UNIVERSITÄT OSNABRÜCK

Effects of auditory priming on the perception of semantic and syntactic inconsistencies in a virtual environment

Group 8: Luna Döring, Eva-Marie von Butler, Johannes Weisen & Nele Werner

Designing and conducting an EEG study in VR, Department of Cognitive Science,
University of Osnabrück
25 September 2022

Abstract:

When we perceive and process our environment, we use the smallest cues to quickly guess what is coming next. However, when the cues contrast with what is really happening, our brain reacts with surprise and we are slower to respond. ERP studies show that for semantic as well as syntactic mismatches the brain potentials N400 and P300 arise as a consequence. In the study presented here, we used VR to analyse the influence of auditory priming on visual stimuli and describe how the results would theoretically show up in EEG based on the reaction time and expectedness of the stimuli. Participants were immersed in a virtual environment, where different objects occured after they got primed by an auditory stimuli, which described the objects. To analyse the effect of auditory priming, the stimuli were either semantically and/or syntactically congruent or incongruent. In line with our hypotheses, the results reveal that auditory priming has an influence on the perception and processing of the stimulus object. Incongruent trials were processed more slowly than congruent trials, which is also indicated in the correlation between the expectedness rating and the response time. Based on our results and prior studies, we assume that the stimuli in our virtual environment would elicit a N400 and P300 component.

Keywords: VR, auditory priming, semantic and syntactic inconsistency, EEG, N400

---

1. Introduction

1.1 Background: Electrophysiological Indicants of Semantic and Syntactic Processing

When we walk through the streets, there are countless environmental and sensory impressions that our brain has to perceive and process. A so-called top-down processing helps us to understand our environment quickly and not to be overwhelmed by all the information (von Stein, Chiang & König, 2000). This process categories environmental impressions and is strongly influenced by our expectations and prior knowledge. We use the smallest cues to anticipate quickly what comes next.
But what happens if our anticipation is not met?

In a series of influential studies, Kutas and Hillyard (Kutas & Hillyard, 1981) investigated ERP components associated with the occurrence of unexpected words in written language. At the time it was believed that unexpected stimuli would elicit the P300 component, however when analysing brain activity in response to semantically inappropriate words at the end of otherwise meaningful sentences (e.g. 'dog' at the end of the sentence 'I take coffee with cream and'), the data revealed a negative component that peaked 400ms after stimulus onset (N400). The findings demonstrate, that the P300 is only elicited by stimuli that are physically deviant from our expectation (i.e. She put on her high-heeled SHOES), whereas semantically incongruent stimuli produce the N400 component. The expectedness of a stimulus relies on the appropriateness of its meaning and its physical appearance in the given context. To assess the interaction effect of these characteristics, a follow-up study was initiated, where a word could match both the physical appearance and the meaning, mismatch either the physical appearance or the meaning, or mismatch both meaning and appearance of the sentence. EEG analysis suggested that the physical appearance of the semantically mismatching word did not significantly affect the N400 amplitude or distribution. Similarly, the semantic appropriateness of a word was irrelevant for the processing of syntactic abbreviation. In other words, no interaction effect could be registered.

That the N400 can be seen as an electrophysiological correlate of semantic integration was again shown in several priming experiments, where the N400 amplitude could be manipulated by pairing the stimulus with a preceding stimulus that was either semantically related or unrelated. For example, if the word "bread" had been preceded by the word "butter", the N400 was significantly lower in comparison to the word pair "cloud "/ "bread" (Holcomb & Neville, 1990).

One issue of these studies is that the information available for processing is limited to the linguistic domain, whereas in our everyday life, we integrate non-linguistic cues from all modalities (visual, tactile etc.) to comprehend verbal information (Tromp et al., 2017). Resultantly, it does not clarify

whether semantic and syntactic processing are modality-specific or whether the system is common to all modalities. In their paper, Connoly et al. (Connolly, Byrne & Dywan, 1994) address this knowledge gap by using a cross-modal priming paradigm. They found that, when pictures were presented as primes and spoken words were presented as targets, the N400 was significantly larger in incongruous stimulus pairs than in congruous pairs. Moreover, word stimuli that were not semantically primed elicited larger N400 components than primed stimuli, supporting the hypothesis that the system responsible for semantic processing can be accessed by all modalities.

In our study, we want to go one step further by using virtual reality (VR) to create a well-controlled, more realistic experiment setting and to collect data that is more accurate to real world scenarios. We thereby expect to find further support for the commonality of semantic and syntactic processing to all modalities.

VR technology is a commonly used method in neuroscience research, because it simulates realistic environments and social interactions while eliminating possible distractions and confounding factors that often have unwanted influence on the data. In an experiment, Tromp et al. (Tromp et al., 2017) recorded electrophysiological brain activity during semantic integration in a naturalistic environment. The subject was placed in a restaurant setting, where he/she could see the food that other guests had ordered. In addition, the subject received auditory information about the dish that the guest has ordered (e.g. “I just ordered this salmon.”). As predicted by the authors and concurrent to the findings illustrated in section 1.1, a mismatch between the auditory information and the visual information elicited a strong N400 effect.

Encouraged by the success of this study, our goal is to validate the use of VR for studying semantic and syntactic processing. Our experiment will involve cross modal priming, where an auditory stimulus is used as a prime and a three-dimensional visual object as a target. A semantic incongruence is a mismatch between the object that is visually presented and the object that is communicated to the participant via speakers. To create syntactic violations, we rotate the object by 180 degrees and let it hover over the ground. Moreover, we measure the reaction time for selecting an object category and to collect behavioural data on the expectedness of our stimuli to correlate this with the ERP using a 5-point likert scale.

1.2 Hypotheses

If the setup is successful, we expect a prolonged reaction time for classifying the object if the audio priming was incongruent to the displayed 3D object. The priming anticipates a certain expectation of what object will appear on the screen and will affect the processing of the incoming stimulus. The

rotation of the visual stimulus should also have an influence on the reaction time, as it differs from the way we would expect an object to be rotated in the real world. Additionally, we think that the participants are more likely to choose the category of the stimulus according to the visual appearance and not on the audio input. Participants will rate a lower expectedness level for stimuli with incongruent audio as well as syntactically inappropriate stimuli. However, we expect a higher expectedness level for incongruent visual stimuli, when the object is up-side down, compared to an audiovisual mismatch. We assume that a mental rotation occurs fast enough ending up in less confusion.

Corresponding to previous research, we assume that the semantically incongruent objects in our experiment will elicit a N400 component and the syntactically incongruent objects a P300 component, which would replicate the results from Kutas and Hillyard (Kutas & Hillyard, 1981). In this study there is no measurement for electrophysiological brain activity (EEG) in use, therefore we are measuring the effects indirectly with the help of the reaction time and the expectedness rating.

2. Experiment

2.1 Method

2.1.1 *Participants*

We recruited 10 participants (7 females, 3 males) aged between 18 and 26. It should be noted that the participants were all students at a university in a german/western society. The participants have received an invitation through the email distribution list of the University of Osnabrück. All participants had an average knowledge of English, had normal or corrected-to-normal vision and normal hearing, and had no history of hearing problems or neurological diseases. Participants provided written informed consent and did not earn any money for participation. Participation was discouraged if the participant had previously suffered from motion sickness in experiments or travelling. Ethical approval for the study by the ethics board of the Cognitive Science Faculty of Osnabrück University was not prescribed and therefore not necessary. No participants were excluded from the analysis due to technical failures during the experiment.

2.1.2 *Materials and design*

The experiment took place in an immersive virtual reality that was custom-made using Unity (Editor version 2021.3.8f1). The environment consisted of a landscape which was green, with some hills, grass and flowers. The participants were instructed to follow the brown path, which was structured as

a circle. Objects appeared along the path at certain points at a certain time. Two different objects were used in the experiment, a tree (Fig.1) and a house (Fig.2). These two objects were chosen because it could be safely said that the western participants in the study know these objects from their natural environment. To avoid distraction no other objects were displayed. The environment can be seen as realistic.

Fig.1: Screenshot of the VR tree object

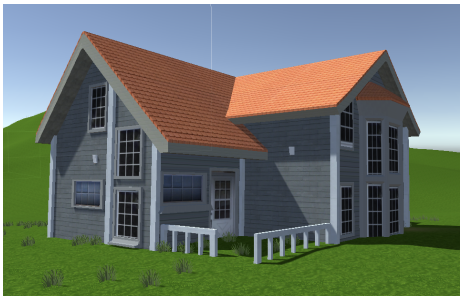

Fig.2: Screenshot of the VR house object

Fig.3: Screenshot of the VR landscape map

The participants controlled their walking by themselves with the help of a controller in their hand, they did not physically walk themselves. This allows as much self determination as possible in an experimental setting and a close to real life experience. It helps to make the experiment more interesting for the participant and to keep their attention on the screen to navigate their way.

The materials consisted additionally of two experimental sentences used for the audio condition:

Sentence 1: “Do you see the tree over there?”

Sentence 2: “Do you see the house over there?”

These sentence stimuli were produced with a free online Text-To-Speech Converter (https://ttsmp3.com). The speaker is represented by a male voice. The artificially compiled sentences allow for better experiment control, since they have similar pitch frequency and pacing across the two sentences. The corresponding sentence was played two seconds before the object was displayed. The sentence was congruent when the correct object was named and otherwise incongruent.

Key presses (Q = tree and E = house) correspond to the feedback of the participant after each stimulus. The response time (= RT) was measured with the help of Unity from time onset of the visual stimulus until the key press of the participant.

A 5-Point Likert Scale was used for the judgement of how much the presented object matched the participants’ expectation after the previous audio sentences.

The experiment is designed to be a VR study using VR goggles. This study is a pilot study, where the participants complete the experiment on their own computer at home.

The experiment had a within-subject design with 2 factors (VISUAL CONDITION) x 2 (AUDIO CONDITION) factorial design. For this 2 x 2 factorial design, there were 2 * 2 = 4 different experimental conditions.
Two different objects were included to make sure that the effect is not specific to the object. To have a within-subject design each participant was presented with both objects and had to go through all possible conditions of each object (four times a tree condition and four times a house condition), resulting in eight different conditions. The OBJECT CONDITION was not included as a factor, as this is only an exploratory investigation of a possible effect of an object, but will not be included in the analysis.

Both variables have two factors and can be either congruent or incongruent. The variable VISUAL CONDITION refers to the object which was shown. In the congruent case, the object is displayed correctly, meaning for the tree the tree trunk is attached to the floor/ grass and the treetop is facing the sky in a realistic manner next to the path. The same holds for the house, meaning that the roof is facing the sky. If the condition is incongruent the object is displayed the other way around and hovers above the ground. The variable AUDIO CONDITION refers to the audio clip which was played. If the condition is congruent the object which is shown is named correctly by the speaker. If the condition is incongruent the object which is shown is named incorrectly.

There were three possible starting points to walk in the circle, and the order of the stimuli differs corresponding to the starting point. The starting point of each participant in the virtual reality is chosen randomly.

		visual condition	
		congruent	incongruent
auditory condition	congruent	A: Tree condition: tree *displayed* the correct way round, *audio information*: 'Do you see the tree over there?'	B: Tree condition: tree displayed the correct way round, *audio information*: 'Do you see the house over there?'
	incongruent	C: Tree condition: tree *displayed* the wrong way round, *audio information*: 'Do you see the tree over there?'	D: Tree condition: tree *displayed* the wrong way round, *audio information*: 'Do you see the house over there?'

Fig.4: Table displaying the 2x2 factorial design with the conditions for the object " tree". For the object "house" the procedure was the following: A: house displayed the correct way round, audio information: 'Do you see the house over there?'; B: house displayed the correct way round, audio information: 'Do you see the tree over there?'; C: house displayed the wrong way round, audio information: 'Do you see the house over there?'; D: house displayed the wrong way round, audio information: 'Do you see the tree over there?'

2.1.3 *Procedure*

First, the participant was welcomed to the study in the lab and they got a brief introduction about the set up and the study. They got the information that they will walk through a VR environment and they will hear a voice talking to them. They were instructed to listen carefully to what was said and to pay close attention to what they saw around them.

Prior to the actual study the participant was introduced to the scene, the controls and the instructions via a test run, in which four stimuli were presented. All objects were syntactically congruent. No data was collected during the test run. The objects are from the same category (tree and house), but different prefabs, so that in the real experiment, it is still a new stimulus. During this practice round, participants could get used to the movement and the situation of being in a VR.

Immediately after the test run, the real experiment started. The participant got the instructions "Welcome to our experiment. In the following scene you will follow a path with another person, who will make you aware of some objects in the scene. Please stay on the path and press Q if it is a tree and E if it is a house. Press Space to begin." They had to actively press space to begin the experiment when they were ready.

Participants walked one round on the map and were presented with eight objects. The objects were placed with a certain distance in between to avoid stimulus overload and irritation between the

stimuli. There have been three different starting positions and therefore three different sequences of the objects which the participants have seen after each other to minimise any order effects.

The order from starting position one was (the corresponding conditions are explained in section 2.1.2): *tree condition A, tree condition D, house condition C, tree condition C, house condition A, house condition D, tree condition B, house condition B*. Starting position two was *house condition C* and the last stimulus *tree condition D*. Starting position three was *house condition A* and the last stimulus *tree condition C*.

The participant walked on the path and no object was visible. After crossing an invisible plane, the audio cue was played. Two seconds later an object appeared either to the left or to the right of the path.

After each object the participants were asked in the form of a text, what object it was ("PLEASE SELECT: this is a … tree = Q or house = E"). