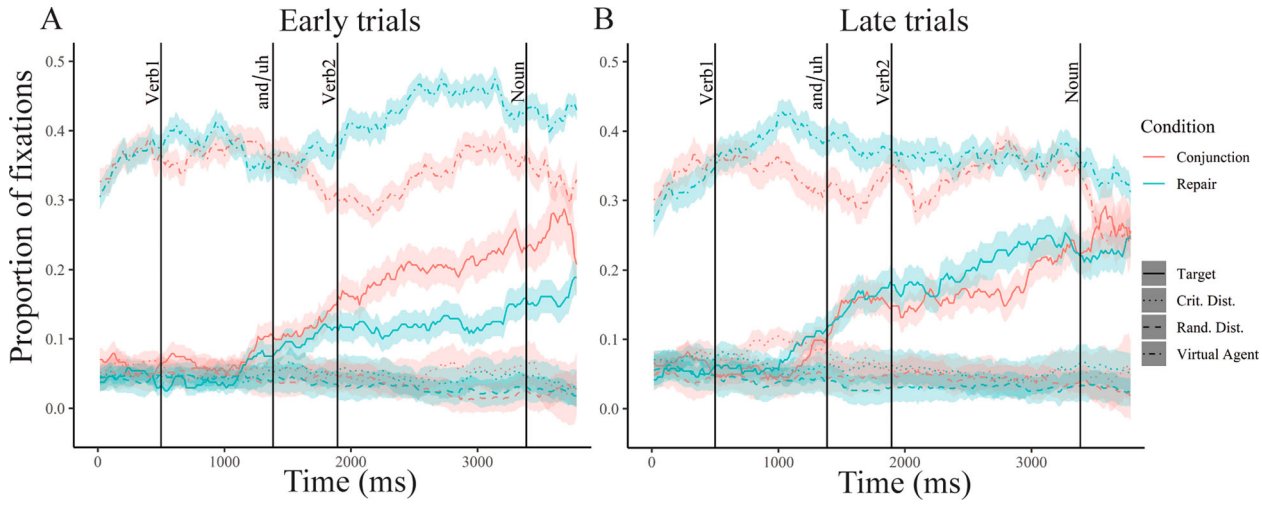


Figure 6. Proportion of fixations towards the target (solid line), critical distractors (Crit. Dist.; dotted line), random distractors (Rand. Dist.; dashed line), and virtual agent (dot-dashed line) in conjoined verb sentences in (A) the first half of trials; and (B) the second half of trials. Vertical lines represent critical time points (mean onsets). Shaded ribbons represent standard error of the mean. Zero indicates 500 ms prior to verb1 onset.

onset and noun onset. The model summary of the target fixation GAMM is presented in Table 3. There was a significant parametric effect of condition and a significant smooth for time for the Conjunction condition, but not the Repair condition, which demonstrates that there was no significant change in the proportion of target fixations over time in the Repair condition between repair and noun onsets. Figure 7 displays the difference between the model-estimated smooth splines of the Repair and Conjunction conditions. The proportion of target fixations was estimated to be significantly lower in the Repair compared to Conjunction condition between 362–1289 ms, and from1695 ms after conjunction onset until noun onset. Visual inspection of Figure 6 suggests that there was a difference between conditions in early (panel A) but not late (panel B) trials.

Table 3. Model summary of GAMM comparing target fixations in Conjunction and Repair conditions after the conjunction (and/uh) onset

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−2.49	0.27	−9.30	<.001
Condition	0.09	0.01	9.26	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: and	4.80	5.84	28.75	<.001
Smooth for Time: uh	0.03	0.04	0.01	.943
Random effect for Subjects	203.01	287.00	5180.22	<.001
Random effect for Sentence	395.82	575.00	9003.04	<.001
Random effect for Trial number	20.99	23.64	273.20	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Deviation contrast-coding: Conjunction (1); Repair (−1).

3.2.2.2. Critical distractor fixations. The results of the model for target fixations (summarised in Table 3) support the notion that hearing a repair disfluency reduces listeners' predictions of a likely referent, compared to when hearing a conjoined verb. It was hypothesised that, if participants discard their prediction and make a new prediction based on the repair disfluency, there would be an increase in the proportion of fixations towards distractor items compatible with the unrestrictive verb (i.e. critical distractor fixations) in the Repair condition, in a time window between conjunction (and/uh) onset and noun onset. The model summary of

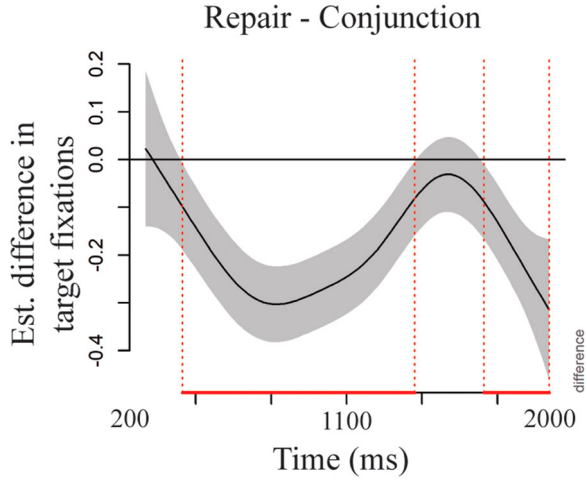


Figure 7. The difference between the model-estimated smooth splines of target fixations in Repair vs Conjunction sentences. Red dashed lines mark windows of significant differences. Time is relative to conjunction (and/uh) onset. Tick marks represent steps of 300 ms.

Table 4. Model summary of GAMM comparing critical distractor fixations in Conjunction and Repair conditions after the conjunction onset

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−3.10	0.25	−12.27	<.001
Condition	0.05	0.01	3.91	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: and	5.77	6.84	93.71	<.001
Smooth for Time: uh	1.01	1.01	0.05	.839
Random effect for Subjects	216.69	287.00	3232.39	<.001
Random effect for Sentence	410.37	575.00	5766.36	<.001
Random effect for Trial number	27.62	28.33	504.96	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Deviation contrast-coding: Conjunction (1); Repair (−1).

the critical distractor fixation model is presented in Table 4. There was a significant parametric effect of condition and a significant smooth for time for the Conjunction condition but not the Repair condition. These findings do not support the hypothesis that there would be an increase in the proportion of critical distractor fixations upon hearing a repair disfluency.

The difference curve in Figure 8 demonstrates that the model-estimated proportion of critical distractor fixations was significantly lower in the Repair compared to the Conjunction condition 1083–1695 ms relative to conjunction onset, but significantly higher from 1822ms after conjunction onset until noun onset.

3.2.2.3. Virtual agent fixations. Contrary to fixating on a new object according to the next most likely continuation, it was hypothesised that the repair disfluency could result in enhanced attention to the (speech of the) virtual agent. In such a scenario it was expected that there would be an increase in the proportion of virtual agent fixations upon hearing a repair disfluency, but not upon hearing a conjoined verb, in the time window between conjunction (and/uh) and noun onset. The model summary of the virtual agent fixation model is presented in Table 5. There was a significant parametric effect of condition and a significant smooth for time for the Repair condition but not the Conjunction condition. These findings support the hypothesis that there would be an increase in the proportion of virtual agent fixations upon hearing a repair disfluency.

The difference curve in Figure 9 demonstrates that the model estimated there to be a significantly higher proportion of virtual agent fixations in the Repair compared to the Conjunction condition during the entire critical window, from 200 ms after conjunction onset until noun onset.

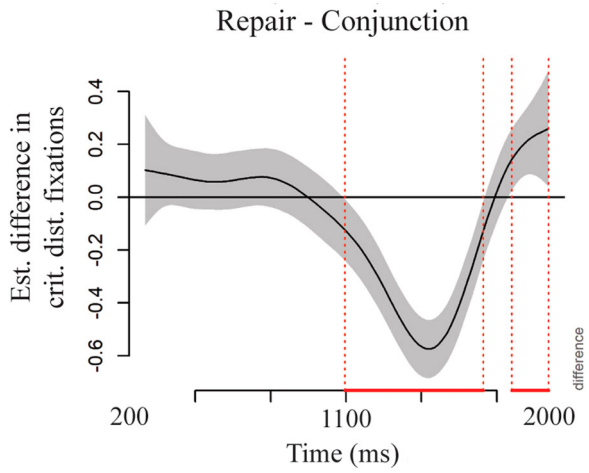


Figure 8. The difference between the model-estimated smooth splines of critical distractor fixations in Repair vs Conjunction sentences. Red dashed lines mark windows of significant differences. Time is relative to conjunction (and/uh) onset. Tick marks represent steps of 300 ms. *Note*: crit. dist. critical distractor.

3.3. Interim Discussion

The aim of Experiment 2 was to investigate the extent that repair disfluencies inform the listener's predictions. In support of our hypotheses, the proportion of anticipatory target fixations was reduced following a repair disfluency (e.g. "… cutting down *uh* moving the tree") compared to a conjunction (e.g. "… cutting down *and* moving the tree"). In addition to supporting Corley (2010), these findings corroborate that listeners in rich, naturalistic environments rapidly update their predictions in response to repaired speech. However, results diverge from Corley (2010) and from expectations, in that looks towards the target object did not *decrease* upon hearing a repair disfluency. It is therefore unclear whether listeners suppressed or abandoned their initial prediction after hearing the repair disfluency here, or whether listeners merely place less weight on their

Table 5. Model summary of GAMM comparing virtual agent fixations in Conjunction and Repair conditions after the conjunction onset

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−0.66	0.19	−3.52	<.001
Condition	−0.16	0.01	−21.60	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: and	0.18	0.24	0.05	.830
Smooth for Time: uh	3.38	4.18	17.70	.002
Random effect for Subjects	199.10	287.00	8660.24	<.001
Random effect for Sentence	370.09	575.00	8205.64	<.001
Random effect for Trial number	28.35	28.72	1000.82	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Deviation contrast-coding: Conjunction (1); Repair (−1).

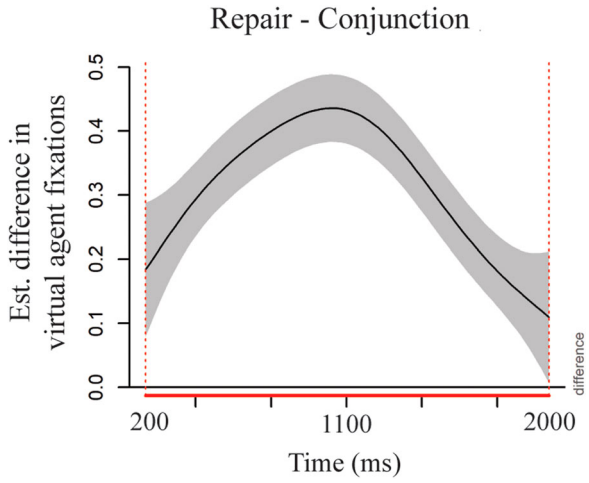


Figure 9. The difference between the model-estimated smooth splines of virtual agent fixations in Repair vs Conjunction sentences. Red dashed lines mark windows of significant differences. Time is relative to conjunction (and/uh) onset. Tick marks represent steps of 300 ms.

original prediction, without suppressing it completely. In contrast to the current results, previous VWP studies have shown that looks towards the initially predicted item decrease if the constraining information is marked as erroneous with a repair disfluency (Corley, 2010; Karimi et al., 2019; Lowder & Ferreira, 2016). In the current experiment, the repair (e.g. " … cutting down *uh* moving … ") never ended in a verb that signalled to the listener that they should update their prediction (i.e. *tree*), which could explain the lack of a *decrease* in target fixations. However, this does not explain why our findings differ from Corley (2010), where similar stimuli were used (e.g. … eat uh move …). It could be that in naturalistic environments, where the visual context is much richer, more time is required for listeners to update their predictive eye movements. Our findings could be in line with evidence to suggest that words spoken in error are not completely suppressed, but, instead, information from the speech error lingers after hearing the repair (Ferreira et al., 2004; Ferreira & Bailey, 2004; Lau & Ferreira, 2005; Slevc & Ferreira, 2013).

To distinguish between different theoretical accounts of how repair disfluencies influence the listener's prediction, we additionally investigated looks towards the virtual agent and critical distractor objects between conjunction onset and noun onset. We hypothesised that, if participants efficiently apply information from the repair to update their prediction, the proportion of fixations towards items compatible with the repaired verb (i.e. the *critical distractors*) would increase. In contrast, if listeners merely inhibit or place less confidence in their initial prediction and wait for the sentence to become disambiguated, the proportion of critical distractor fixations should remain constant over time. Our findings supported the latter, that participants wait for the sentence to become disambiguated after hearing a repair disfluency. There was no significant change in the proportion of critical distractor fixations over time in the Repair condition. This contrasted with the Conjunction condition, in which there was a brief increase in the proportion of critical distractor fixations, resulting in a significantly greater proportion of fixations to critical distractors in the Conjunction condition compared to the Repair condition in a period of 1083–1695 ms relative to conjunction onset. There was, however, a significant change in the proportion of virtual agent fixations over time in the Repair condition, but not the Conjunction condition. To summarise, upon hearing a repair disfluency, participants' attention no longer moved towards the originally predicted item, possibly as they inhibited (Ryskin et al., 2020) or placed less confidence in their initial prediction, and instead their attention moved towards the virtual agent. Although it appears as though participants reconsidered their original prediction, we saw no indication that a new prediction was made based on the repaired verb. Instead, listeners seem to realign their attention with the speaker.

So that spoken stimuli sounded as natural as possible, each sentence was recorded separately by a native Dutch speaker. This meant that the duration of "uh" in the Repair condition was longer than the duration of "en" in the Conjunction condition. It could therefore be argued that, potentially, participants' attention towards the target object was lost during this more extended period of time. However, previous research supports that a temporal delay alone cannot account for the effect of disfluencies on sentence processing (Fraundorf & Watson, 2011). In Figure 6 panel A it can be seen that target fixations continue to increase from shortly after the first verb onset until after noun onset in the Conjunction condition, but plateau shortly before the second verb onset in the Repair condition, where it does not increase again until after noun onset. If the disfluency were to result in a brief lapse of attention during the temporal delay, rather than a change in the listener's prediction, one might expect fixations towards the predicted item to continue to increase once the sentence continues to unfold. Instead, listeners seem to wait for the sentence to be disambiguated before their attention is reengaged with the target object. Furthermore, the increased looks towards the virtual agent upon hearing a repair speak against an interpretation of the data as a result of a temporal delay. If listeners' attention was lost due to the temporal delay, it might be expected that different participants would look away towards different areas of the scene, rather

than the speaker systematically capturing the listeners' attention. For these reasons we do not believe that a temporal delay alone can explain our findings. However, such arguments are speculative, and require confirmation with further empirical work.

Our findings differ from Karimi et al. (2019), in that we found no evidence to suggest that participants made a new prediction upon hearing the repair disfluency. However, it is important to note that the current experimental design differed in important ways. Firstly, whereas the current design, in line with (Corley, 2010), constrained possible upcoming nouns through the properties of the (repaired) verb, Karimi et al. (2019) constrained the repaired noun through the properties of the erroneous noun. Hence, the questions being asked across these separate studies were slightly different. Here, we wanted to know whether listeners use information from the *repaired* verb to update their predictions. In contrast, Karimi et al. (2019) set out to establish the extent to which the listener uses information from the *error* to predict the *repair*.

The question still remains, however, as to whether more subtle cues in speech, such as hesitations, more generally inform the listener's predictions, as has been suggested by traditional laboratory experiments (Arnold et al., 2007). As outlined in the Introduction, hesitations are more likely to occur when upcoming speech is difficult to conceptualise or produce (Bortfeld et al., 2001; Brennan & Williams, 1995; Fraundorf & Watson, 2014; Schachter et al., 1991; Smith & Clark, 1993). A hesitation could therefore signal to the listener that the upcoming speech is unlikely to be the most predictable outcome. Although the extent to which viewing an object automatically activates its lexical representations is a current topic of debate (Huettig et al., 2011; Magnuson, 2019), lexical retrieval should be faster and less demanding for a referent that is visible in the scene compared to a referent that is absent from the scene, as the semantic features of visible objects are already activated (Huettig et al., 2011). Moreover, filled hesitations in speech have been shown to reduce the listener's confidence in the speaker's utterance (Brennan & Williams, 1995; Lowder & Ferreira, 2019) and the listener updates their prediction accordingly (Lowder & Ferreira, 2019). It could therefore be the case that, upon hearing a hesitation, the listener re-evaluates their initial prediction, instead anticipating a less predictable outcome, for example, a referent that is not visible in the scene.

4. Experiment 3. Hesitations

The aim of Experiment 3 was therefore to investigate the extent to which hesitations in speech influence the listener's predictive target fixations in a naturalistic environment. We recorded fixations while participants listened to Restrictive– and Unrestrictive-Fluent sentences (sentences 3a-b below) in addition to Restrictive– and Unrestrictive-Disfluent sentences (sentences 3c-d below).

- 3a. De gemeente vergadert over het *kappen* van de boom.
  The council has a meeting about *cutting down* the tree.
- 3b. De gemeente vergadert over het *verzetten* van de boom.
  The council has a meeting about *moving* the tree.
- 3c. De gemeente vergadert over het *kappen* van *uhh* de boom.
  The council has a meeting about *cutting down uhh* the tree.
- 3d. De gemeente vergadert over het *verzetten* van *uhh* de boom.
  The council has a meeting about *moving uhh* the tree.

In line with the findings from Experiments 1 and 2, it was hypothesised that there would be an increase in the proportion of target fixations in the Restrictive conditions (sentence 3a and 3c above) but not the Unrestrictive conditions (sentence 3b and 3d above) in a time window between verb and noun/hesitation onset (Fluent/Disfluent respectively). Secondly, it was hypothesised that, in a time window directly preceding noun onset (matched in spoken content across all four conditions), there would be an increase in the proportion of target fixations over time in only the Restrictive-Fluent condition (sentence 3a above).

Both *predictive* and *attentional* accounts of a listener's response to a hesitation predict an increase in fixations towards the speaker rather than the continued scanning of objects in the environment upon hearing a hesitation. According to a predictive account, a hesitation signals to the listener that the upcoming content of speech is difficult for the speaker to retrieve or produce, and is, therefore, less likely to be an item present in the room. Similarly, according to an attentional account, the hesitation enhances the listener's attention, either through an automatic capture of attention, or due to informed priors to pay attention to the upcoming (potentially challenging) speech (Bosker, 2014). However, so far current theories of how hesitations influence predictions have been tested with paradigms that lack an integral aspect of language; the presence of a visible speaker who produces the communicative message. To confirm that this fundamental property of current theories

holds in naturalistic environments, when the speaker is visible, we investigated looks towards the virtual agent during the critical time window before noun onset. It was hypothesised that, in the time window before noun onset, which was exactly matched in spoken words across conditions (a) there would be a higher proportion of virtual agent fixations in both of the Hesitation conditions (Restrictive and Unrestrictive) compared to the Restrictive-Fluent condition, and (b) that there would be a steeper increase in the proportion of virtual agent fixations over time in the Restrictive-Hesitation condition compared to all other conditions.

In summary, we expected hesitation disfluencies (compared to fluent sentences) to result in a reduced proportion of anticipatory target fixations in the Restrictive condition, as listeners lose confidence their prediction, and instead to see an increased proportion of virtual agent fixations, as listeners wait for the sentence to become disambiguated.

4.1. Methods

4.1.1. Participants

Thirty-six native Dutch speakers participated, of which 34 were included in the analysis (28 female, 6 male, mean age = 24.35, age range = 19–43), in exchange for a standard fee. One participant did not complete the experiment due to cyber sickness and one participant was excluded due to poor eye tracker calibration. The same ethical procedures and exclusion criteria were applied as outlined in Experiment 1 Methods (section 2.1).

4.1.2. Materials, procedure, and design

The same stimuli and procedure were used as in Experiment 1, with the added manipulation of Fluency (Fluent/Disfluent) yielding two additional conditions. The hesitation "uhh" was always placed directly before the article/pronoun that preceded the object noun. There were 16 trials in each condition. Sentence sets were separated into four lists that participants were randomly assigned to, so that participants only heard one sentence from a set. Only 50% of the sentences referred to an object present in the scene. The remaining 50% of sentences were filler trials. Consistent with Experiment 1, participants filled in an Object questionnaire and a Verb questionnaire after completing the experiment (see Experiment 1 Methods, section 2.1 for details).

4.1.3. Data analysis

Differing from Experiments 1 and 2, two time windows were analysed. To enable the inclusion of data from all four conditions in a single model, the same length time windows were used for Fluent and Disfluent conditions. The first time window was from 200 ms after verb onset until mean noun onset (calculated from Fluent sentences). Thus, the critical window ended an average of 447 ms prior to the mean hesitation onset in Disfluent sentences, or 160 ms prior to the onset of the silent pause that preceded the hesitation. There was a mean duration of 605 ms (SD = 218 ms) and 627 ms (SD = 239 ms) between verb offset and the onset of the silent pause that preceded the hesitation in Restrictive and Unrestrictive sentences respectively. The analysis of the first time window was the same as the analysis used for Experiment 1 and for single verb sentences in Experiment 2, with the exception of two additional (Disfluent) conditions.

The second time window was the 683 ms preceding noun onset, which corresponded to the mean hesitation offset (with 200 ms adjustment for the time to programme a fixation) and captured the word preceding the noun. The mean time between hesitation offset and noun onset for Restrictive and Unrestrictive sentences was 888 ms (SD = 164 ms) and 901 ms (SD = 189 ms) respectively (note that the 5–8 ms difference between these values and the length of the analysis time window was caused by the analysis window having been calculated from the number of samples between the hesitation offset and noun onset, and then converted to ms).

The analysis of the second time window was the same as the analysis for the Paired Verb sentences in Experiment 2. The model included *Condition* (Restrictive-Fluent/Restrictive-Disfluent/Unrestrictive-Fluent/Unrestrictive-Disfluent) as a parametric component and factor smooth interactions of *Time* × *Constraint*, *Time* × *Sentence*, *Time* × *Subject*, and *Time* × *Trial Number*. Including the random smooth of *Time* × *Trial Number* improved the model fit compared to a model without such a random smooth (AIC difference of 98.94, $p < .001$).

4.2. Results

All participants achieved over 66% accuracy on the Object questionnaire (mean = 84%, SD = 0.07%), suggesting that all participants were attentive. If participants reported that they did not know the meaning of a verb in the Verb questionnaire, the trial containing the corresponding sentence was excluded from the analysis (mean = 3.93%, SD = 2.11%). No further analysis was conducted on the questionnaire data. Figure 10 displays the proportion of fixations towards the target, distractors and virtual agent in Restrictive and Unrestrictive, Fluent (panel A) and Disfluent sentences (panel B).

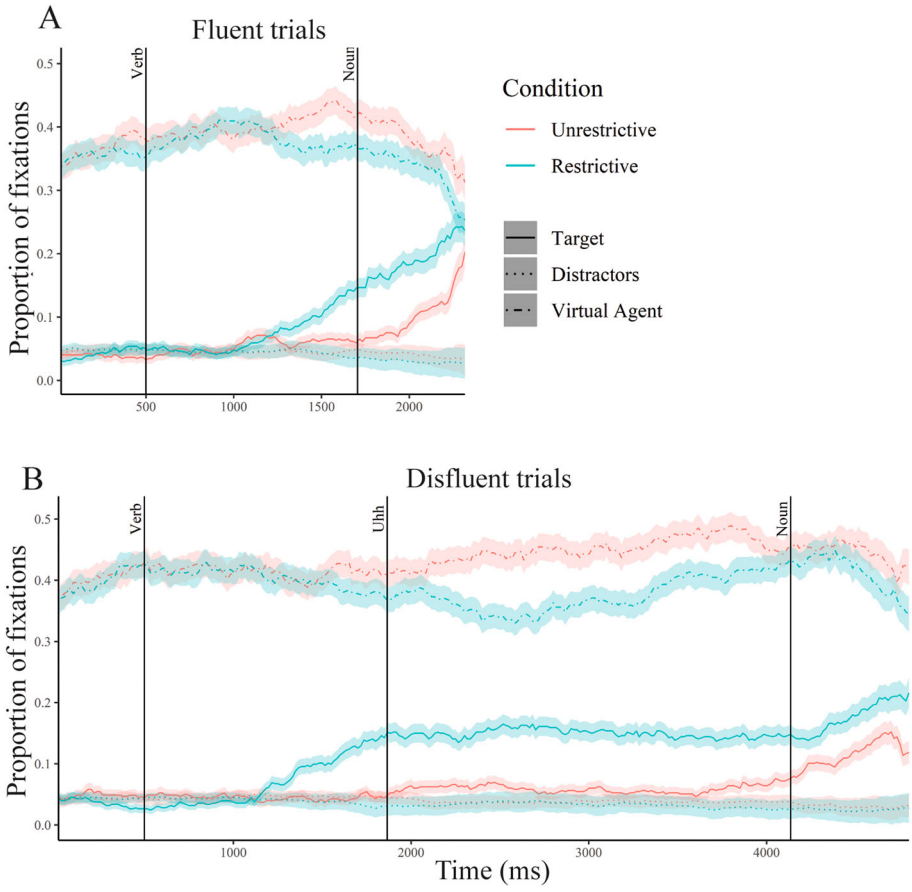


Figure 10. Proportion of fixations towards the target (solid line), distractors (dotted line), and virtual agent (dot-dashed line) for (A) Fluent sentences; (B) Disfluent sentences. Vertical lines represent critical time points (mean onsets). Shaded ribbons represent standard error of the mean. Zero indicates 500 ms prior to verb onset.

4.2.1. Post verb time window

Consistent with Experiment 1, Experiment 2, and our hypothesis, the model confirmed that there was a significant parametric effect of condition and a significant smooth for time in both the Restrictive-Fluent and –Disfluent conditions but not in the Unrestrictive-Disfluent condition in a time window preceding noun/hesitation onset (summary presented in Table 6). In contrast to Experiments 1 and 2, a significant smooth for time was also seen in the Unrestrictive-Fluent condition. On inspection of the model estimated smooths in Figure 11 panel A, this resulted from a brief rise and fall in the proportion of target fixations in this condition (also visible in Figure 10 panel A).

Consistent with Experiment 1 and 2, the model-estimated difference curves, presented in Figure 11, confirm that there was a greater proportion of target fixations in the Restrictive compared to Unrestrictive conditions from 707 ms after verb onset until noun onset in the Fluent conditions (Figure 11 panel B), and from 538 ms after verb onset until the end of the critical window (1200 ms after verb onset) in the Disfluent conditions (Figure 11 panel C). The model additionally estimated that there was a significantly lower proportion of target fixations in the Restrictive – compared to Unrestrictive-Disfluent condition in a time window

Table 6. Model summary for target fixations in Unrestrictive– and Restrictive-Fluent, and Unrestrictive– and Restrictive-Disfluent conditions post verb onset.

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−3.74	0.23	−16.34	<.001
Restrictive Fluent	0.28	0.02	12.79	<.001
Restrictive Disfluent	0.08	0.02	3.55	<.001
Unrestrictive Fluent	−0.03	0.02	−1.26	.21
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: Restrictive Fluent	1.74	2.12	30.67	<.001
Smooth for Time: Restrictive Disfluent	1.00	1.01	31.35	<.001
Smooth for Time: Unrestrictive Fluent	3.08	3.55	21.61	<.001
Smooth for Time: Unrestrictive Disfluent	1.00	1.00	2.23	.136
Random effect for Subjects	161.26	305.00	2264.25	<.001
Random effect for Sentence	303.43	575.00	6231.53	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Parametric effects present the comparison of each level of condition with the overall mean across levels.

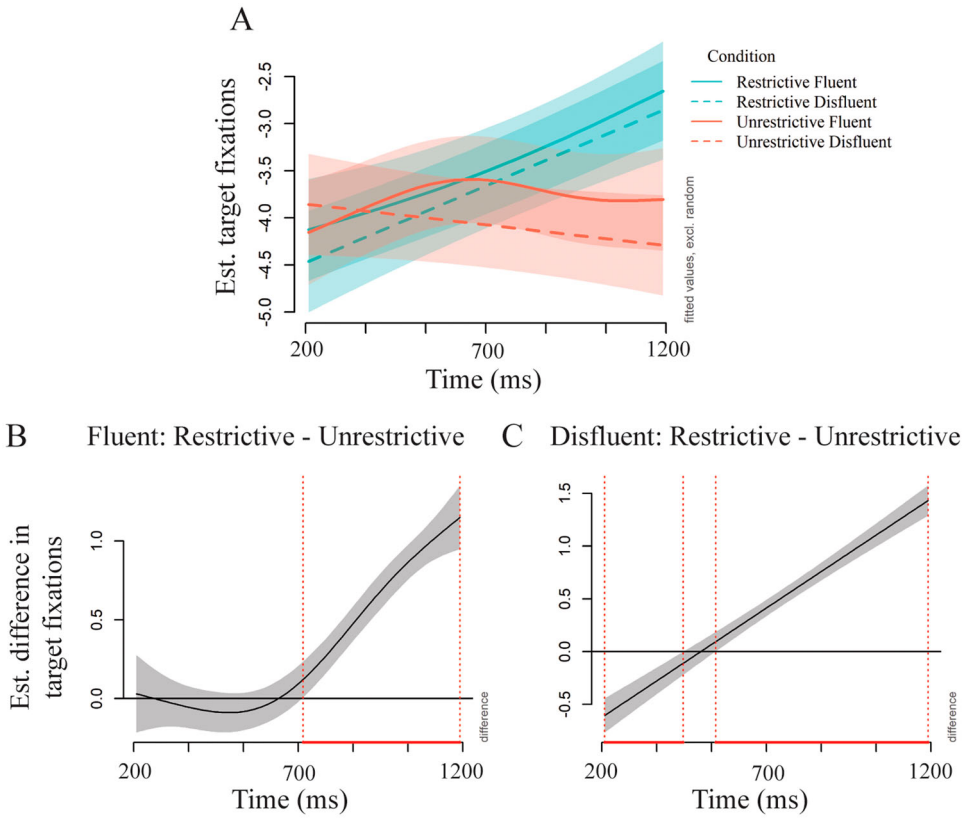


Figure 11. The target fixation model estimated smooths for each condition (Panel A), and the difference between the target model-estimated smooth splines of target fixations in the Restrictive vs Unrestrictive Fluent conditions (Panel B) and Restrictive vs Unrestrictive Disfluent conditions (Panel C). Red dashed lines mark windows of significant differences. Time is relative to verb onset. Tick marks represent steps of 166.67 ms.

from the beginning of the critical window (200 ms after verb onset) until 438 ms after verb onset (see Figure 11 panel C). Overall, these findings further support that there is a change in the proportion of target fixations over time in the Restrictive but not Unrestrictive conditions.

4.2.2. Post hesitation time window

In the time window immediately preceding noun onset, which succeeded the hesitation onset in Disfluent sentences, and was exactly matched in words spoken across conditions, it was hypothesised that there would be (a) an increase in the proportion of target fixations over time in only the Restrictive-Fluent condition, (b) a higher proportion of virtual agent fixations in both Disfluent conditions compared to the Restrictive-Fluent condition, and (c) a steeper increase in the proportion of virtual agent fixations over time in the Restrictive-Disfluent condition compared to all other conditions.

4.2.2.1. Target fixations. The model investigating target fixations (summary presented in Table 7) revealed that, in the time window preceding noun onset, there was a significant parametric effect of condition and significant smooths for time for the Restrictive-Fluent, but not for the Unrestrictive-Fluent or the Disfluent conditions.

The model estimated difference curves presented in Figure 12 panel B demonstrate that there was a significantly greater proportion of fixations towards the target in the Restrictive-Disfluent compared to Restrictive-Fluent condition from the beginning of the critical window (683 ms preceding noun onset) until −43 ms relative to noun onset. There was a significantly lower proportion of target fixations in the Unrestrictive-Disfluent compared to Unrestrictive-Fluent condition −568 ms– −171 ms relative to noun onset (see Figure 12 panel C).

The model-estimated smooths presented in Figure 12 panel A demonstrate that there was no increase in the proportion of target fixations over time during the critical time window in the Unrestrictive-Fluent condition (red solid line Figure 12 panel A). Similarly, there was no change in the proportion of target fixations over time in either of the Disfluent conditions (dashed lines Figure 12 panel A). In contrast, there was an increase

Table 7. Model summary of GAMM comparing target fixations in Unrestrictive– and Restrictive-Fluent, and Unrestrictive– and Restrictive-Disfluent conditions post hesitation onset.

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−3.22	0.23	−13.85	<.001
Restrictive Fluent	0.04	0.02	1.63	.103
Restrictive Disfluent	0.70	0.02	33.51	<.001
Unrestrictive Fluent	−0.30	0.03	−11.86	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: Restrictive Fluent	1.01	1.01	11.04	<.001
Smooth for Time: Restrictive Disfluent	1.01	1.01	0.68	.415
Smooth for Time: Unrestrictive Fluent	0.81	1.25	0.44	.440
Smooth for Time: Unrestrictive Disfluent	1.01	1.02	0.09	.767
Random effect for Subjects	96.40	305.00	2647.89	<.001
Random effect for Sentence	220.99	575.00	3348.43	<.001
Random effect for Trial number	23.16	26.06	109.00	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Parametric effects present the comparison of each level of condition with the overall mean across levels.

in the proportion of target fixations over time in the Restrictive-Fluent condition (cyan solid line Figure 12 panel A) supporting our hypothesis.

4.2.2.2. Virtual agent fixations. In the time window immediately preceding noun onset there was a significant parametric effect of condition on the proportion of virtual agent fixations, and significant smooths for time for the Restrictive-Fluent and –Disfluent conditions, but not for the Unrestrictive conditions (see Table 8).

The model estimated difference curve presented in Figure 13 panel B demonstrates that there was a significantly greater proportion of virtual agent fixations in the Restrictive-Fluent compared to Restrictive-Disfluent condition from the beginning of the critical window (683 ms preceding noun onset) until −555 ms relative to noun onset, but a significantly greater proportion of fixations towards the virtual agent in the Restrictive-Disfluent compared to Restrictive-Fluent condition from −420 ms relative to noun onset until noun onset. Visualisation of the smooths in Figure 13 panel A demonstrates that there was a steeper increase in the proportion of virtual agent fixations in the Restrictive-Disfluent (cyan dashed line) condition compared to the Restrictive-Fluent condition (cyan solid line).

The model estimated difference curves presented in Figure 13 panel C show that there was a significantly

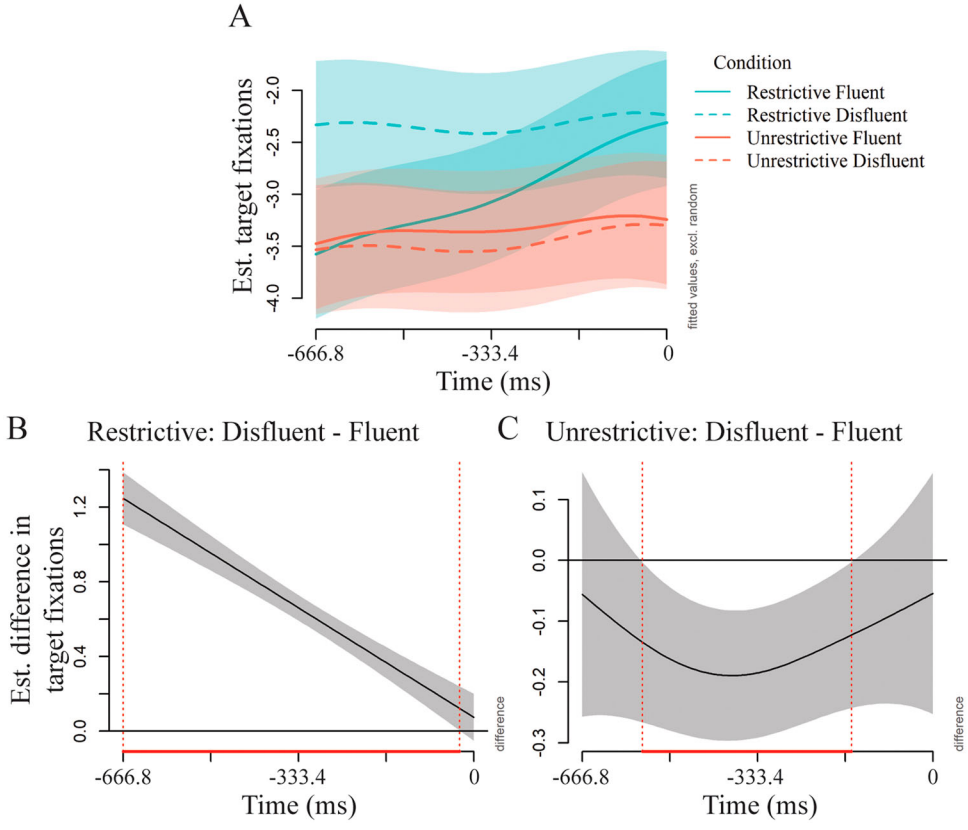


Figure 12. The target fixation model estimated smooths for each condition (Panel A), and the difference between the model-estimated smooth splines of the Disfluent and Fluent conditions for the Restrictive conditions (Panel B) and Unrestrictive conditions (Panel C). Red dashed vertical lines mark windows of significant differences. Time is relative to mean noun onset. Tick marks represent steps of −166.70 ms.

Table 8. Model summary of GAMM comparing virtual agent fixations in Unrestrictive– and Restrictive-Fluent, and Unrestrictive– and Restrictive-Disfluent conditions after hesitation onset.

Parametric coefficients	Estimate	Standard error	Z value	*P* value
Intercept	−0.34	0.14	−2.41	.016
Restrictive Fluent	−0.18	0.01	−13.87	<.001
Restrictive Disfluent	−0.07	0.01	−5.45	<.001
Unrestrictive Fluent	−0.07	0.01	−5.03	<.001
Smooth terms	**edf**	**Ref. df**	**Chi sq.**	***P* value**
Smooth for Time: Restrictive Fluent	1.01	1.02	8.20	.004
Smooth for Time: Restrictive Disfluent	1.01	1.01	21.51	<.001
Smooth for Time: Unrestrictive Fluent	0.02	1.03	<0.01	.974
Smooth for Time: Unrestrictive Disfluent	1.34	1.60	5.39	.115
Random effect for Subjects	91.58	305.00	8571.88	<.001
Random effect for Sentence	160.79	575.00	3042.05	<.001
Random effect for Trial number	27.63	28.62	681.28	<.001

edf, effective degrees of freedom; Ref. df, reference degrees of freedom. Parametric effects present the comparison of each level of condition with the overall mean across levels.

higher proportion of virtual agent fixations in the Unrestrictive-Disfluent compared to Unrestrictive-Fluent condition from the beginning of the critical window (683 ms preceding noun onset) until noun onset. Figure 13 panel A illustrates that, although there was a greater proportion of virtual agent fixations in the Unrestrictive-Disfluent condition compared to the Unrestrictive-Fluent condition, the change in fixations over time is similar across (Unrestrictive) conditions.

4.3. Interim Discussion

The aim of Experiment 3 was to investigate to what extent hesitations in speech inform the listener's predictions. Experiment 3 confirmed our hypothesis that there would be an increase in the proportion of target fixations in a time window directly preceding noun onset, which succeeded the hesitation in Disfluent sentences, in only the Restrictive-Fluent condition. In contrast, there was no longer a change in the proportion of target fixations over time in the Restrictive-Disfluent condition during this time window. This pattern of fixations is consistent with those observed in response to a repair disfluency in Experiment 2, and may reflect the listener losing confidence in their initial prediction.

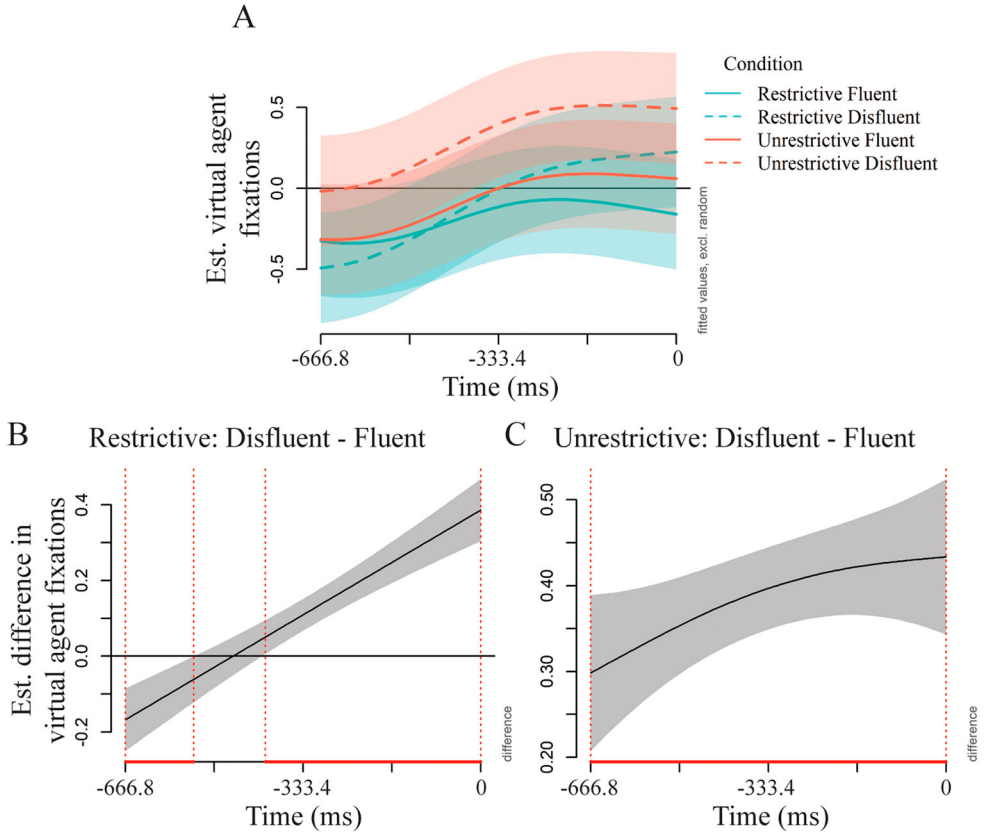


Figure 13. The virtual agent fixation model estimated smooths for each condition (Panel A), and the difference between the model-estimated smooth splines of the Disfluent and Fluent conditions for the Restrictive conditions (Panel B) and Unrestrictive conditions (Panel C). Red dashed lines mark windows of significant differences. Time is relative to mean noun onset. Tick marks represent steps of −166.70 ms.

Current theoretical accounts of a listener's response to a hesitation predict an increase in fixations towards the speaker rather than the continued scanning of the environment, as outlined in the Introduction and the introduction to Experiment 3 (section 3). Previously, current theories of how hesitations influence predictions have only been tested with traditional paradigms in which sentences are presented through a disembodied voice. Virtual reality provided a platform in which we could incorporate a virtual agent into the paradigm and investigate whether listeners' eye gaze behaviour is consistent with current theories in naturalistic environments.

Our findings supported the hypothesis that there would be an increase in the proportion of virtual agent fixations over time in the Restrictive-Disfluent condition compared to all other conditions, thereby supporting that theoretical accounts of how hesitations influence sentence processing do indeed hold in more naturalistic environments. In contrast, the change in fixations over time was very similar in the Unrestrictive-Disfluent condition compared to the Unrestrictive-Fluent condition. Importantly, investigating looks towards the virtual agent has demonstrated that, when fluent speech breaks down, the listener's attention moves towards the speaker to aid comprehension, rather than to their environment in search for a referent. Such findings raise new and intriguing theoretical questions as to what information is obtained from looks towards the speaker to help speech comprehension, and under what conditions (if any) listeners adopt an environment-oriented focus of attention. Are listeners using information from the speaker's facial expressions, eye gaze and/or gestures to help to disambiguate the sentence, or are they passively waiting for a disambiguation?

5. General discussion

The ability to predict future behaviour is fundamental to human cognition. In the domain of language research, there is now a large body of literature to support that listeners can predict upcoming linguistic input, which may help the rapid processing of speech (e.g. Ehrlich & Rayner, 1981; Schwanenflugel & Shoben, 1985) and allow for efficient turn-taking (Levinson, 2016). Recent work from our VR laboratory provided initial evidence that listeners not only predict in relatively artificial experiments, but that prediction is also engaged in more naturalistic environments (Heyselaar et al., 2020). In a series of three experiments, we here expanded on this work to test whether such findings can generalise to new stimuli, and to provide novel insights into whether subtle cues in speech, such as repair disfluencies and hesitations, can be used to inform one's predictions under naturalistic circumstances. Finally, we provide evidence to distinguish between different theoretical accounts of how repair disfluencies inform predictions.

