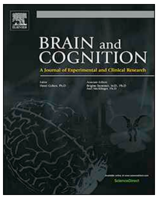
Brain and Cognition 135 (2019) 103571

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c



Influence of speakers' gaze on situated language comprehension: Evidence from Event-Related Potentials

Torsten Kai Jachmann*, Heiner Drenhaus, Maria Staudte, Matthew W. Crocker

Language Science and Technology, Campus C7, Saarland University, 66123 Saarbrücken, Germany
Cluster of Excellence Multimodal Computing and Interaction (MMCI), Campus E1.7, Saarland University, 66123 Saarbrücken, Germany

ARTICLE INFO

Keywords:
ERP
Speaker gaze
Referential expectation
N200
N400
P600

ABSTRACT

Behavioral studies have shown that speaker gaze to objects in a co-present scene can influence listeners' sentence comprehension. To gain deeper insight into the mechanisms involved in gaze processing and integration, we conducted two ERP experiments (N = 30, Age: [18, 32] and [19, 33] respectively). Participants watched a centrally positioned face performing gaze actions aligned to utterances comparing two out of three displayed objects. They were asked to judge whether the sentence was true given the provided scene. We manipulated the second gaze cue to be either Congruent (baseline), Incongruent or Averted (Exp1)/Mutual (Exp2). When speaker gaze is used to form lexical expectations about upcoming referents, we found an attenuated N200 when phonological information confirms these expectations (Congruent). Similarly, we observed attenuated N400 amplitudes when gaze-cued expectations (Congruent) facilitate lexical retrieval. Crucially, only a violation of gaze-cued lexical expectations (Incongruent) leads to a P600 effect, suggesting the necessity to revise the mental representation of the situation. Our results support the hypothesis that gaze is utilized above and beyond simply enhancing a cued object's prominence. Rather, gaze to objects leads to their integration into the mental representation of the situation before they are mentioned.

1. Introduction

In face-to-face spoken interactions, interlocutors are presented with rapidly unfolding information from the speech signal that is oftentimes accompanied by non-linguistic information. These speaker cues can be used to facilitate the understanding of the speaker's intended message.

For example, the gaze of a speaker toward objects present in a shared scene provides a visual cue that expresses the speaker's focus of visual attention and may draw the listener's attention as well (Emery, 2000; Flom, Lee, & Muir, 2007). Such cues can be used by the listener to ground and disambiguate referring expressions and infer the speaker's intentions and goals, which can facilitate comprehension (Hanna & Brennan, 2007).

As an example, one could imagine an everyday situation such as a breakfast scenario in which the table is set. When the speaker's gaze falls onto a mug while saying "Could you hand me …" already at this point the listener may anticipate the mug being the desired object. In the very same scenario, however, a contextually also valid continuation might be "…the plate." Such a situation, in which the continuation – even though being valid given the discourse – is not supported by visual cues, has been shown to lead to comprehension difficulties (Staudte & Crocker, 2011). However, the precise source of these difficulties is so far not entirely clear.

An eye-tracking study by Staudte and Crocker (2011) provided evidence that speaker gaze is interpreted by listeners as revealing speakers' referential intentions. In their study, participants were presented with videos of a robot performing gaze cues toward objects time-aligned to sentences that compared those objects with one-another, e.g. 'The cylinder is taller than the pyramid that is pink.' The target object additionally had a competitor that was a same type object with different size and color (e.g.: two pyramids were present in the visual scene). Thus, the linguistic point of disambiguation (LPoD) occurs only once the adjective is encountered. The gaze cues preceded the naming of the object by 1000 ms providing an early visual point of disambiguation (VPoD). The results showed that participants used that early VPoD to disambiguate the sentence as soon as the gaze cue was provided, expressed by a higher inspection rate of the gazed at object compared to the competitor before the onset of the spoken referent. Furthermore, a misleading gaze cue led to an elevated reaction time when judging whether the heard sentence was true or false given the visual scene.

As this and most other research regarding the influence of speakers' gaze on listeners' language comprehension has used behavioral

[*] Corresponding author.
E-mail address: jachmann@coli.uni-saarland.de (T.K. Jachmann).

https://doi.org/10.1016/j.bandc.2019.05.009
Received 17 January 2019; Received in revised form 24 May 2019; Accepted 25 May 2019

methods (e.g.: reaction times, eye-tracking), little is known regarding the mechanisms that support the integration of visual cues with speech. In order to better understand the neurocognitive processes that underlie the reported effects, we conducted two ERP-studies examining how listeners exploit speech-aligned speaker gaze to incrementally understand situated utterances. Specifically, we consider the extent to which gaze influences three stages of language processing that are known to be indexed by distinct ERP components, namely phonological expectations (N200), lexical retrieval (N400), and semantic integration and updating (P600).

We further outline predictions for these ERP components, based on two possible accounts of utterance and gaze cue integration: (a) a more shallow, prominence-based processing of gaze that simply increases the prominence of the gazed-at object, resulting in gaze cue driven reflexive focal shifts to the gazed-at location (Driver et al., 1999; Ricciardelli, Carcagno, Vallar, & Bricolo, 2013; Senju, Tojo, Dairoku, & Hasegawa, 2004), and (b) a deeper processing in which gaze cues are integrated with the linguistic signal (Staudte & Crocker, 2011). We interpret the latter in terms of a situational integration account in which listeners form a mental representation of the situation (Zwaan & Radvansky, 1998) that is rapidly updated and revised utilizing not only linguistic information but speech-aligned gaze cues as well.

Recalling the aforementioned breakfast scenario, based on previous work in the field, we outline how three stages of language processing identified above, would plausibly be affected when hearing ‘plate’ following a gaze toward the mug. Firstly, if gaze leads to a prediction that ‘mug’ will be heard, then a word which fails to confirm this expectation may be expected to result in an increase in N200 amplitude (Connolly, Stewart, & Phillips, 1990; Hagoort & Brown, 2000). Similarly, expectation for ‘mug’ might result in more difficult retrieval of ‘plate’ from semantic memory, as expressed by the N400 component (Kutas & Federmeier, 2011; Van Berkum, Koornneef, Otten, & Nieuwland, 2007). Finally, if gaze further informs the mental representation of the situation – such that the gaze to the mug is sufficient to instantiate ‘mug’ as the intended referent – then encountering ‘plate’ will entail a revision of this meaning representation, which has been shown to be indexed by the P600 component (Burkhardt, 2006, 2007). We outline below that, while the situational integration account predicts all of these effects, the prominence account would predict only the N400, and possibly N200 effect.

The N200 component as a Phonological Mapping Negativity (PMN) (Spivey, Joanisse, & McRae, 2012) has been previously observed when there is a mismatch between the expected word form given the context and the actual word candidates that are consistent with the speech signal listeners perceive (Connolly et al., 1990; Hagoort & Brown, 2000). In the breakfast scenario, such a mismatch would arise on the first phoneme of the uttered word ‘plate’ where the onset of ‘mug’ would be expected based on the preceding gaze cue. The presence of an N200 modulation would provide support for the situational integration account: Given that the N200 expresses a phonological mismatch, this component could only occur if a specific phoneme was expected. This, in turn, could only be the case if the name of the gazed-at object was retrieved before it is mentioned. If however gaze would merely increase an objects prominence, this would not necessarily be the case.

The N400 is known to be modulated by a word’s retrieval cost relative to its expectability in a context (Kutas & Federmeier, 2011; Van Berkum et al., 2007). As such, in the breakfast scenario, we expect an increased N400 when hearing ‘plate’ as it is less expected given the preceding gaze to the mug. However, factors that elicit an N400 modulation are more diverse than those related to the N200. The N400 effect has been shown to also be influenced by a more conceptual context such as music (Daltrozzo & Schön, 2009; Koelsch et al., 2004) as well as by expectations derived from a broader variety such as world knowledge (Hagoort, Hald, Bastiaansen, & Petersson, 2004), lexico-semantic information (Federmeier & Kutas, 1999) or information about the speaker (Van Berkum, 2009). As such, the expectations formed that influence the N400 are not necessarily constrained to a specific word form, whereas the N200 as a PMN would entail specific lexical predictions. Additionally, it has been shown that the occurrence of the N400 does not necessarily require strategic semantic processes but can even be elicited by unconsciously perceived stimuli (Kiefer, 2002). Therefore, in terms of the two accounts considered in this paper, we expect the N400 modulation to occur for both the prominence-based processing of gaze, as the object is more prominently attended to, as well as for the situational integration processing, as both prominence and linguistic expectations should modulate lexical retrieval. Recalling the aforementioned breakfast scenario, both accounts predict that encountering an unexpected/uncued object (‘plate’) would - in comparison to the gazed-at and, hence, expected/cued object (’mug’) - elicit increased N400 amplitude.

Although the P600 was originally linked to syntactic violations (see Friederici, 2002) more recent studies observed it’s occurrence in syntactically non-anomalous contexts and were able to link it to a semantic integration function (for instance Burkhardt, 2006; Burkhardt, 2007). While the N400 component could possibly be explained by both the shallow, prominence-based account and the situational integration account, only the latter additionally predicts integration difficulties with the mental representation of the situation for the mentioning of ‘plate’. Under this account, we assumed that any kind of information – linguistic information as well as information provided by gaze – is incrementally integrated into the mental representation as soon as it is provided. Information that can not be integrated with the already constructed representation in turn leads to the necessity of a reevaluation of that mental representation. In the aforementioned scenario, gaze to the ‘mug’ will cause this referent to be integrated into the utterance’s mental representation as the (anticipated) direct object. If ‘plate’, rather than ‘mug’, is then mentioned this mental representation must be revised. We expect this revision of the unfolding meaning representation to be expressed by an increased P600 (e.g., Burkhardt, 2007).

