The Influence of Speaker's Gaze on Sentence Comprehension: An ERP Investigation

Torsten Kai Jachmann (jachmann@coli.uni-saarland.de)

Heiner Drenhaus (drenhaus@coli.uni-saarland.de)

Maria Staudte (masta@coli.uni-saarland.de)

Matthew W. Crocker (crocker@coli.uni-saarland.de)
Department of Language Science and Technology, Campus Saarbrücken, Germany
Cluster of Excellence MMCI, Saarland University, Germany

Abstract

Behavioral studies demonstrate the influence of speaker gaze in visually-situated spoken language comprehension. We present an ERP experiment examining the influence of speaker's gaze congruency on listeners' comprehension of referential expressions related to a shared visual scene. We demonstrate that listeners exploit speakers' gaze toward objects in order to form sentence continuation expectations: Compared to a congruent gaze condition, we observe an increased N400 when (a) the lack of gaze (neutral) does not allow for upcoming noun prediction, and (b) when the noun violates gaze-driven expectations (incongruent). The later also results in a late (sustained) positivity, reflecting the need to update the assumed situation model. We take the combination of the N400 and late positivity as evidence that speaker gaze influences both lexical retrieval and integration processes, respectively (Brouwer et al., in press). Moreover, speaker gaze is interpreted as reflecting referential intentions (Staudte & Crocker, 2011).

Keywords: ERP; N2; N400; late sustained positivity; gaze; prediction; referential expressions

Introduction

The gaze of a speaker toward objects present in a shared scene in a face-to-face interaction 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). This visual cue can be used by the listener to ground and disambiguate referring expressions, infer the speaker's intentions and goals, and thus facilitate comprehension (Hanna & Brennan, 2007). As most research conducted on the influence of speakers' gaze on listeners' language comprehension focused on behavioral data (e.g.: reaction times, eye movements), little is known about the precise time course for the integration of visual and linguistic information, or which underlying mechanisms are involved.

We therefore conducted an ERP-study to investigate how listeners integrate cues provided by the speaker's gaze when it is time-aligned to an utterance containing statements about the visual context. We monitored listeners' event-related potentials (ERPs) as they observed a stylized face performing gaze actions toward simple objects preceding their mentioning in a simultaneously presented utterance that compared objects in the scene with one-another. The gaze cue corresponding to the second object in the sentence was either Congruent (toward the named object), Incongruent (toward the object that remains unnamed in the sentence) or Neutral (toward an empty position at the bottom of the screen). This manipulation was intended to shed light on how listeners use speakers' gaze to anticipate and integrate subsequently mentioned referents.

Previous eye-tracking studies have shown that, when a visual context is present, speakers orient their gaze toward an object about 800 - 1000ms before mentioning it (Griffin & Bock, 2000; Kreysa, 2009). However, it is less clear to what extent such speaker gaze affects listeners' sentence comprehension.

Staudte and Crocker (2011) showed in an eye-tracking study that participants used gaze cues to disambiguate a sentence with multiple same-type referents as soon as it was provided, expressed by a higher inspection rate to the gazed-at object compared to the competitor. Furthermore, a misleading gaze cue lead to a longer reaction time while judging whether the sentence was true or false given the visual scene.

For our ERP study, we hypothesized that listeners integrate gaze cues in a situation model to anticipate which objects are likely to be subsequently mentioned, influencing the retrieval and integration of the noun. Specifically, we expect an N400 modulation will occur as a function of predictability, such that neutral gaze, and even more so incongruent gaze, will increase the amplitude of the N400 compared to a congruent gaze condition. Additionally, we hypothesized that the naming of objects that were previously eliminated based on (incongruent) speaker gaze should lead to a higher cost of integration, possibly reflected by a increased late positivity (Brouwer, Crocker, Venhuizen, & Hoeks, in press).

Experiment

In our experiment, German native speakers judged the truthfulness of an auditorily presented sentence 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. Every gaze action was performed 800ms prior to the naming of the corresponding noun. The first gaze was always Congruent toward the object that was named first in the sentence. Gaze to the second object was manipulated such that is was either toward the second named object (Congruent), toward a distractor object that was never named during the course of the trial (Incongruent) or toward the bottom of the screen where no object was situated (Neutral).

We hypothesized that (1) Congruent gaze toward the upcoming object leads to facilitated retrieval of the corresponding noun, and a reduced N400, as it is highly predictable given the visual scene. (2) Incongruent gaze on the other hand is hypothesized to evoke an increased N400 modulation, as the visual information favors predictions of the unnamed object and thereby hinders word retrieval. Additionally, the elimination of the named object based on the visual information demands an update of the situation model and thereby might increases integration costs reflected by a late positivity. (3) As Neutral gaze does not highlight one object more than the other, both remaining objects are equally predictable in the sentence. This might lead to an intermediate retrieval cost of the noun.

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.

Stimulus Materials and Procedure

We created 24 pictures of objects of masculine, feminine and neuter gender (8 per gender). The pictures were pretested to ensure that they (a) were recognized as the intended objects and (b) were equally complex in their appearance.

Participants were presented with a picture containing three objects of the same gender that varied either in size or brightness arranged in positions above, left and right of the center of the screen. Each screen contained a large, medium, and small object (or bright, medium, and dark object respectively). After 3000ms, 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 systems 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 listen 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.

In order to keep the influence of the first noun on the second gaze cue as well as on the second noun the same across all items, a pause of variable length was introduced after the first noun, so that the distance of the onset of the first noun to the onset of the second half of the sentence always was about 1000ms. At sentence 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 first gaze cue appeared approximately 800ms before the first noun was mentioned. This gaze cue was always Congruent toward the 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 object (or object of medium brightness when brightness was manipulated). If the first mentioned object were the smallest/brightest or biggest/darkest object in the scene, it would not matter which of the other objects were named second, as for both the same comparative adjective would render the sentence true or false.

An example of the visual scene provided in Figure 1 displays the time line of an example trial, with a small house, medium car and a large t-shirt. If the t-shirt was mentioned first in this context, both of the remaining objects would be smaller. The second, manipulated gaze cue then appeared again 800ms prior to the onset of the second noun. The gaze was redirected toward the participant 400ms before the end of the sentence, and the mouth closed on the offset of the sentence.

Each item appeared in three conditions (Congruent / Incongruent / Neutral). In the Congruent condition, the gaze preceding the second noun was directed toward the subsequently named object. In the Incongruent condition, the gaze cue went toward the object that remained unnamed in the sentence. In the Neutral condition, gaze was directed toward the bottom of the screen where no object was present, in order to still present a gaze cue induced by the eye-movement of the face. Additionally, we created versions of those manipulations that were counterbalanced for naturalness. Naturalness was defined as the truth value of the utterance in reality. For example, the in-reality invalid utterance "compared to the car, the house is relatively small, I think" was counterbalanced with the utterance "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. Using a Latin-square design, this led to a total of six lists.

Each list contained 72 experimental items (24 per condition) and 72 fillers with mentioning of an object other then the medium object as the first noun and gaze patterns different from the gaze patterns in the experimental items. As in the experimental items only the second gaze cue was manipulated, 25% of the fillers (18) contained a manipulation of the first gaze cue instead of the second gaze cue. This version of the fillers still started with a mentioning of the medium object as the first noun in the sentence. The first gaze cue was always neutral and never incongruent in order to enforce the validity of the gaze cues. The remaining fillers were of the same form as the experimental items with the difference

[1] For a concrete description of the removal see Section "Data Analysis" on the following page.

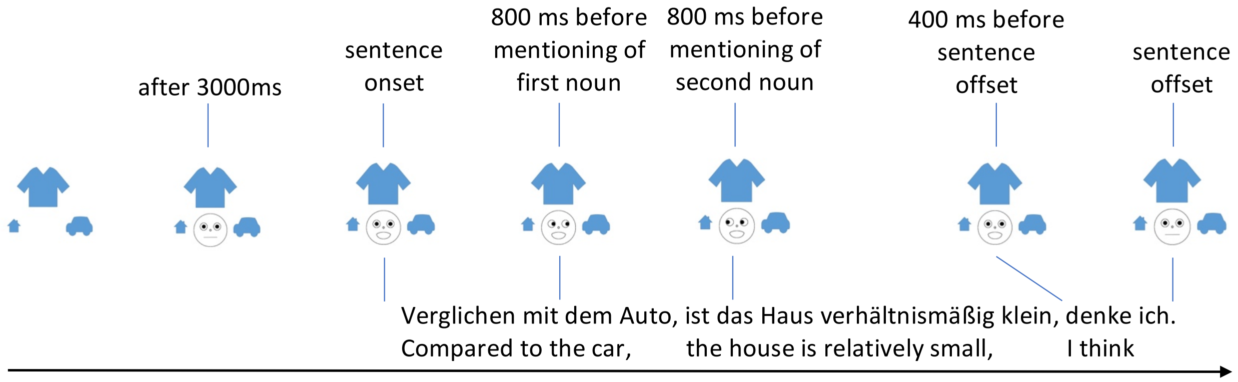
after 3000ms
sentence onset
800 ms before mentioning of first noun
800 ms before mentioning of second noun
400 ms before sentence offset
sentence offset
Verglichen mit dem Auto, ist das Haus verhältnismäßig klein, denke ich.
Compared to the car, the house is relatively small, I think

Figure 1: Timeline of an item in the Congruent condition.

that the first mentioned object was either the small or large (bright/dark) object. The second named object in these fillers was the medium object half of the time and the remaining size/brightness the other half. The gaze patterns performed on these fillers always started with a congruent gaze, as in the experimental items, followed by another congruent gaze towards the second named object half of the times (36) and a quarter of the times each by an incongruent or neutral gaze cue (18). This distribution of gaze patterns throughout the experiment led to an overall ratio of congruent gaze actions[2] of 70% (204). Another 18% (51) of the gaze actions were Neutral and only about 12% (33) of the gaze actions were Incongruent. This way, the validity of the gaze cue was strongly enforced 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 1280x1024 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 breaks 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. Answers were recorded using a Response Pad RB-834 (Cedrus Corporation). The experiment lasted approximately 45 min.

Data Analysis

The EEG was recorded by 24 Ag/AgCl scalp electrodes (actiCAP, BrainProducts) and amplified with a BrainAmp (BrainVision) amplifier. Electrodes were placed according to the 10-20 system (Sharbrough et al., 1995). Impedances were kept below 5kΩ. 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 250Hz was used. The EEG data was band pass filtered offline at 0.01-40Hz 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 1100ms window per condition relative to the acoustical onset of the noun following the manipulated gaze cue and the manipulated gaze cue itself. All segments were aligned to a 100ms pre-stimulus baseline. We semi-automatically screened offline for 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 artefacts 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 removed. On average 5.3 trials per participant and condition (22%) were rejected due to eye movements). This led to the removal of 12 participants from the analysis. Additionally, the data of 3 participants was removed due to their behavioral data. Participants’ data was removed if they gave wrong answers to more than 10% of the questions. Overall, the two criteria led to the removal of the data of 15 participants. The averaged data of the remaining 30 participants (Mean age: 23.7; Age range: [18, 32]; SD: 3.49; Male: 4) was

[2] As every trial contained two gaze actions, one aligned to the first noun and one aligned to the second noun, the total number of gaze actions throughout the course of the experiment was 288 per list/participant.

exported using BrainVision Analyzer (Version 2.1) BESA export function.

We analyzed the ERP data time-locked to the onset of the second 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^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.

Critical Region (Second Noun)

We analyzed the influence of experimental condition (Congruent, Incongruent, Neutral gaze) time locked to the onset of the second noun, including electrode site (frontal/central/parietal) as within-subject factors. An ANOVA of time window between 150 and 400ms showed a main effect of condition ($F(2,58) = 3.99, p < 0.05, \eta^2 = 0.12$). There is a globally distributed, significantly larger negativity for both the Incongruent and Neutral condition compared to the Congruent condition (($F(1,29) = 5.66, p < 0.05, \eta^2 = 0.16$) and ($F(1,29) = 4.86, p < .05, \eta^2 = 0.14$) respectively). However, further Visual inspection revealed that the Neutral condition contains two distinct, frontally distributed peaks within this time window (see Figure 2 for comparison). This is coherent with previous findings by, e.g., Hagoort and Brown (2000). This led to the analysis using a moving time window in this epoch in order to determine whether those peaks are indeed distinct. We split the data of the previous time-window in four overlapping sub-time-windows of 100ms length with a distance of 50ms each and introduced those time-windows as a factor in an ANOVA, where the interaction of time-window, longitude and condition showed a significant effect ($F(12,348) = 1.95, p < 0.05, \eta^2 = 0.06$). We find a main effect of condition only in the time windows between 150 - 300ms ($F(2,58) = 3.93, p < 0.05, \eta^2 = 0.12$) and 300 - 400ms ($F(2,58) = 3.38, p < 0.05, \eta^2 = 0.1$). This indicated that the two peaks are indeed distinct. A pairwise comparison of the conditions in the time window of the earlier peak (150 – 300ms) showed that both the Incongruent and Neutral condition retain their significantly larger negativity ((($F(1,29) = 5.82, p < 0.05, \eta^2 = 0.17$) and ($F(1,29) = 4.93, p < .05, \eta^2 = 0.15$) respectively). A pairwise comparison in the time window of the second peak also shows that both the Incongruent and Neutral condition contain a significantly larger negativity (($F(1,29) = 4.22, p < 0.05, \eta^2 = 0.13$) and ($F(1,29) = 5.99, p < .05, \eta^2 = 0.17$) respectively). Additional to the findings in the early time-window, we analyzed the time-window between 600 - 1000ms. The analysis revealed a main effect of experimental condition ($F(2,58) = 3.38, p < 0.05, \eta^2 = 0.16$). A pairwise analysis of the conditions showed that the late long-lasting positivity is only present in the Incongruent condition ($F(1,29) = 7.24, p < 0.05, \eta^2 = 0.2$).

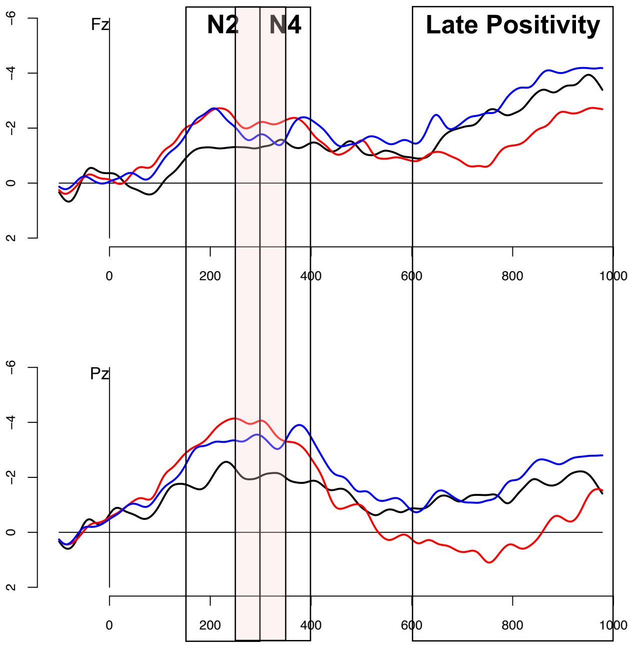

Figure 2: ERP time-locked to the Second Noun Onset separated by the Experimental Conditions (Congruent (black), Incongruent (red) and Neutral (blue)). Reported regions are highlighted by boxes. The data presented shows the electrode subset Fz and Pz filtered at 20Hz (low-pass) for presentation purposes only.

Table 1: Summary of the pair-wise computed differences between the (C)ongruent, (I)ncongruent and (N)eutral conditions split by the analyzed epochs. Significance is indicated by *

Time Window	C - I	C - N	N - I
150 - 400ms	*	*	-
150 - 300ms	*	*	-
250 - 350ms		–	
300 - 400ms	*	*	-
600 - 1000ms	*	-	*

Discussion

Research from Van Berkum, Koornneef, Otten, and Nieuwland (2007) suggests that comprehenders predict the upcoming course of a sentence based on the previously gathered information that they integrated in a situation model (Zwaan & Radvansky, 1998). Various studies have further shown that not only linguistic information is used to form predictions about upcoming sentence content but also visual information (Staudte & Crocker, 2011; Ferreira, Foucart, & Engelhardt, 2013). It therefore seems reasonable to suggest that such visual cues contribute to the construction of the situation model. We interpret the two main components (N400, late positivity) found in terms of the retrieval-integration approach (Brouwer et al., in press): The N400 is modulated by the retrieval difficulty of the upcoming noun, influenced by its predictability given the visual context. Further, the positivity is influenced by the integration difficulties founded in the need to update the situation model.

N400

Parviz, Johnson, Johnson, and Brock (2011) showed that the N400m, representing the N400 as measured by magnetoencephalography (MEG), is modulated by the information content that is conveyed by a word in a given context. A similar interpretation can be attributed to findings from Willems, Frank, Nijhof, Hagoort, and Van den Bosch (2015). In their study participants listened to spoken narratives. Their results show that words with a higher surprisal let to an increased activation in the left temporal lobe, which has been identified as source of the N400 effect (e.g., Van Petten and Luka (2006)). Additionally, an ERP study from Frank, Otten, Galli, and Vigliocco (2015) revealed a correlation between the amplitude of the N400 and word surprisal. In the current study, the gaze cue preceding the second noun leads to predictions for the upcoming noun. In the Congruent condition, those predictions are fulfilled, which leads to an effortless retrieval of the noun and thus leads to a reduced N400 amplitude compared to both Neutral and Incongruent conditions. In the Neutral condition, participants have two possible upcoming nouns active. The information conveyed by the noun therefore is higher than in the Congruent condition, as the set of candidates is reduced to the actual target. In the Incongruent condition, the information conveyed by the noun contradicts the prediction made using the visual information. This results in an increase of the retrieval cost of the noun and thus to an increased N400 effect compared to the congruent condition.

The early onset of the negativity (150ms after noun onset) suggests we may in fact be observing modulation of the N2 component, as well as N400. This is supported by the analysis using a moving time window, which revealed two distinct peaks between 150-300ms and 300-400ms respectively, especially prominent in the Neutral condition (see Figure 2 for comparison), and is discussed below.

N2

The globally distributed, early negative component between 150 and 300ms can be interpreted as a reminiscent of the Phonological Matching Negativity (PMN) as described by Connolly and Phillips (1994). Similar results have been found by Hagoort and Brown (2000). They explain this early effect with a peak around 250ms as a mismatch between the expected word form given a context and the actual activated word candidates given the speech signal listeners perceive. In this study, the context was built up by the gaze towards an object present in the visual scene. The following word now either confirms the expectation (Congruent), which in turn leads to no PMN modulation, or disconfirms them (Incongruent), which evokes a large PMN modulation.

Following the account of Hagoort and Brown (2000), which states that 'the N250 effect might reflect the lexical selection process that occurs at the interface of lexical form and contextual meaning', the effect in our Neutral condition could also be explained as such an selection process. Given our visual scene, at the second noun, two of the three objects are still valid targets. The Neutral gaze cue, directed downwards, does not provide any further information about the upcoming word. As both remaining objects are equally plausible, a decision for either one has to be made using the first phoneme of the uttered word, which leads to the discard of one of the two predictions. This selection process elicits the negativity in the N2 region found in the Neutral condition. It is important to highlight that all of the previously named studies establish the predictive context using language. Our study differs in that predictive context is determined solely on the basis of visual information: the linguistic context does not contribute a preference for either of the valid nouns.

Positivity (600 - 1000ms)

The relation between updating of a situation model and the occurrence of a late positivity has been demonstrated in various studies (Burkhardt, 2007; Donchin, 1981). Following those accounts, we can interpret our findings in the later time window starting at 600ms as similarly reflecting the cost of updating the situation model, and integration more generally (Brouwer et al., in press). In both the Congruent gaze and Incongruent gaze condition participants can exploit the gaze cue towards an object in their situation model to make predictions about the upcoming noun. In the Congruent condition, this leads to no violation of those predictions and thereby doesn't require an update of the situation model.

In the Incongruent condition however, the violation of the predictions leads to the need to update the situation model: the gaze cue toward an object leads to the listener's interpretation of the gazed-at object to be the upcoming noun. This in turn leads to the elimination of the remaining object as relevant to the situation model. As the upcoming noun however shows that the previously discarded object is in fact relevant, the situation model has to be updated, which leads to higher integration costs expressed by the late positivity.

The Neutral condition does not draw the focus to one single object but leaves two objects (the so far unnamed objects) as equally possible targets. The prediction of this set of objects is not violated and therefore does not require an update of the situation model.

Conclusion

We suggest that the N400 and late positivity are most naturally interpreted in terms of the retrieval-integration approach (Brouwer et al., in press): The N400 findings suggest that gaze is used to anticipate the upcoming noun, resulting in increased retrieval cost when gaze is absent or incongruent. Interestingly, the late positivity for incongruent gaze, suggests that gaze is interpreted as conveying referential intentions, resulting in integration difficulty only when gaze is misleading. This is consistent with eye-tracking data from Staudte and Crocker (2011). Additionally, we found an N2 modulation preceding the N400, suggesting gaze leads listeners to anticipate specific word forms. More specifically, we argue that the gaze cue preceding the second noun is used in combination with the unfolding situation model to make predictions about the continuation of the sentence. Those predictions are then matched with the auditory input. If the initial phoneme of the input is in line with the prediction, this phoneme provides little new information and therefore facilitates word retrieval. If however the phoneme provides more information, either by helping to reduce the set of predictions to a single target (Neutral condition) or through violation of the predictions (Incongruent condition), an N2 modulation is elicited. In both cases, a subsequent N400 modulation is evoked. If the predictions are completely violated (Incongruent condition), the situation model needs to be updated, which increases the integration cost of the corresponding noun, expressed by a late positivity. Given the findings in the N4 time-window and the late positivity, a classical semantic integration (N4) and reanalysis (P6) account seems unlikely. The integration of the word in the Neutral condition should not lead to a strong N4 modulation as both possible words fit the context without a semantic violation. This predicted modulation for only the Incongruent condition however can be found in the later time-window reflected in the late positivity. Additionally, the late positivity should not be evoked, according to the classic account, as no syntactic reanalysis is needed in any condition. In sum, our findings demonstrate a robust influence of non-verbal gaze cues on several underlying processes, including auditory processing, lexical retrieval, and integration with sentence meaning.

Acknowledgments

This work was funded by the Cluster of Excellence ”Multimodal Computing and Interaction” of the German Excellence Initiative.

References

Brouwer, H., Crocker, M. W., Venhuizen, N., & Hoeks, J. C. J. (in press). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*.

Burkhardt, P. (2007). The p600 reflects cost of new information in discourse memory. *Neuroreport*, *18*(17), 1851–1854.

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*(3), 256–266.

Donchin, E. (1981). Surprise!... surprise? *Psychophysiology*, *18*(5), 493–513.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, *24*(6), 581–604.

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.

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

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). Brain & Language The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological science*, *4*(11), 274–279.

Hagoort, P., & Brown, C. M. (2000). Erp effects of listening to speech: Semantic erp effects. *Neuropsychologia*, *38*(11), 1518–1530.

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*(4), 596–615.

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

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

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the australasian language technology association workshop 2011* (pp. 38–46).

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, *120*(2), 268–291.

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.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and language*, *97*(3), 279–293.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, bhv075. doi: 10.1093/cercor/bhv075

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.