Consistent with our hypotheses and the previous literature (Altmann & Kamide, 1999; Corley, 2010; Eichert et al., 2018; Heyselaar et al., 2020), in Experiment 1 there was an increase in the proportion of anticipatory fixations towards the referent before noun onset when the verb in the sentence was restrictive to a single object, but not when verb constraints were unrestrictive (i.e. the verb was applicable to multiple objects). Consistent with Heyselaar et al.'s (2020) findings, anticipatory fixations towards the target were significant from ~400 ms (368 ms) after verb onset and increased to a proportion of ~0.15 of all looks. The pattern of anticipatory target fixations in fluent sentences was consistent across Experiments 1, 2, and 3, and is plotted collapsed across experiments in Figure 14. Overall, the proportion of target fixations was lower compared to traditional VWP research (e.g. see Hintz et al., 2017 for verb mediated fixations), a pattern that is typically observed when presenting more complex visual displays (Eichert et al., 2018; Heyselaar et al., 2020; Sorensen & Bailey, 2007; Staub et al., 2012). Fewer fixations per object are particularly expected in VR compared to 2D displays, as participants are immersed in 3D scenes that are interesting to visually explore. Furthermore, in the current paradigm the speaker (i.e. the virtual agent) was present in the scene. The current data emphasise that it is natural for the listener to align with the person who is speaking during communication, rather than to objects in the environment (see also virtual agent fixations in Figure 4, Figures 6 and 10). Despite the overall lower proportion of target fixations in the current work, the proportion remains substantially and consistently higher in the Restrictive compared to Unrestrictive condition. As such, our results provide further confirmatory evidence that people indeed predict upcoming speech in naturalistic environments.

Experiments 2 and 3 investigated the extent to which different types of disfluencies in speech influence predictive eye movements. In Experiment 2 we compared the pattern of fixations in response to hearing sentences that contained a repair disfluency (e.g. " … cutting down uh moving the tree") with hearing sentences that contained a conjunction (e.g. " … cutting down and moving the tree"). Firstly, Experiment 2 replicated the findings of increased anticipatory target fixations in the Restrictive but not the Unrestrictive condition in single critical verb sentences, in a time window that preceded noun onset. Secondly, consistent with our

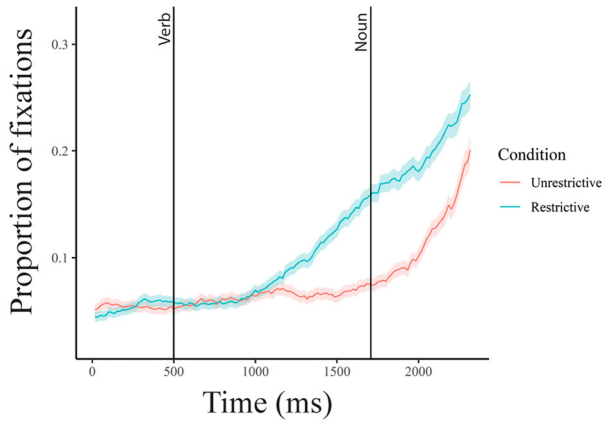


Figure 14. Proportion of fixations towards the target in fluent sentences collapsed across Experiments 1, 2 and 3. Vertical lines represent critical time points (mean verb onset and mean noun onset). Shaded ribbons represent standard error of the mean. Zero indicates 500 ms prior to verb onset.

hypotheses and expanding on the findings of Corley (2010), it provided novel findings that hearing a repair disfluency in naturalistic settings reduced the proportion of fixations towards the predicted item compared to when participants heard the same verbs conjoined with an "and". In contrast to Corley (2010), however, the proportion of target fixations did not *decrease* upon hearing a repair disfluency, which challenges whether listeners inhibit their prediction in response to a repair in naturalistic settings, or wait for the sentence to become disambiguated. Thirdly, contrary to our hypotheses and to a predictive account of repair disfluencies, we saw no indication that a new prediction was made based on the repaired verb, as there was no change in the proportion of fixations towards items that were compatible with the repaired verb (i.e. critical distractors) in the Repair condition. Instead, an increase in the proportion of fixations to the (virtual) speaker was observed, thereby supporting an attentional account of the influence of repair disfluencies on sentence comprehension.

In Experiment 3 we compared participants' target fixations in predictable (restrictive verb constraints) and unpredictable (unrestrictive verb constraints) sentences in both fluent utterances and after hearing a hesitation (e.g. " … cutting down uhh the tree"). Experiment 3 firstly replicated the findings of increased target fixations in restrictive but not unrestrictive sentences in a time window that preceded noun/hesitation onset (Fluent/Disfluent respectively). Consistent with our hypotheses, there was an increase in the proportion of target fixations in only the Restrictive-Fluent condition (see Figure 12 panel A) in a time window directly preceding noun onset, which crucially succeeded the onset of the hesitation in disfluent sentences. In contrast, after hesitation onset there was no longer an increase in the proportion of target fixations in the Restrictive-Disfluent condition. This pattern of fixations statistically differed from the Restrictive-Fluent condition, and qualitatively differed from the time windows both before the hesitation onset (when the prediction is strong) and after the noun onset (when the prediction is confirmed; see Figure 11), where the proportion of target fixations continued to increase (visualised in Figure 10). Furthermore, the hypothesis that there would be an increase in the proportion of fixations to the speaker over time in the Restrictive-Disfluent condition was supported. Hesitations in speech, therefore, seem to alter the trajectory of anticipatory fixations away from the predicted referent towards the speaker.

Together, Experiments 2 and 3 hence demonstrate that disfluencies in speech alter the trajectory of anticipatory fixations towards a predicted item and speaker, which indicates that disfluencies influence ongoing sentence processing. However, the cognitive mechanisms underlying this altered pattern of fixations remain open for discussion. In the following paragraphs we discuss the possible interpretations (attentional *vs* predictive accounts) of our findings in repair and hesitation disfluencies in turn.

5.1. Repair disfluencies

Previous research has been unable to determine whether listeners simply inhibit their prediction upon hearing a repair disfluency, or use information from the error to inform further predictions. Gaze shifts towards semantic and phonological competitors of an erroneous noun could either reflect an informed prediction, interpreting the error as an intrusion from a competitor (Karimi et al., 2019), or an automatic priming mechanism diverting attention to the erroneous noun's lexical competitors after the initial prediction is inhibited. Although Karimi et al. (2019) provided some evidence to support a prediction account, in that looks towards the semantic competitor were greater in the repair condition compared to an " … and also … " coordination condition, these findings could also be explained by the first noun being suppressed, allowing for a greater effect of semantic priming, in the repair condition but not the coordination condition. Contrary to a prediction account, the results of Experiment 2 provided no evidence that participants made a new prediction in response to the repair disfluency, as there was no increase in the proportion of fixations towards items that were compatible with the repaired verb (i.e. the critical distractors). Instead, there was an increased proportion

of fixations towards the virtual agent. Our findings, therefore, provide novel evidence that a repair disfluency drives the listener to realign their attention with the speaker. Such findings highlight the need to consider the presence of the speaker in theories of speech comprehension, and raise questions about what looks towards the speaker reflect. For example, the listener could be passively waiting for the sentence to become disambiguated, or searching the speaker's facial expression, eye gaze and gestures for visual cues to inform their predictions. As a third alternative, participants may indeed make a new prediction based on the repaired verb, but evidence for this could have been masked by the salience of looks towards the virtual speaker. Although this last alternative is impossible to determine from eye tracking data alone, we are currently investigating this possibility in a follow-up electroencephalography study. To create a naturalistic environment, in our paradigm the virtual speaker was visible in every scene. When a speaker's message becomes ambiguous, it may be natural for the listener to look towards the speaker for additional information. In contrast, if the speaker is not visible, listeners may instead attend to the environment to resolve the ambiguity. Future research could benefit from comparing listeners' eye movement behaviour when the speaker is visible compared to absent from (or occluded in) the scene to distinguish between passive versus predictive accounts of eye movement behaviour upon hearing a repair disfluency.

The current findings highlight an important ongoing question regarding the mechanisms underlying the prediction of language. It has recently been proposed that priming could be one of several mechanisms through which prediction can occur (Huettig, 2015; Pickering & Gambi, 2018). Huettig (2015) highlights that priming could be a fast, efficient and automatic mechanism through which prediction can take place. Karimi et al.'s (2019) findings of increased anticipatory fixations towards semantic and phonological competitors of an erroneous noun could therefore reflect prediction through this fast, automatic process. One important difference between the current work and Karimi et al. (2019) is that the repair was placed on the verb in the former and on a noun in the latter study. This meant that, whereas the purpose of the current study was to understand whether listeners use information from the *repaired* verb to update their predictions, Karimi et al. (2019) set out to establish the extent to which the listener uses information from the *error* to predict the *repair*. A strong semantic association between the erroneous word and the referent was therefore only present in Karimi et al. (2019). In contrast, in the current paradigm participants needed to perform a higher-level integration of semantic priors to form a new prediction based on the verb. Participants were required to evaluate which objects were compatible with the repaired verb before they could form a new prediction, thereby implementing a slow and effortful "active reasoning" approach to prediction (Huettig, 2015) via the production system (Pickering & Gambi, 2018; Rommers et al., 2020). It could therefore be that, upon hearing a repair disfluency, listeners are able to update their predictions with automatic (priming) predictive mechanisms (or "prediction-by-association"; Pickering & Gambi, 2018), but not with effortful predictive mechanisms that rely on semantic priors. This could particularly be the case in naturalistic environments, where there is a larger and richer context to consider. Future research should aim to determine whether listeners formulate a new prediction if given sufficient time, and if so, how much time is required.

5.2. Hesitation disfluencies

As outlined in the introduction, current theories of how hesitations influence ongoing sentence processing agree that eye gaze should move towards the speaker upon hearing a hesitation. However, fixations towards the speaker upon hearing a hesitation have not yet been measured. The pattern of fixations in Experiment 3 demonstrated that participants' eye gaze did indeed move towards the virtual agent, thereby supporting that current theories of how hesitations influence ongoing sentence processing hold in the presence of a speaker.

Which cognitive mechanisms underlie the shift in eye gaze towards the virtual agent remain an open question. It could be that the plateau in target fixations and increase in virtual agent fixations upon hearing a hesitation reflect a suppression of the weight placed on the listener's initial prediction. Listeners have been shown to place less confidence in the speaker's utterance if it is preceded by a hesitation (Brennan & Williams, 1995; Lowder & Ferreira, 2019). Moreover, there is evidence to suggest that listeners utilise the distribution of disfluencies occurring in natural speech to inform their predictions about upcoming utterances (Arnold et al., 2003, 2007, 2004). In the current work, participants may have interpreted the hesitation as a cue that the upcoming speech was more difficult to conceptualise and produce, and was unlikely to be the most predictable item present in the scene (see Introduction and section 4 for a detailed discussion). Importantly, previous work has shown that hesitations do not seem to influence predictions when there is an alternative

explanation for the hesitation, for example, if the participant is a non-native speaker (Bosker et al., 2014) or has object agnosia (Arnold et al., 2007). The linguistic processing of the virtual agent's speech, therefore, seems to have been similar to that of a native speaking human, supporting earlier evidence that people's linguistic behaviour does not substantially differ in interactions with human versus realistic virtual interlocutors (Heyselaar et al., 2017).

An alternative account for our current findings is that the participant's prediction did not change, but their visual attention was simply guided away from the referent. There are three mechanisms under which an attentional shift could occur. Firstly, it could be that the delay caused by the lengthy hesitation led to participants losing interest in the target object and averting their attention away from the referent. According to Huettig et al.'s (2011) model of the interaction of visual and linguistic information, information about a referent is stored in relation to a spatial location in visuospatial working memory. It is likely that once the participant has fixated on the object, they hold the item in visuospatial working memory and have no need to continue to look at the object. This first explanation is consistent with a temporal delay hypothesis of disfluencies (Corley & Hartsuiker, 2011; Wester et al., 2015). Secondly, rather than a loss of attention, it could be that the hesitation enhances attention, through an automatic, bottom-up-driven capture of attention to the salient interruption in the flow of speech. Finally, heightened attention could occur in anticipation of complex information after hearing a disfluency, either as a learned automatic response to pay attention after hearing a disfluency, or as a top-down driven response with the anticipation of complex information (Bosker, 2014; Fraundorf & Watson, 2014). An enhanced attention account is supported by both our current findings in the pattern of virtual agent fixations, in addition to findings of improved memory for information spoken after a disfluency (Collard et al., 2008; Corley et al., 2007; Fraundorf & Watson, 2011; MacGregor et al., 2010) and faster responses on a picture recognition task when the spoken picture name followed a disfluency (Corley & Hartsuiker, 2011).

5.3. Adapting to context

Earlier work suggests that listeners may flexibly adapt their predictive behaviour to distributional aspects of the linguistic input they receive (Bosker et al., 2019; Heyselaar et al., 2020). When comparing the paired critical verb conditions in Experiment 2, we observed that the model was improved when a random smooth for trial number was included. In an exploratory analysis (see Experiment 2 supplementary material) we, therefore, compared the change in the proportion of target fixations over time in the Repair and Conjunction conditions in early trials (first half of trials) and late trials (second half of trials). Although we found significant smooths for time for all conditions (see Table S1), the parametric effect of condition and the difference curve in Figure S1 panel C both demonstrate that there was a significantly greater proportion of target fixations in the late Repair trials compared to the early Repair trials (see also Figure 6). The smooths presented in Figure S1 panel A illustrate that the late Repair trials resembled a pattern of target fixations much closer to those of the Conjunction condition (both early and late trials). Throughout the experiment, participants seem to have learned that the reparation never led to a sentence ending that was incompatible with their initial prediction and adjusted their predictions accordingly. Similarly, when Heyselaar et al. (2020) reduced the number of predictable sentences to only 25%, participants stopped making anticipatory target fixations in the second half of the experiment. The authors proposed that participants learned that predicting the upcoming speech was no longer beneficial. Furthermore, Bosker et al. (2019) recently demonstrated that people quickly adapt to the distribution of disfluencies in speech within the present context. Increasing the proportion of disfluencies ("uh") that occurred before highly frequent words increased anticipatory fixations towards highly frequent referents. Our findings further corroborate that participants quickly adapt to the predictability of the current situation and rapidly adapt their predictive behaviour as a function of recent experience. We contribute novel insights by showing that listeners not only stop predicting when they learn it is no longer beneficial (Heyselaar et al., 2020), but also continue to predict when the current context renders a typically ambiguous sentence ending predictable.

5.4. Methodological considerations

VR provided a platform to present our stimuli with increased ecological validity while maintaining the high level of experimental control provided by a programmed experiment. The importance of studying language processing embedded within more naturalistic, interactive contexts is becoming increasingly salient (Hasson et al., 2018; Redcay & Schilbach, 2019). Until recently, theories promoting the prediction of linguistic input have been based on relatively artificial laboratory experiments that lack both the richness of the 3D visual world and the interactive, communicative

component of language. The current work provides confirmatory evidence that predictive eye movements are also made in rich, 3D, dynamic environments, when sentence stimuli are spoken to the listener by a virtual agent. Crucially, VR allowed us to go beyond the investigation of fixations towards the referent, and investigate fixations towards the speaker (i.e. the virtual agent). In doing so, we (a) provided novel empirical evidence supporting hypotheses of current theories of how hesitations effect ongoing sentence processing, and (b) separated competing theories of whether a listener uses the content of a repair disfluency to make new predictions. However, the current work provides only an early step in a trajectory of research into the prediction of upcoming linguistic input in naturalistic contexts. Prediction mechanisms are thought to utilise information from the different features of multimodal communication, including the speaker's gestures, facial expressions, eye gaze and posture (ter Bekke et al., 2020; Tromp et al., 2018). VR provides the scope to investigate how different features of multimodal communication are integrated to make predictions, while maintaining a high degree of experimental control.

In accordance with the naturalistic context rendered with VR, participants were not required to perform a task. Such a "look-and-listen" paradigm could have reduced participants' predictive behaviour compared to if participants were asked to actively identify or respond to objects mentioned by the speaker (see Corley & Hartsuiker, 2011; Wester et al., 2015 for examples of response-based paradigms). Despite any reduction in predictive behaviour that may have resulted from a free viewing rather than a response-based paradigm, we observed anticipatory eye movements in highly constraining sentences. However, under different task goals with an increased emphasis on prediction, listeners may also have displayed signs of making a new prediction after hearing a repair disfluency (as reflected in an increase in critical distractor fixations, rather than virtual agent fixations), which we failed to find here (see Experiment 2). Nevertheless, it is important to note that, although there are circumstances in natural speech in which a listener would be expected to identify the object mentioned by the speaker, in most of daily communication that is not the case. In light of the main goals of the paper, we preferred to keep our task instructions similar to everyday situations. Related to a lack of task goals, it has previously been suggested that look and listen paradigms could induce "good behaviour" in participants, where participants try to produce the behaviour that the experimenter is looking for. Magnuson (2019) argues that it is unlikely look and listen paradigms are particularly susceptible to strategy, as gaze shifts in visual scenes are a semi-automatic behaviour (Mishra et al., 2013). The present study indeed confirms the generalizability of earlier findings to rich, communicatively meaningful settings in which artificial tasks can be avoided.

5.5. Conclusions

There is an increasing body of literature to support that prediction is important for the rapid processing of speech. In a series of three VR experiments we here tested whether such findings hold in naturalistic settings (Experiment 1) and provided novel insights into whether disfluencies in speech, such as repair disfluencies and hesitations, can be used to inform one's predictions in naturalistic environments (Experiments 2–3). Experiment 1 provided further confirmatory evidence that listeners predict upcoming speech in naturalistic environments. Experiments 2–3 provided novel findings that disfluencies in speech alter the trajectory of anticipatory fixations towards a predicted referent in naturalistic environments. The proportion of target fixations was lower when hearing a repair disfluency (e.g. " … cutting down uh moving the tree") compared to when hearing an added verb (e.g. " … cutting down and moving the tree"). Similarly, after hearing a more ambiguous hesitation ("uhh") preceding a noun phrase, the pattern of target fixations was the same in the Restrictive (predictable) and Unrestrictive (unpredictable) conditions, in that there was no change in the proportion of target fixations over time. This contrasted both with fluent sentences and with the time window preceding the hesitation onset, where there was an increase in the proportion of target fixations over time in the Restrictive condition but not the Unrestrictive condition. Experiment 2 provided no evidence that participants made new predictions based on the repaired verb – there was no increase in the proportion of fixations towards objects compatible with the repaired verb but, instead, an increase in fixations towards the (virtual) speaker – thereby supporting an *attention* rather than a *predictive* account of effects of *repair* disfluencies on sentence processing. Experiment 3 provided novel evidence that the proportion of virtual agent fixations increased upon hearing a *hesitation*, supporting current theories of the effects of hesitations on sentence processing. Future research is needed to establish whether listeners indeed update their predictions, or whether this pattern of fixations merely reflects a shift in visual attention alone.

Acknowledgements

The authors would like to thank research assistants Iris Schmits and Eva Poort for helping to develop the Dutch sentence

stimuli and with data collection. The authors are grateful for the Max Planck Institute for Psycholinguistics' Technical Group, particularly Reiner Dirksmeyer and Albert Russel for providing technical support, and Jeroen Derks for programming the experiment and creating the visual stimuli. Many thanks to Evelien Heyselaar for sharing R scripts for data preprocessing and to Phillip Alday for statistical advice.

Disclosure statement

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

Funding

This work was supported by the NWO Gravitation Grant [grant number 024.001.006] to the Language in Interaction Consortium.

ORCID

Eleanor Huizeling http://orcid.org/0000-0001-7338-5966

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken Word Recognition using Eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

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., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518. https://doi.org/10.1016/j.jml.2006.12.004

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

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, Um, New information [journal article]. *Journal of Psycholinguistic Research*, *32*(1), 25–36. https://doi.org/10.1023/a:1021980931292

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Arnold, J. E., & Tanenhaus, M. K.. (2011). Disfluency effects in comprehension: How new information can become accessible. *The processing and acquisition of reference, 197-217*.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The Old and thee, uh, New:Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582. https://doi.org/10.1111/j.0956-7976.2004.00723.x

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of Age, relationship, topic, role, and gender. *Language and Speech*, *44*(2), 123–147. https://doi.org/10.1177/00238309010440020101

Bosker, H. R. (2014). Research Note: The processing and evaluation of fluency in native and non-native speech.

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language*, *75*, 104–116. https://doi.org/10.1016/j.jml.2014.05.004

Bosker, H. R., van Os, M., Does, R., & van Bergen, G. (2019). Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension. *Journal of Memory and Language*, *106*, 189–202. https://doi.org/10.1016/j.jml.2019.02.006

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

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

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 696–702. https://doi.org/10.1037/0278-7393.34.3.696

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Corley, M. (2010). Making predictions from speech with repairs: Evidence from eye movements. *Language and Cognitive Processes*, *25*(5), 706–727. https://doi.org/10.1080/01690960903512489

Corley, M., & Hartsuiker, R. J. (2011). Why Um helps auditory Word Recognition: The temporal delay hypothesis. *Plos One*, *6*(5), 1–6https://doi.org/10.1371/journal.pone.0019792.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., & Hart, J. C. (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM*, *35*(6), 64–72. https://doi.org/10.1145/129888.129892

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507–534. https://doi.org/10.1080/01690960143000074

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Memory and Language*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

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

Ferreira, F., & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, *8* (5), 231–237. https://doi.org/10.1016/j.tics.2004.03.011

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, *69*(3), 165–182. https://doi.org/10.1016/j.jml.2013.06.001

Ferreira, F., Lau, E. F., & Bailey, K. G. D. (2004). Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science*, *28*(5), 721–749. https://doi.org/10.1016/j.cogsci.2003.10.006

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, *40*(2), 280–295. https://doi.org/10.1006/jmla.1998.2613

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175. https://doi.org/10.1016/j.jml.2011.03.004

Fraundorf, S. H., & Watson, D. G. (2014). Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, *29*(9), 1083–1096. https://doi.org/10.1080/01690965.2013.832785

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135–157. https://doi.org/10.1016/j.cognition.2018.06.018

Heyselaar, E., Hagoort, P., & Segaert, K. (2017). In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior Research Methods*, *49*(1), 46–60. https://doi.org/10.3758/s13428-015-0688-7

Heyselaar, E., Peeters, D., & Hagoort, P. (2020). Do we predict upcoming speech content in naturalistic environments? *Language, Cognition and Neuroscience*, *36*(4), 1–22. https://doi.org/10.1080/23273798.2020.1859568.

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

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). 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

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

Karimi, H., Brothers, T., & Ferreira, F. (2019). Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions. *Cognitive Psychology*, *112*, 25–47. https://doi.org/10.1016/j.cogpsych.2019.04.001

Kukona, A., Altmann, G. T. M., & Kamide, Y. (2014). Knowing what, where, and when: Event comprehension in language processing. *Cognition*, *133*(1), 25–31. https://doi.org/10.1016/j.cognition.2014.05.011

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616. https://doi.org/10.1080/23273798.2015.1130233

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*(5), 633–666. https://doi.org/10.1080/01690960444000142

Levinson, S. C. (2016). Turn-taking in human communication – origins and Implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. Proceedings of the 2008 conference on empirical methods in natural language processing.

Lowder, M. W., & Ferreira, F. (2016). Prediction in the processing of repair disfluencies: Evidence from the visual-world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1400–1416. https://doi.org/10.1037/xlm0000256

Lowder, M. W., & Ferreira, F. (2019). I see what you meant to say: Anticipating speech errors during online sentence processing. *Journal of Experimental Psychology: General*, *148*(10), 1849–1858. https://doi.org/10.1037/xge0000544

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982–3992. https://doi.org/10.1016/j.neuropsychologia.2010.09.024

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, 3 (2), 113–139. https://doi.org/10.1007/s41809-019-00035-3

Mishra, R. K., Olivers, C. N. L., & Huettig, F. (2013). Chapter 8 - spoken language and the decision to move the eyes: To what extent are language-mediated eye movements automatic? In V. S. C. Pammi, & N. Srinivasan (Eds.), *Progress in Brain research* (Vol. 202, pp. 135–149). Elsevier. https://doi.org/10.1016/B978-0-444-62604-2.00008-3

Pan, X., & Hamilton, A. F. d. C. (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 [review]. *Frontiers in Human Neuroscience*, *9*(660), 1–19https://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 & Review*, *26*(3), 894–900. https://doi.org/10.3758/s13423-019-01571-3

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

Porretta, V., Kyröläinen, A.-J., van Rij, J., & Järvikivi, J. (2018). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In I. Czarnowski, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent decision technologies 2017* (pp. 268–277). Springer.

Rayner, K., Slowiaczek, M. L., Clifton, C., & Bertera, J. H. (1983). Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(6), 912–922. https://doi.org/10.1037/0096-1523.9.6.912

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Rommers, J., Dell, G. S., & Benjamin, A. S. (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition*, *198*, 1–8, 104206. https://doi.org/10.1016/j.cognition.2020.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*(3), 437–447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002

Ryskin, R., Levy, R. P., & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, *136*, 1–12, 107258. https://doi.org/10.1016/j.neuropsychologia.2019.107258.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89. https://doi.org/10.1016/S0010-0277(03)00139-2

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*(3), 362–367. https://doi.org/10.1037/0022-3514.60.3.362

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232–252. https://doi.org/10.1016/0749-596X(85)90026-9

Slevc, L. R., & Ferreira, V. S. (2013). To err is human; To structurally prime from errors is also human. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 985–992. https://doi.org/10.1037/a0029525

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Sorensen, D., & Bailey, K. (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

Steinberg, J., Truckenbrodt, H., & Jacobsen, T. (2012). The role of stimulus cross-splicing in an event-related potentials study. Misleading formant transitions hinder automatic phonological processing. *The Journal of the Acoustical Society of America*, *131*(4), 3120–3140. https://doi.org/10.1121/1.3688515

ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. *PsyArXiv*, https://doi.org/10.31234/osf.io/b5zq7.

Teruya, H., & Kapatsinski, V. (2019). Deciding to look: Revisiting the linking hypothesis for spoken word recognition in the visual world. *Language, Cognition and Neuroscience*, *34*(7), 861–880. https://doi.org/10.1080/23273798.2019.1588338

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 Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22. https://doi.org/10.1177/2331216519832483.

Van Rij, J., Wieling, M., Baayen, R. H., & Van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using GAMMs. *R package version*, *2*.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*(2), B63–B72. https://doi.org/10.1016/j.cognition.2005.07.001

Wester, M., Corley, M., & Dall, R. (2015). The temporal delay hypothesis: natural, vocoded and synthetic speech. *Proceedings DiSS Edinburgh, UK*.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and Generalized additive modeling. *Journal of Language Evolution*, *1*(1), 7–18. https://doi.org/10.1093/jole/lzv003

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.