In the present experiments, we monitor listeners’ ERPs as they observe a stylized face performing gaze actions toward simple objects preceding their mentioning in a simultaneously presented utterance comparing objects in the scene with one-another. For example, participants were presented with a scene containing three different objects varying in depicted size; a house, a car and a t-shirt. They then heard sentences of the form ‘The car is bigger than the house, I think’ comparing two of the three present objects with the third remaining unmentioned (see time-line in Fig. 1 for the corresponding scene). The stylized depictions of the face and objects was used to reduce effects of visual complexity on the speaker, while the simple comparisons that listeners were presented with were used to avoid a preference for certain objects based on the linguistic content of the sentence.

Previous eye-tracking studies have shown that, if visual context is provided, speakers direct their gaze toward an object about 800–1000 ms before mentioning it (Griffin & Bock, 2000; Kreysa, 2009; Meyer, Sleiderink, & Levelt, 1998). In our experiments, we used these findings to place the occurring gaze actions in a natural way, so that they precede the mentioning of an object by 800 ms.

Furthermore, the gaze cue preceding the second object in the sentence was either Congruent (directed toward the consequently mentioned object), Incongruent (toward the unmentioned object) or Uninformative (Averted - toward the bottom of the screen – in experiment 1, and Mutual – redirected toward the listener – in experiment 2), with the Congruent condition serving as a baseline. This manipulation is intended to shed light on how listeners exploit speakers’ gaze to anticipate and/or integrate mentioned referents.

In summary, we hypothesized that gaze modulates listeners’ expectations for a referent to be mentioned, possibly even anticipating a specific word, predicting the modulation of three established ERP components. Previous research has shown that, if a specific word form is predicted, an attenuated N200 is observed when this prediction is

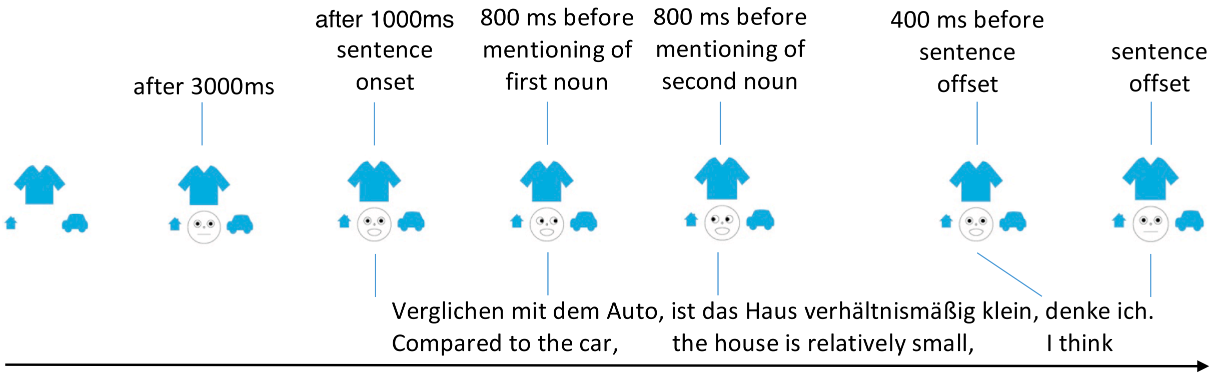

after 3000ms
after 1000ms sentence onset
800 ms before mentioning of first noun
800 ms before mentioning of second noun
400 ms before sentence offset
sentence offset
Verglichen mit dem Auto, ist das Haus verhältnismäßig klein, denke ich.
Compared to the car, the house is relatively small, I think

Fig. 1. Timeline of an item in the Congruent condition in Experiment 1.

confirmed (Connolly et al., 1990; Hagoort & Brown, 2000). Based on the interpretation of the N200 as a Phonological Mapping Negativity (Spivey et al., 2012), we hypothesized that the N200 effect is driven by the amount of information conveyed by the incoming phoneme. If the perceived signal is consistent with the name of the gaze-cued referent (Congruent) – hence, conveying little new information – we expected an attenuated N200. In the remaining two conditions however the phoneme is providing new information, be it to select the correct target out of a set of still active candidates (Uninformative) or to disconfirm the currently active, expected candidate (Incongruent). We argue that this effect would be more reasonably explained by the situational integration account than by the prominence account, as the prediction of a specific word form suggests that gaze not only cues an object, but also results in naming that object.

We further predict that Congruent gaze leads to facilitated retrieval of the named object compared to Incongruent and Uninformative gaze as either expectations that were formed based on the gaze cue toward an object are violated (Incongruent) or no expectations could be formed (Uninformative). The lexical retrieval difficulty for the Incongruent and Uninformative conditions is hypothesized to be expressed by an increase in N400 amplitude (Kutas & Federmeier, 2011; Van Berkum et al., 2007). We argue that both the prominence account as well as the situational integration account would equally predict this effect. Either because the named object is outside of the listeners' focus of attention, or because linguistic processing of the gazed-at object encouraged stronger expectations that were violated respectively.

A summary of the predictions for each of the discussed components in relation to the proposed accounts can be found in Table 1.

Table 1
Summary of the predictions for each discussed component in relation to the proposed accounts. +: modualtion explained by account, –: modulation not expected.

Account	N200	N400	P600
Prominence	(+)	+	–
Situational Integration	+	+	+

2. Experiment 1

In the first experiment, German native speakers judged whether a sentence played to them was true given a visual context while their EEG was recorded. Each trial contained a stylized face that performed gaze actions timed to the sentence that was to be evaluated. Trials were constructed following the before-mentioned criteria, so that every gaze action was performed 800 ms prior to the naming of the corresponding noun. The first gaze was always congruent toward the object that was named first in the sentence, whereas the second gaze was either toward the second named object (Congruent), toward a distractor object that remained unmentioned throughout the course of the trial (Incongruent) or toward the bottom of the screen where no object was situated (Averted). An example of a screen presented in the experiment can be found in the time-line depicted in Fig. 1.

2.1. Participants

Forty-five right-handed native speakers of German (Mean age: 24; Age range: [18, 32]; SD: 3.39; Male: 8; Female: 37) took part in the ERP experiment. 15 participants were removed from the analysis due to their behavioral data (3) and too high numbers of eye artifacts (12).[1] Participants gave informed consent. All participants had normal or corrected-to-normal vision and had no hearing problems. All participants were compensated with €15 for their participation.

2.2. Stimulus materials and procedure

We created 60 pictures of objects of masculine (25), feminine (18) and neuter grammatical gender (17) in respect to their naming in German. Those pictures were presented to seven participants using Google Forms with the task to name the objects and indicate how complex they appear to the participant on a scale from 1 (low complexity) to 5 (high complexity). Out of the 47 objects that were named identically by all participants, we chose eight objects per gender for the experiment. In the selection of the objects, we further used the complexity rating of the participants, so that only objects with a similar complexity ranging from 1.5 to 2.5 were chosen. All objects used are summarized in Table 2.

Participants were presented with a picture containing three objects of the same gender that varied either in size or shading arranged in positions above, left and right of the center of the screen. Each screen contained a large, medium sized, and small object (or light, medium, and dark object respectively). After 3000 ms, a stylized face appeared in the middle of the screen with a straight gaze toward the participant. The face then performed gaze actions timed to an auditory presented sentence of the form "Verglichen mit dem Auto, ist das Haus verhältnismäßig klein, denke ich" (Compared to the car, the house is relatively small, I think). The utterance was a synthesized German sentence using the CereVoice TTS system's Alex voice (Version 3.2.0). We created different versions of example utterances that varied in intonation contour and turn internal pause length. A Google Form was used to collect responses of seven participants, who listened to those examples with the task to rate their naturalness and order them from most natural to least natural. We selected the version with the most natural rating for the experiment. The nouns had an average length of 560 ms with the shortest noun lasting for 383 ms (Tisch) and the longest noun lasting for 791 ms (Flugzeug).

[1] For a concrete description of the removal see Section 2.3 Data Analysis.

Table 2
Summary of the objects presented in Experiment 1 with their English translation separated by grammatical gender.

Masculine	Feminine	Neuter
Baum (tree)	Blume (flower)	Auto (car)
Blitz (bolt)	Brezel (pretzel)	Blatt (leaf)
Fisch (fish)	Gießkanne (watering can)	Boot (boat)
Handschuh (glove)	Hand (hand)	Flugzeug (airplane)
Hut (hat)	Lampe (lamp)	Haus (house)
Stern (star)	Maske (mask)	Kreuz (cross)
Stiefel (boot)	Tasche (bag)	Rad (wheel)
Tisch (table)	Wolke (cloud)	T-Shirt (t-shirt)

On speech onset, the face retained its straight gaze but opened the mouth to evoke the impression of the face being the speaker of the sentence. The mouth remained in this position for the time the sentence was spoken. The first gaze cue appeared approximately 800 ms before the first noun was mentioned. This gaze cue was always Congruent toward the first named object for all experimental trials. Also, in order to ensure the participants' attention throughout the entire sentence, the first named object in the experimental items was always the medium sized/shaded object. An example of an experimental trial can be seen in Fig. 1, which displays the time line of a congruent trial, containing a small house, medium sized car and a large t-shirt. There, if the first gaze action were directed toward the t-shirt, both remaining objects would be smaller and, hence, would no longer require the participant to pay attention to the upcoming noun in order to evaluate the sentence. The second, manipulated gaze cue appeared 800 ms prior to the onset of the second noun. The gaze was redirected toward the participant 400 ms before the end of the sentence, and the mouth closed on the offset of the sentence. Each item appeared in three conditions (Congruent (baseline)/Incongruent/Averted). In the Congruent condition, the gaze preceding the second noun was directed toward the subsequently named object (Haus). In the Incongruent condition, the gaze cue was instead directed toward the object that remained unmentioned in the sentence (T-Shirt). In the Averted condition, the gaze was directed toward the bottom of the screen where no object was present. This led to three lists using a latin square design. Additionally, we created versions of those lists that were counterbalanced for realism. Realism was defined based on the truth value of the performed utterance in the real world. For example, in the experiment, some trials contained utterances like "compared to the car, the house is relatively small, I think". In the real world, such a statement would usually be false. Therefore, such 'unrealistic' statements were counterbalanced with their 'realistic' version (e.g. "compared to the house, the car is relatively small, I think"). This counterbalancing also led to a swap of the size of the named objects in the visual scene, resulting in a total of six lists. Each list contained 72 experimental items (24 per condition) and 72 fillers that mentioned an object other then the medium object as the first noun, and gaze patterns different from the gaze patterns in the experimental items. Both the experimental items as well as the filler items contained the same number of true and false statements relative to the visual scene. Importantly, however, the truth value of the sentence was not revealed before the naming of the adjective at the end of the clause. Hence, neither the gaze region nor the noun region are affected.

25% of the fillers (18) contained a manipulation of the first gaze cue instead of the second gaze cue. This subset of the fillers still started with a mentioning of the medium object as the first noun in the sentence. However, the first gaze cue was always directed toward the empty position. We didn't use an incongruent first gaze cue in order to maintain the overall reliability of the gaze cues. The remaining fillers were of the same form as the experimental items with the difference that the first mentioned object was either the small or large (light/dark) object, followed by the naming of either of the remaining two objects. The gaze patterns performed on these fillers always started with a congruent gaze, as in the experimental items, followed by another congruent gaze toward the second named object half of the time (36) and a quarter of the time by either an incongruent or averted gaze cue (9/9). This distribution of gaze patterns throughout the experiment led to an overall ratio of congruent gaze actions of 70.8% (204). Every trial contained two gaze actions, one preceding the first noun and one preceding the second noun, the total number of gaze actions throughout the course of the experiment was 288 per list/participant. Another 17.7% (51) of the gaze actions were Averted and only about 11.5% (33) of the gaze actions were Incongruent. This way, the validity of the gaze cue was kept high in order to avoid that participants would start to ignore the gaze cues altogether throughout the course of the experiment.

The stimuli were presented using the E-prime software (Version 2.0.10. Psychology Software Tools, Inc.). Each participant was seated in a sound-proof, electro-magnetically shielded chamber in front of a 24" Dell U2410 LCD monitor (resolution of 1280 × 1024 with a refresh rate of 75 Hz). The distance between the participant and the screen was always 100 cm in order to keep all objects in a 5° visual angle from the center of the screen. This was done to minimize eye-movements throughout the experiment. While the participants were prepared for the recording, they were presented with all objects that occurred throughout the experiment and their naming. The Alex voice of the CereVoice TTS was also used for the naming of the objects. After this, participants were presented with written instructions and completed six practice trials. The items were pseudo randomized for each list and presented in 7 blocks with fs after each block. After each item, the participants were asked to indicate whether the sentence was true given the visual context they were presented with by pressing one of two buttons to assure the participants attention. Answers were recorded using a Response Pad RB-834 (Cedrus Corporation). The experiment lasted approximately 45 min.

2.3. Data analysis

The EEG was recorded by 24 Ag/AgCl[2] scalp electrodes (actiCAP, BrainProducts) and amplified with a BrainAmp (BrainVision) amplifier. Electrodes were placed according to the 10–20 system (Sharbrough et al., 1991). Impedances were kept below 5 kΩ. The ground electrode was placed at AFz. The signal was referenced online to the reference electrode FCz and digitized at a sampling rate of 500 Hz. The EEG files were re-referenced offline to the average of the mastoid electrodes. The horizontal electrooculogram (EOG) was monitored with two electrodes placed at the right and left outer canthi of each eye and the vertical EOG with two electrodes below both eyes paired with Fp1 and Fp2. During recording an anti-aliasing low-pass filter of 250 Hz was used. The EEG data was band pass filtered offline at 0.01–40 Hz in order to attenuate skin potentials and other low voltage changes as well as line noise and EMG noise (Luck, 2014). Single-participant averages were computed for a 1100 ms window per condition relative to the acoustical onset of the noun following the manipulated gaze cue. All segments were aligned to a 100 ms pre-stimulus baseline. We semi-automatically screened offline for electrode drifts, amplifier blocking, eye-movements and muscle artifacts.

Due to the nature of the task and the experimental setup containing various eye-movements performed by the displayed face, the number of eye artifacts was relatively high. Therefore, we set a threshold of 30% rejection rate per condition for participant exclusion (i.e., participants' data with more than 7 rejected trials out of 24 in one or more conditions were entirely removed). This led to the removal of 12 participants from the analysis. Additionally, participants' data was removed if they gave wrong answers to more than 10% of the questions. However, this

[2] This excludes the electrodes used for the electrooculogram and offline re-reference: Fp1, Fp2, T7, T8, TP9, TP10, PO9 and PO10.

was only the case for two participants that had already been removed due to eye artifacts. Overall, participants performed very well in the task with an average of 94.8% of correct answers. There was no difference in accuracy between conditions ($F(2, 58) = 0.96$, $p = .39$). After artifact rejection and participant exclusion 85% of the trials on average per participants were included in the analyses. The averaged data of the remaining 30 participants (Mean age: 23.7; Age range: [18, 32]; SD: 3.49; Female: 26) was exported using BrainVision Analyzer (Version 2.1) BESA export function. We analyzed the onset of noun following the manipulated gaze cue. We used R (R Core Team, 2015) to perform repeated measures analysis of variance (ANOVA) using Greenhouse-Geisser correction. We report F values, Greenhouse-Geisser corrected p values and η_p^2 (partial eta-squared) values as a measure of effect size. All ANOVAs were computed on the F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1 and O2 electrodes including ROIs for frontal (F3, Fz, F4, FC5, FC1, FC2, FC6), central (C3, Cz, C4, CP5, CP1, CP2, CP6) and posterior (P7, P3, Pz, P4, P8, O1, O2) distributions.

2.3.1. *Results*

We conducted an analysis of the mean amplitudes of the differences between the three experimental conditions (Congruent, Incongruent, Averted) time locked to the onset of the second noun, for which the preceding gaze cue was manipulated. Each reported ANOVA was computed with experimental condition (3-levels) and electrode site (frontal, central, parietal) as within-subject factors. Similar to other studies presenting auditory stimuli consisting of continuous speech (Connolly et al., 1990, Connolly, Phillips, Stewart, & Brake, 1992; Hagoort & Brown, 2000; O'Halloran, Isenhart, Sandman, & Larkey, 1988), we did not find the N100-P200 complex, which is a usual response to the abrupt onset of auditory stimuli.

Visual inspection revealed that the Averted condition contains two distinct, mostly frontally distributed peaks within this time-window (see Fig. 3 for comparison). In order to isolate the involved components and to establish the time-windows for the analyses more precisely, we followed the approach utilizing difference waves as proposed by Kappenman and Luck (2016). We created an Incongruent-minus-Congruent difference wave as well as an Averted-minus-Congruent difference wave that can be found in Fig. 2. The two distinct peaks revealed in the visual inspection were also globally distinguishable in the Averted-minus-Congruent difference wave. This is consistent with the predictions to find distinct N200 and N400 patterns and is supported by previous findings, e.g., Connolly et al. (1990) and Hagoort and Brown (2000). Taking both difference waves into account, the established time-window for the N200 lies between 150 and 300 ms which is in line with findings from Praamstra and Stegeman (1993). The N200 is followed by the N400 time-window lasting from 300–450 ms, which falls into the typical N400 time-window (300–500 ms). Lastly, the time-window for the P600 was established between 600 and 900 ms which is consistent with previously established time-windows for the P600 (Burkhardt, 2006; Brouwer, Crocker, Venhuizen, & Hoeks, 2017).

An ANOVA of the N200 time-window between 150 and 300 ms showed a main effect of condition ($F(2, 58) = 5.91$, $p < .01$, $\eta_p^2 = 0.17$). There was a globally distributed, significantly larger negativity for both the Incongruent ($M = -2.76\,\mu\text{V}$, $SD = 1.87$) and Averted ($M = -2.34\,\mu\text{V}$, $SD = 1.75$) condition compared to the Congruent ($M = -1.41\,\mu\text{V}$, $SD = 1.39$) baseline (($F(1, 29) = 10.33$, $p < .01$, $\eta_p^2 = 0.26$) and ($F(1, 29) = 7.55$, $p < .05$, $\eta_p^2 = 0.2$) respectively). An ANOVA of the following N400 time-window between 300 and 450 ms also showed a main effect of condition ($F(2, 58) = 3.37$, $p < .05$, $\eta_p^2 = 0.1$), with a globally distributed, significantly larger negativity for both the Incongruent ($M = -1.96\,\mu\text{V}$, $SD = 2.01$) and Averted ($M = -2.04\,\mu\text{V}$, $SD = 2.32$) condition compared to the Congruent ($M = -1.12\,\mu\text{V}$, $SD = 1.69$) baseline ($F(1, 29) = 5.79$, $p < .5$, $\eta_p^2 = 0.17$ and $F(1, 29) = 4.26$, $p < .05$, $\eta_p^2 = 0.13$ respectively). In order to assess whether the distinctiveness of the N200 and N400 components is also statistically supported, we additionally ran an ANOVA for the time-window between 250 and 350 ms. Indeed, no effect of condition was found in this time-window ($F(2, 58) = 2.89$, $p = .07$) supporting the previously determined time-windows. The analysis of the P600 time-window as well revealed a main effect of experimental condition ($F(2, 58) = 5.69$, $p < .01$, $\eta_p^2 = 0.16$). A pairwise analysis of the conditions to the Congruent baseline ($M = -1.28\,\mu\text{V}$, $SD = 2.23$) showed that the P600 modulation is only present for the Incongruent ($M = -0.22\,\mu\text{V}$, $SD = 2.99$) condition ($F(1, 29) = 8.49$, $p < .05$, $\eta_p^2 = 0.23$), but not in the Averted ($M = -1.7\,\mu\text{V}$, $SD = 3.33$) condition ($F(1, 29) = 0.1$, $p = .75$). Table 3 summarizes the findings for the reported time-windows.

2.4. *Discussion*

Research from Koornneef and Van Berkum (2006) and Van Berkum et al. (2007) suggests that comprehenders generate expectations about the unfolding sentence based on the previously gathered information that they integrated in a situation model (Zwaan & Radvansky, 1998). Various studies further suggest that not only linguistic information is used to form such expectations about upcoming sentence content but also visual information provided by the combination of provided scene and gaze cues (Ferreira, Foucart, & Engelhardt, 2013; Staudte & Crocker, 2010, 2011; Staudte, Crocker, Heloir, & Kipp, 2014). It is therefore reasonable to assume that this visual information also contributes to form the situation model. We interpret the earlier peak (150–300 ms) as an N200 reflecting an auditory matching process that is driven by the amount of information the incoming phoneme contains. This results in an attenuated N200 for the Congruent gaze condition only as no new information is provided. In both the Averted and Incongruent condition however, the phoneme provides additional information. In the former case by supporting only one of two possible referents, and in the latter case by mismatching with the expected word form due to the highly lexically specific expectations the gaze cues elicit. The N400 (300–450 ms) is interpreted to be reflecting a word's retrieval cost, influenced by how strongly supported or expected a word is given a visual context, such as situated gaze. Finally, we interpret the P600 to be reflecting the cost of revising the situational model formed on prior contextual information. Taken together, the results provide support for the situational integration account over the prominence account. While the effects in the N400 time-window could equally be explained by both accounts, the other time-windows provide evidence for both the prediction of a specific word form (N200) and the integration of the gazed-at object in the mental situational representation (P600) supporting the situational integration account.

In order to assure that the results from Experiment 1 were replicable and robust, we ran a follow-up experiment with some changes to address possible concerns, which will be discussed in greater detail in the following section.

3. Experiment 2

In this follow-up experiment, we adjusted the positioning of the objects relative to the face. As the cross-wise positioning in Experiment 1 (up, down, left and right of the face) led to significantly different ERP responses when participants were presented with the gaze cues, we shifted to a diagonal positioning. All objects in Experiment 2 were positioned 30° above and below the horizontal axis to the eyes of the face (see time-line in Fig. 4 for comparison). The new positioning of the objects also required an increased distance between the participant and the screen of 114 cm in order to keep all objects in a 5° visual angle from the center of the screen. In line with the change in object positioning, the empty position also rotated through the four possible

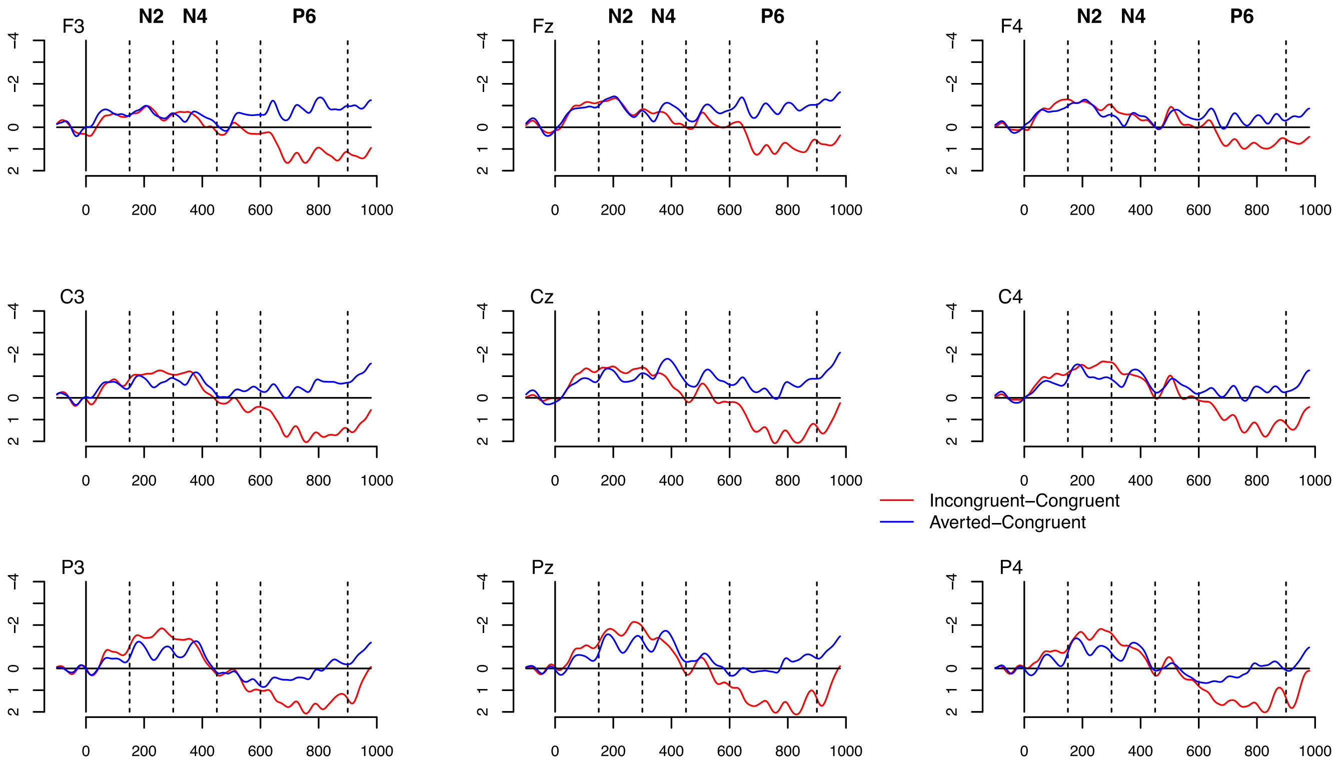


Fig. 2. Difference waves of Incongruent-minus-Congruent (red) and Averted-minus-Congruent (blue). The data presented shows the electrode subset F3, Fz, F4, C3, Cz, C4, P3, Pz and P4 filtered at 20 Hz for presentation purposes only. Negativity is plotted upwards. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

positions in Experiment 2 instead of being always below the face as in Experiment 1. We used the same objects as in Experiment 1 with one change. As we had similar onsets for the words Stern (star) and Stiefel (boot) in the first experiment, we exchanged the star with the equally well-performing object Mond (moon). We define ’well-performing’ as similar complexity ratings in combination with all participants naming

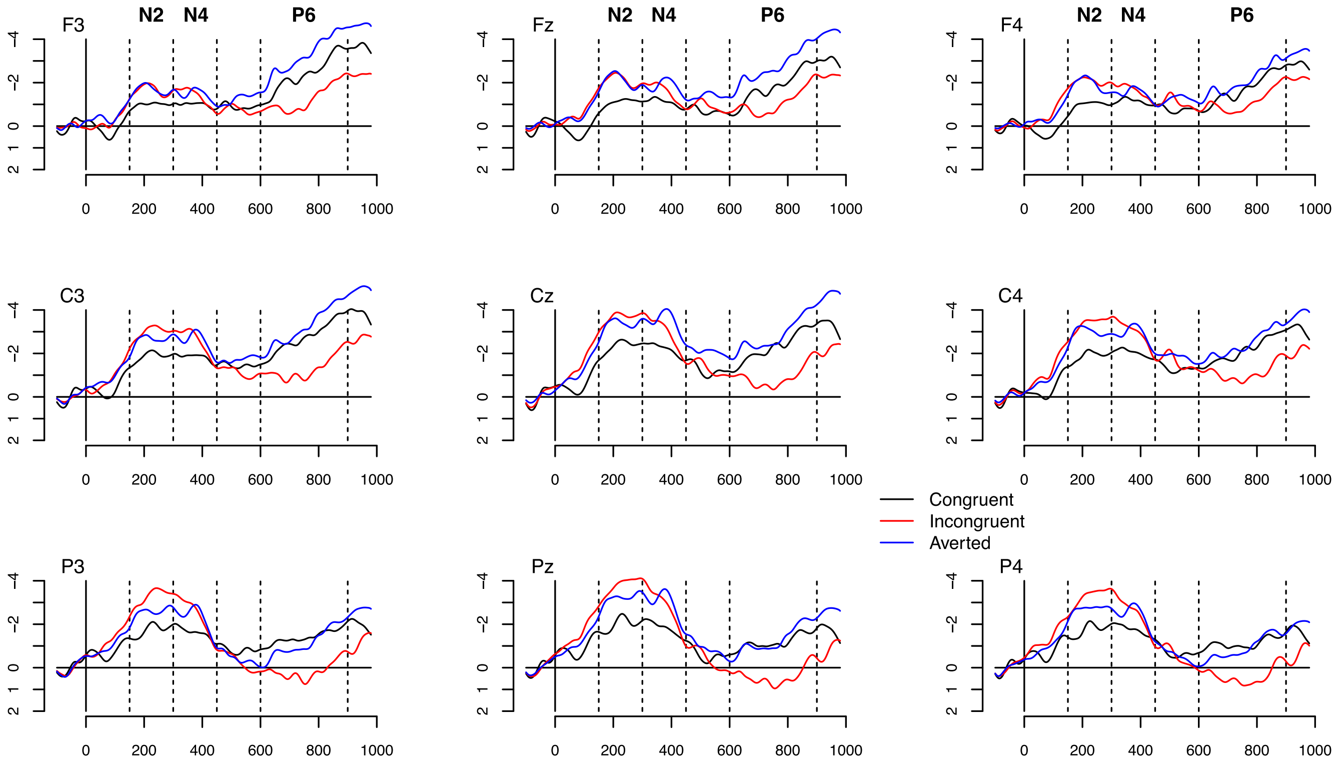


Fig. 3. ERP time-locked to the Second Noun Onset in Experiment 1 separated by the Experimental Conditions (Congruent (black), Incongruent (red) and Averted (blue)). The data presented shows the electrode subset F3, Fz, F4, C3, Cz, C4, P3, Pz and P4 filtered at 20 Hz for presentation purposes only. Negativity is plotted upwards. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Summary of the statistics in Experiment 1 (left) and 2 (right). C - Congruent , I - Incongruent , M - Mutual, A - Averted, – - not significant , ↑ - medium effect (η_p^2>.06) , ↑↑ - strong effect (η_p^2>.14) (Cohen, 1988).

	Experiment 1			Experiment 2		
	N200	N400	P600	N200	N400	P600
C : I	↑↑	↑↑	↑↑	↑↑	↑↑	↑↑
C : M				↑↑	–	–
C : A	↑↑	↑	–	↑↑	↑	–

the object identically in the pre-test.

The sentences presented to the participants were of the same form as in Experiment 1 (e.g., "Verglichen mit dem Haus, ist das Auto verhältnismäßig klein, denke ich" Compared to the house, the car is relatively small, I think). Fig. 4 shows a time-line of a Congruent trial in Experiment 2.

It is also perhaps debatable how 'uninformative' a gaze cue toward a position is, even if that position does not contain an object (Averted). In order to address this concern, we introduce an additional version of an 'uninformative' gaze cue to the three conditions as present in the first experiment. Different from the previous Averted gaze cue, this version moves the eyes of the face back to the straight gaze position instead of directing gaze to the empty position. The new gaze cue redirected straight toward the listener will be referred to as Mutual.[3] The Averted gaze cue from Experiment 1 with the gaze being directed to the empty position was demoted to a control condition that was added as a filler type. In order to still achieve comparable data, the number of this filler type was matched with the number of items per experimental condition. Each item in Experiment 2 appeared in three conditions (Congruent/Incongruent/Mutual) with an additional filler type that provides comparability between the two experiments (Averted). This led to three lists using a Latin square design. Additionally, we created versions of those lists that were counterbalanced for the truth value of a sentence. This means that the scene displayed in Fig. 4 was paired with two sentences: (1) "compared to the house, the car is relatively small, I think" and (2) "compared to the house, the car is relatively big, I think". This led to a total of six lists. As we found no effect of realism in the first experiment, this was no longer counterbalanced across lists in the second experiment. It should be noted though that the number of 'realistic' and 'unrealistic' sentences within a list was still balanced.

Each list contained 126 experimental items (42 per condition) and 126 fillers. 42 fillers were created identical to the Averted condition in Experiment 1 to retain comparability. This means that the first gaze was always congruent toward the object named first in the sentence followed by a gaze toward the empty position 800 ms before the mentioning of the second noun. In the remaining fillers (84) the first gaze cue was manipulated followed by a Congruent second gaze cue. In these fillers, the first gaze cue was either directed toward the empty position (42) or toward the object that remained unmentioned throughout the course of the sentence (42).

This distribution of gaze actions additionally led to a slight adjustment of the reliability of the gaze cue. The overall percentage of Congruent gaze actions was lowered from 70.8% to 58.3%, whereas the Incongruent and uninformative gaze actions – represented by Mutual and Averted gaze actions – both were increased (from 11.5% to 16.6% and from 17.7% to 25% respectively). We used these adjustments to test for the robustness of the effects found in the first experiment.

We hypothesized that we would replicate the modulations in the N200, N400 and P600 time-windows found in the first experiment. Specifically, we expected a stronger modulation of the N200 and N400 in the Mutual and Incongruent conditions compared to the Congruent condition related to the expectability of the noun given the visual context. Additionally, we predicted a P600 in the Incongruent condition related to the necessity to revise the mental model formed by utilizing the visual input.

3.1. Participants

Forty-four right-handed native speakers of German, who did not participate in Experiment 1 (Mean age: 24.6; Age range: [18, 35]; SD: 3.65; Female: 34), took part in the ERP experiment. 14 participants were removed from the analysis due to their behavioral data (4), technical errors (3) and too high numbers of eye artifacts (7).[4] Participants gave informed consent. All participants had normal or corrected-to-normal vision and had no hearing problems. All participants were compensated with €15 for their participation.

3.2. Data analysis

The technical setup and EEG recording sites were the same as in Experiment 1 (see Section 2.3 for comparison).

We kept the 30% threshold for the rejection rate per condition for participant exclusion due to eye-movements and other artifacts. This led to the removal of 7 participants from the analysis. Additionally, the data of 4 participants was removed due to their behavioral data with more than 10% of wrong answers to the question. Again, remaining participants performed very well in the task with an average of 97.4% of correct answers. As in experiment 1, there was no difference in accuracy between conditions ($F(2, 58) = 1.98$, $p = .15$) Another 3 participants had to be removed due to technical errors. Overall, the three criteria led to the removal of the data of 14 participants. After artifact rejection and participant exclusion 94.3% of the trials on average per participants were included in the analyses. The analysis of the data of the remaining 30 participants (Mean age: 24.3; Age range: [19, 33]; SD: 3.2; Female: 24) was conducted in the same way as in Experiment 1 (see Section 2.3 for comparison).

3.2.1. Second noun

We conducted an analysis of the mean amplitudes of the differences between the three experimental conditions (Congruent, Incongruent, Mutual) time locked to the onset of the second noun, for which the preceding gaze cue was manipulated. We used the same time-windows as established in the first experiment for the analyses.

N200 (150–300 ms): We again found a main effect of experimental condition ($F(2, 58) = 14.63$, $p < .01$, $\eta_p^2 = 0.33$). There was a globally distributed, significantly larger negativity for both the Incongruent ($M = -2.08\ \mu\text{V}$, $SD = 1.09$) and Mutual ($M = -1.84\ \mu\text{V}$, $SD = 1.19$) condition compared to the Congruent ($M = -1.04\ \mu\text{V}$, $SD = 1.26$) condition (($F(1, 29) = 27.49$, $p < .01$, $\eta_p^2 = 0.49$) and ($F(1, 29) = 8.15$, $p < .01$, $\eta_p^2 = 0.22$) respectively).

N400 (300–450 ms): Similar to the first experiment, we found a main effect of experimental condition ($F(2, 58) = 7.72$, $p < .01$, $\eta_p^2 = 0.21$). Consistent with the first experiment, the Incongruent ($M = -1.65\ \mu\text{V}$, $SD = 1.37$) condition is significantly more negative in this time-window compared to the Congruent ($M = -0.89\ \mu\text{V}$, $SD = 1.79$) condition ($F(1, 29) = 12.55$, $p < .01$, $\eta_p^2 = 0.3$). However, the newly introduced Mutual ($M = -1.18\ \mu\text{V}$, $SD = 1.77$) condition utilizing a straight gaze back to the listener does not significantly differ from the Congruent condition ($F(1, 29) = 1.44$, $p = .24$).

P600 (600–900 ms): Experimental condition showed a main effect ($F(2, 58) = 10.38$, $p < .01$, $\eta_p^2 = 0.26$). As in the first experiment, only the Incongruent ($M = 0.6\ \mu\text{V}$, $SD = 1.9$) condition shows a significantly

[3] Both Averted and Mutual gaze are considered to be 'uninformative', in contrast with object-directed gaze (Congruent/Incongruent).

[4] For a concrete description of the removal see Section 3.2 Data Analysis.

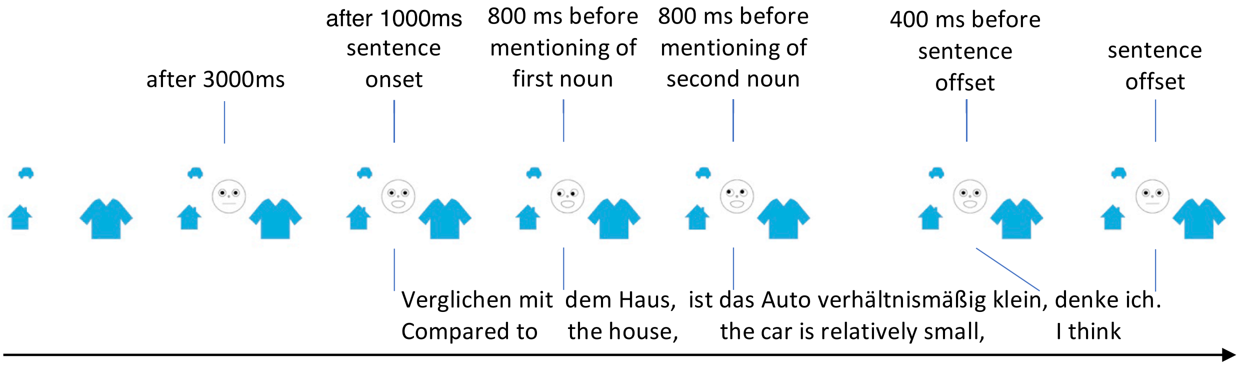

after 3000ms
after 1000ms sentence onset
800 ms before mentioning of first noun
800 ms before mentioning of second noun
400 ms before sentence offset
sentence offset
Verglichen mit dem Haus, ist das Auto verhältnismäßig klein, denke ich.
Compared to the house, the car is relatively small, I think

Fig. 4. Timeline of an item in the Congruent condition in Experiment 2.

more positive deviation from the Congruent ($M = -0.46\ \mu\text{V}$, $SD = 1.46$) baseline ($F(1, 29) = 18.37$, $p < .01$, $\eta_p^2 = 0.39$). There was no difference between the Congruent baseline and the Mutual ($M = -0.29\ \mu\text{V}$, $SD = 1.86$) condition ($F(1, 29) = 0.29$, $p = .59$).

Additionally, we compared the Averted filler condition only with the Congruent baseline in order to allow for a direct comparison between the two experiments. Recall that the Mutual condition stands for a straight gaze toward the listener, whereas Averted stands for a gaze toward the empty position as it was the case in the first experiment. Overall, the Averted data shows similar patterns as the Averted condition in the first experiment, including the visually distinct peaks for the N200 and N400 (see Fig. 5). The comparison showed significant effects in the N200 ($M = -1.73\ \mu\text{V}$, $SD = 1.46$) and N400 ($M = -1.44\ \mu\text{V}$, $SD = 1.64$) time windows but no effect in the P600 ($M = -0.45\ \mu\text{V}$, $SD = 1.96$) time-window (($F(1, 29) = 6.79$, $p < .05$, $\eta_p^2 = 0.19$), ($F(1, 29) = 6.79$, $p < .05$, $\eta_p^2 = 0.19$) and ($F(1, 29) = 0.13$, $p = .71$) respectively). (see Fig. 6).

Table 3 summarizes the findings for the reported time-windows for Experiment 1 on the left side and Experiment 2 on the right side.

3.3. Discussion

The findings in the second experiment replicated findings from the first experiment. This holds especially for the N200 region (150–300 ms) and the P600 region (600–900 ms). For both time-windows the results from the second experiment replicated the findings from the first experiment. However, the results in the N400 time-window (300–450 ms) deviated from the findings in the first experiment for the two ‘uninformative’ conditions indicating differences in the perception of these gaze cues. The difference found for these two conditions provides further support for the claim that the N200 and N400 are indexing different processes.

4. General discussion

Evidence from behavioral data in the literature suggests that speech aligned speaker gaze facilitates comprehension. Our experiments shed light on the underlying mechanisms involved in the integration of gaze into the situation model as reflecting speaker intentions resulting in expectations.

The results from our experiments suggest that the gaze cue preceding the critical second noun is used to predict a word form and is integrated into the situation model as reflecting the speakers referential intentions. This results in clear expectations regarding how the sentence will continue. We identified three ERP components that are involved in these processes, indexing an auditory matching mechanism (N200), word retrieval (N400) and the integration into, as well as the revision of a mental representation of the situation (P600). The latter indicates that gaze toward objects leads to the integration of the gazed-at object into the mental representation, thus going beyond a simple increase of the objects prominence. In the following, we discuss each of these components separately.

4.1. N200

The early negative component between 150 and 300 ms can be plausibly interpreted as a Phonological Mismatch/Matching Negativity (PMN) as described by, e.g., Connolly and Phillips (1994). Similar results have been found by Hagoort and Brown (2000). They explain this early effect peaking at around 250 ms as a mismatch between the expected word form given a context and the actual activated word candidates given the speech signal listeners perceive and additionally suggest that the ‘effect might reflect the lexical selection process that occurs at the interface of lexical form and contextual meaning’ (p. 1528).

Importantly, while the above studies established the context based on linguistic information alone, in our study, the expectations were established by speaker gaze toward an object present in the visual scene. The linguistic context alone supports no preference for either of the valid nouns/referents.

In our experiment, the objects appearing together all had names that began with different phonemes,[5] which is an important factor to elicit a PMN (D’Arcy, Connolly, Service, Hawco, & Houlihan, 2004; Connolly & Phillips, 1994; Hagoort & Brown, 2000). If the initial phoneme of the input matches the onset of the predicted word (i.e. named the gazed at object), this phoneme provides little new information and is therefore easily processed (Congruent). If however the phoneme provides more information, either contradicting the prediction of a specific noun (Incongruent), or by helping to reduce the set of possible nouns to a single target (Averted/Mutual), an N200 modulation is elicited. The Averted and Mutual gaze cues do not provide any further information about the upcoming word. Therefore, lacking an early visual point of disambiguation (VPoD), the earliest possibility to identify and select the actual target is provided by the first phoneme of the actual noun as the linguistic point of disambiguation (LPoD). This in turn increases the information load conveyed by this phoneme. In sum, we interpret the N200 to reflect the processing of information provided by the phoneme, given gaze-driven word-form expectations.

4.2. N400

The N400 effect has been reported to reflect retrieval effort based on

[5] With the exception of the pair ‘Stiefel’ - ‘Stern’ in experiment 1, which was present in three experimental trials. In two of those three cases one of the two objects was the medium sized object, making it the first gazed at object in those sentences independent of the condition. Thereby, they are no longer a valid target for the second gaze cue or naming as the second noun in the sentence in any condition. ‘Stern’ was replaced with ‘Mond’ in Experiment 2.

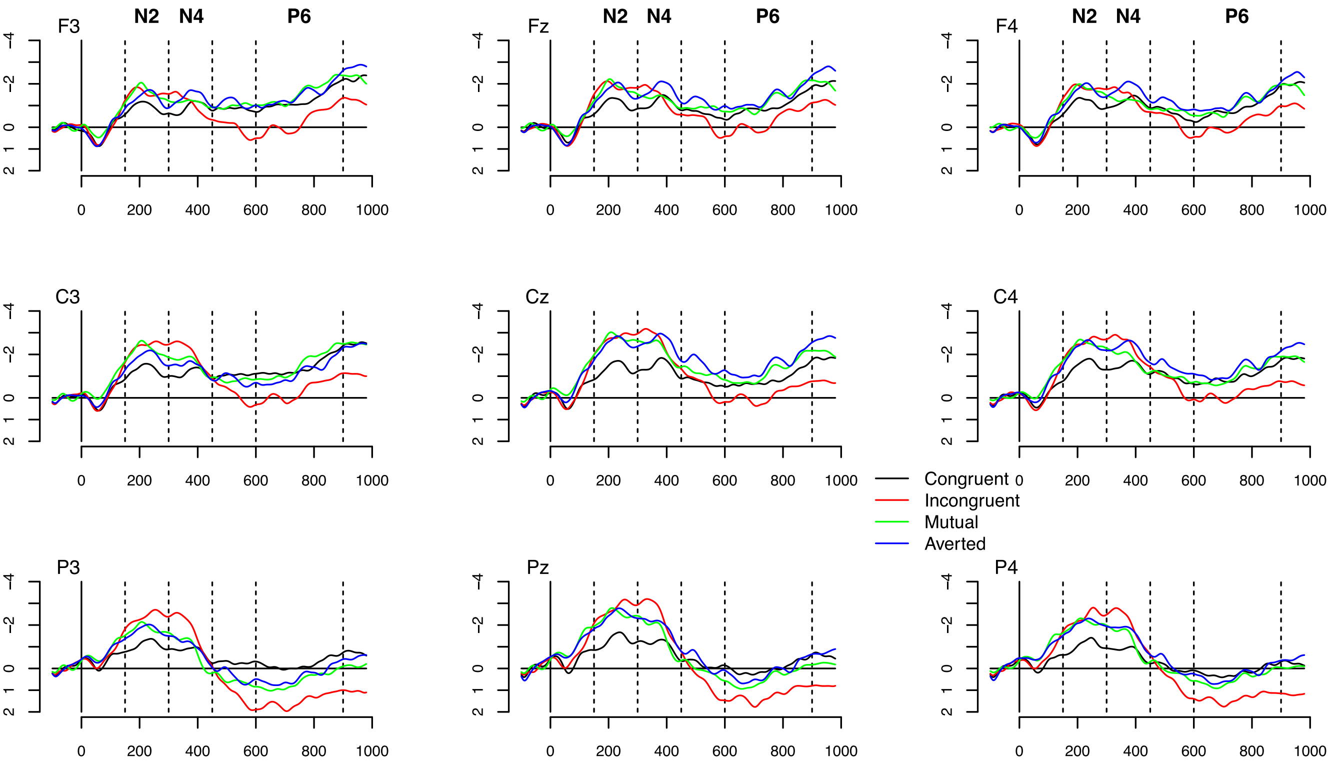


Fig. 5. ERP time-locked to the second noun onset in Experiment 2 separated by the experimental conditions (Congruent (black), Incongruent (red), Mutual (green), and Averted (blue). The data presented shows the electrode subset F3, Fz, F4, C3, Cz, C4, P3, Pz and P4 filtered at 20 Hz for presentation purposes only. Negativity is plotted upwards. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

expectations arising from contextual information (Kutas & Federmeier, 2011; Schumacher, 2012; Van Berkum, 2009). In the current study, the gaze cue preceding the second noun leads to expectations for the upcoming noun. In the Congruent condition, those (matched) expectations lead to facilitated retrieval of the noun, as revealed by an attenuated N400 amplitude. In the Averted and Mutual conditions,

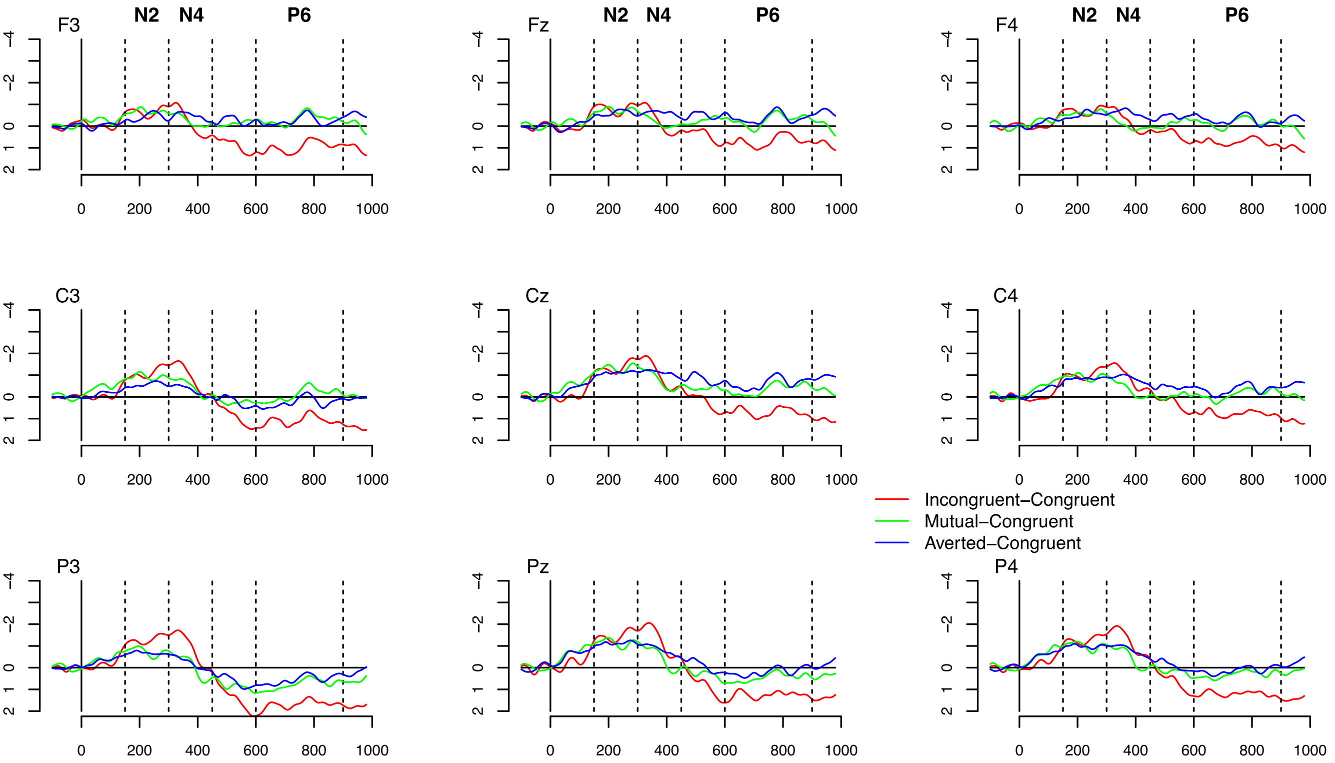


Fig. 6. Difference waves of Incongruent-minus-Congruent (red), Mutual-minus-Congruent (green) and Averted-minus-Congruent (blue). The data presented shows the electrode subset F3, Fz, F4, C3, Cz, C4, P3, Pz and P4 filtered at 20 Hz for presentation purposes only. Negativity is plotted upwards. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

participants have two possible upcoming nouns activated. The noun alone is used to identify the referent, resulting in a significantly larger N400 effect compared to the Congruent condition. In the Incongruent condition, the noun is not consistent with the expectations formed using the gaze cue. This increases the retrieval cost of the noun, as manifest by the significantly larger N400 effect, compared to the Congruent condition. The difference in the effect size of the N400 between the Averted (medium sized effect) and Incongruent (large sized effect) condition[6] in Experiment 1 could also explain the morphological differences observed in the ERPs in this time region. The Incongruent condition displays a single negative movement in central and parietal regions compared to the two peaks observed in the Averted condition. However, in frontal electrodes, the two peaks are visually distinguishable in both conditions. We interpret this as a stronger N400 effect in the Incongruent condition that interacts/overlaps with the N200.

The findings in the N400 time-window in the second experiment replicate the results from the first experiment for the three conditions that were also present in the first experiment (Congruent, Incongruent and Averted), as summarized in Table 3. However, the added Mutual condition utilizing a straight gaze back to the participant instead of being directed toward the empty position (Averted) shows no significant difference from the Congruent baseline condition unlike its Averted counterpart.

The significant difference in the N200 time-window followed by a lack of difference in the N400 time-window only for the Mutual condition provides further evidence that the two peaks (N200 and N400) are indeed separable and expressing two distinct processes. In both the Mutual and Averted conditions, the auditory input can be utilized to select the target from a set of expected objects early on. However, the reduced N400 effect following the listener-directed Mutual gaze compared to the Averted gaze suggests that the two different gaze cues might introduce qualitatively different expectations. Although the precise nature of the difference between the two ‘uninformative’ cues in relation to the Congruent baseline remains to be investigated, we propose two possible explanations. Firstly, it could be argued that every position-oriented gaze action is interpreted as being meaningful throughout the experiment based on the higher number of object-oriented gaze actions (75%). As a consequence, even gaze to an empty position could bind the listeners’ attention or pull their attention away from the objects provided in the scene, hence, hindering word retrieval for any object outside of the attentional focus.

Alternatively, it may be the case that Mutual gaze expresses a higher amount of certainty about the upcoming word compared to an Averted gaze cue toward an empty position, which might rather imply some degree of uncertainty and might even open the space of suitable candidates beyond those objects present on the screen. The speaker’s gaze away from the interlocutor or away from discourse relevant objects is often described as a disengagement from the environment and used to facilitate remembering or, more generally, to lower the cognitive load of the speaker (Doherty-Sneddon, Bruce, Bonner, Longbotham, & Doyle, 2002; Glenberg, Schroeder, & Robertson, 1998). Such a gaze behavior is often understood to display uncertainty or disfluency (Griffin, 2004). Work from Swerts and Krahmer (2005) has shown that interlocutors pick up on such cues that display uncertainty and interpret these in relation to the so called Feeling of Another’s Knowing (FOAK). In their study they showed that participants presented with videos of speakers displaying such cues of uncertainty rated those sentences with a lower FOAK score than videos in which those cues were not displayed. If the averted gaze cue in our experiment is interpreted along these lines, it is possible that this Averted condition leads to different expectations or predictions than the Mutual condition utilizing a straight gaze toward the participant. It is possible that word retrieval for a small set of expectable objects (Averted and Mutual conditions) is benefiting from a higher FOAK (Mutual).

4.3. P600

The update of the mental situation model following the violation of the comprehenders expectations can elicit a P600 effect (Burkhardt, 2006, 2007; Van Berkum et al., 2007). Following those accounts, we interpret our findings in the P600 region as revision/integration costs of the situation model. In both the Congruent and Incongruent gaze condition participants can exploit the gaze cue toward an object to integrate the identified referent into their situation model, and establish expectations for the upcoming noun. In both the Congruent and Averted/Mutual condition, there is no violation of expectations: Either the expected referent was named (Congruent), or no expectations have been formed (Averted/Mutual). In the Incongruent condition however, the violation of the expectations leads to the need to revise the situational model, by replacing the expected referent in the model with the actually named referent that was discarded based on the visual information.

4.4. Summary

We have reported evidence for the utilization of speaker’s gaze by interlocutors to form expectations of the unfolding sentence.

Our findings suggest that gaze is used to form expectations about the upcoming referent, resulting in increased retrieval costs when gaze is uninformative or misleading, as indicated by a stronger N400 modulation in these cases. The attenuated N200 for Congruent gaze preceding the N400 time-window further suggests that predictions are not only formed on a conceptual level but also about the concrete lexical form when a single object is highlighted. The additional findings in the second experiment regarding the Mutual gaze cue as displayed by a straight gaze to the participant provides further evidence to distinguish between the two processes. It is important to note, however, that our results also suggest a strong interplay between these two components. The relatively short lived N400 (300–450 ms) following the N200 indicates that the retrieval of the full word benefits from the phonological matching as indexed by the N200. We speculate that the presence or absence of the N200 might have an effect on the strength of the N400. This, however, requires further investigation.

In the P600 time-window, our results suggest that the visual scene and speech signal as well as gaze are used to form a mental representation of the discourse. This is consistent with the view that gaze is interpreted as conveying referential intentions (Staudte & Crocker, 2011). The first gaze action in each experimental trial correctly provided evidence about the upcoming referent. In case of a following Incongruent gaze, participants were led to believe that the gazed at object actually would be the upcoming noun, eliminating the remaining objects in the scene as likely referents. The upcoming noun however forces the participant to reintegrate the formerly dismissed object into the mental representation. This in turn is then reflected by a P600 modulation representing the (re-) integration difficulty. No such difference is induced in the ‘uninformative’ conditions as either upcoming referent is still possible and, hence, does not require a revision of the situation model. While the N400 for both the Incongruent and Averted conditions was consistent with both the prominence and situated integration accounts, the observation of a P600 only in the Incongruent condition was predicted by the situated integration account alone.

Results from the second experiment replicated the results from the first experiment. When comparing only the three conditions that were present in both experiments, the second experiment shows similar patterns in the N200, N400 and P600 time-windows. This demonstrates the robustness of the observed effects to variation in object position and gaze cue validity. We further used two different types of ‘uninformative’ gaze cues in our experiments, either being directed toward an empty position on the screen or back toward the participant. Our results showed a significant N400 modulation in the noun region for the Averted gaze condition compared to Congruent condition that is absent

[6] Both compared to the Congruent baseline condition.

in the Mutual condition, while both 'uninformative' gaze versions replicate the effects in every other time-window.

It is perhaps worth emphasizing that – despite the use of a stylized speaker, gaze cues, and objects – participants none the less integrate gaze with the speech signal. Therefore, it seems plausible to assume that our findings may be generalized to other time-aligned speaker cues as for example gestures rather than gaze, as well as more natural human faces. Thus, we see these experiment as a solid base for future work investigating the integration of linguistic with non-linguistic aspects of the signal.

Taken together, our findings are consistent with the retrieval-integration account (Brouwer et al., 2017), such that retrieval difficulty (N400) is observed for the Incongruent and Averted conditions, while integration difficulty (P600) is found only when revision of the situation model is necessary in the Incongruent condition.

While EEG does not directly reveal the specific neural substrates that underlie the reported N200, N400 and P600 effects we have reported here, we can offer some speculation based on existing proposals. With regard to the phonological expectations and mismatch, work by e.g., Poeppel and colleagues suggests the involvement of the superior temporal sulcus (STS) in relevant aspects of phonological processing Poeppel (2003, 2014, 2007). Further, as we interpret our results in terms of the Retrieval-Integration account, we follow Brouwer and Hoeks (2013) in suggesting that N400 modulations – which are argued to index lexical retrieval - reflect involvement of the left posterior part of the Middle Temporal Gyrus (lpMTG; Brodmann area 21) (see also Lau, Phillips, & Poeppel, 2008). By contrast, increased P600 amplitude – which is taken to index semantic integration difficulty – is argued to reflect involvement of the left Inferior Frontal Gyrus (lIFG; Brodmann areas 44/45/47) (see evidence reviewed in Hagoort, Baggio, & Willems, 2009). For an alternative account, see Friederici, 2011). While a more detailed understanding of the neural generators of our observed effects constitutes an important and ongoing area of investigation, we suggest that the ERP effects alone offer compelling neurophysiological evidence for rapid integration of speaker gaze and speech across all stages of listener comprehension.

5. Conclusion

We demonstrated a robust and replicable influence of speech-related gaze cues on a range of underlying cognitive processes, including auditory word processing, lexical retrieval, and integration with sentence meaning, as expressed by an N200, N400 and P600 effect respectively. The distinct N200 and N400 components suggest that gaze elicits predictions on word form level which are matched with the incoming phonological information whereas the N400 indicates a broader expectation-driven retrieval mechanism. The P600 results indicate that listeners utilize speakers' gaze above and beyond any increase in prominence, such that the information provided by gaze is used to update the situation model even in advance of hearing the gazed at referent.

Acknowledgments

This work was supported by the Cluster of Excellence "Multimodal Computing and Interaction" and SFB/CRC 1102 "Information Density and Linguistic Encoding", funded by the Deutsche Forschungsgemeinschaft (DFG).

Appendix A. Gaze cue preceding the second noun

In addition to the reported results in the noun region following the manipulated gaze cues, we also analyzed the gaze region itself. As the results found in the noun region show responses to gaze induced expectations, we were interested in possible effects during the formation of these expectations. More precisely, we expected differences between object-directed gaze compared to uninformative gaze.

A.1. Experiment 1

In the analysis of Object-oriented gaze (up, left and right collapsed) compared with an Averted gaze cue (down), a significantly more positive ERP response starting at 200 ms after onset was found for Averted gaze cues ($F(1, 29) = 6.54, p < .05, \eta_p^2 = 0.18$). However, post hoc analyses of the gaze directions indicated that this difference is driven by the position of the object relative to the centrally presented face. We performed an ANOVA using gaze cue direction (vertical/horizontal) as a factor. There was a long lasting globally distributed positivity for vertical gaze cues compared to horizontal gaze cues in the same time-window as in the aforementioned comparison ($F(1, 29) = 16.03, p < .05, \eta_p^2 = 0.36$).

However, there was no significant difference between gaze cue direction when comparing only upward and downward gaze cues ($F(1, 29) = 0.02, p > .05$), nor for a comparison of only leftward and rightward gaze cues ($F(1, 29) = 0.48, p > .05$). The lack of a difference between upward and downward cues possibly implies that the meaningfulness of a gaze cue, in this case the gaze toward a possibly named object (up) versus an Averted gaze (down), did not seem to influence listeners' integration of such cues.

A.2. Experiment 2

The changed positioning of the objects with a fully counterbalanced rotation of the empty position allows for more interpretable analysis of the gaze region. We analyzed the difference between gaze cues that were directed toward an object (Congruent and Incongruent condition) compared to uninformative gaze cues (Averted filler and Mutual condition) by running an ANOVA in the time-window between 300 and 500 ms. There was a significantly larger negativity for both uninformative gaze cues compared to gaze cues toward an object ($F(1, 29) = 5.73, p < .05, \eta_p^2 = 0.17$).

A.3. Results and Discussion

In Experiment 1, we could not reliably interpret the results found in the gaze region preceding the second noun as the positioning of the objects in positions besides (horizontal) and above the face (vertical), confounded the results. This led to a direct comparison only being possible for object-directed gaze cues upward and Averted gaze cues downward. In Experiment 2, however, we addressed this confound by rearranging the objects diagonally to the face, in order to gain better insights into effects in this region. We interpret the reported differences in the N400 time-window as a violation of expectations for an object-oriented, informative gaze cue rather than an uninformative gaze cue, in line with other expectation related reports of the N400 (e.g., Ganis, Kutas, & Sereno, 1996). The lack of a difference in the first experiment could be caused by the aforementioned placement of objects relative to the face.

The N400-like negativity found in the gaze region for uninformative gaze cues compared to gaze cues toward objects possibly hint toward a form of expectation violation. Throughout the course of the experiment, participants were much more often exposed to gaze cues that were directed toward an object. If all gaze actions were taken into account, three out of four gaze cues were directed toward an object whereas only one fourth of the gaze cues were not directed toward an object. Therefore, it could be argued that participants have a higher expectation for an informative, object-directed gaze cue and, hence, a violation of this expectation may elicit an N400 modulation.

References

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the n400 and the p600 in language processing. *Cognitive Science, 41*, 1318–1352.

Brouwer, H., & Hoeks, J. C. (2013). A time and place for language comprehension: Mapping the n400 and the p600 to a minimal cortical network. *Frontiers in Human Neuroscience, 7*, 758.

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language, 98*, 159–168.

Burkhardt, P. (2007). The p600 reflects cost of new information in discourse memory. *Neuroreport, 18*, 1851–1854.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Routledge.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6*, 256–266.

Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language, 43*, 1–18.

Connolly, J. F., Stewart, S., & Phillips, N. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language, 39*, 302–318.

Daltrozzo, J., & Schön, D. (2009). Conceptual processing in music as revealed by n400 effects on words and musical targets. *Journal of Cognitive Neuroscience, 21*, 1882–1892.

D'Arcy, R. C., Connolly, J. F., Service, E., Hawco, C. S., & Houlihan, M. E. (2004). Separating phonological and semantic processing in auditory sentence processing: A high-resolution event-related brain potential study. *Human Brain Mapping, 22*, 40–51.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology, 38*, 438.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*, 509–540.

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews, 24*, 581–604.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*, 469–495.

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

Flom, R. E., Lee, K. E., & Muir, D. E. (2007). *Gaze-following: Its development and significance.* Lawrence: Erlbaum Associates Publishers.

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

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews, 91*, 1357–1392.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for common sense? An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience, 8*, 89–106.

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition, 26*, 651–658.

Griffin, Z. M. (2004). The integration of language, vision, and action: Eye movements and the visual world. In J. M. Henderson, & F. Ferreira (Eds.). *The integration of language, vision, and action: Eye movements and the visual world chapter Why look? Reasons for eye movements related to language production.* New York: Psychology Press.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 4*, 274–279.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. *The cognitive neurosciences* (pp. 819–836). (4th ed.). MIT Press.

Hagoort, P., & Brown, C. M. (2000). Erp effects of listening to speech: Semantic erp effects. *Neuropsychologia, 38*, 1518–1530.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*, 438–441.

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

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393.

Kappenman, E. S., & Luck, S. J. (2016). Best practices for event-related potential research in clinical populations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*, 110–115.

Kiefer, M. (2002). The n400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of n400 priming effects. *Cognitive Brain Research, 13*, 27–39.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience, 7*, 302.

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language, 54*, 445–465.

Kreysa, H. (2009). *Coordinating speech-related eye movements between comprehension and production.* The University of Edinburgh.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology, 62*, 621–647.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the n400. *Nature Reviews Neuroscience, 9*, 920.

Luck, S. J. (2014). *An introduction to the event-related potential technique.* MIT Press.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*, 25–33.

O'Halloran, J. P., Isenhart, R., Sandman, C. A., & Larkey, L. S. (1988). Brain responses to semantic anomally in natural, continuous speech. *International Journal of Psychophysiology, 6*, 243–254.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*, 245–255.

Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Current Opinion in Neurobiology, 28*, 142–149.

Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory n400 event-related brain potential. *Cognitive Brain Research, 1*, 73–86.

R Core Team (2015). *R: A language and environment for statistical computing.* Austria: R Foundation for Statistical Computing Vienna.

Ricciardelli, P., Carcagno, S., Vallar, G., & Bricolo, E. (2013). Is gaze following purely reflexive or goal-directed instead? Revisiting the automaticity of orienting attention by gaze cues. *Experimental Brain Research, 224*, 93–106.

Schumacher, P. B. (2012). Context in neurolinguistics. *What is a Context?: Linguistic Approaches and Challenges, 196*, 33.

Senju, A., Tojo, Y., Dairoku, H., & Hasegawa, T. (2004). Reflexive orienting in response to eye gaze and an arrow in children with and without autism. *Journal of Child Psychology and Psychiatry, 45*, 445–458.

Sharbrough, F., Chatrian, G. E., Lesser, R. P., Lüders, H., Nuwer, M., & Picton, T. W. (1991). American electroencephalographic society guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology, 8*, 200–202.

Spivey, M., Joanisse, M., & McRae, K. (2012). *The Cambridge handbook of psycholinguistics.* Cambridge University Press.

Staudte, M., & Crocker, M. W. (2010). When robot gaze helps human listeners: Attentional versus intentional account. *Proceedings of the 32nd annual meeting of the cognitive science society* (pp. 1637–1642). .

Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition, 120*, 268–291.

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

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53*, 81–94.

Van Berkum, J. J. (2009). The neuropragmatics of 'simple' utterance comprehension: An erp review. *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

Van Berkum, J. J., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research, 1146*, 158–171.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162.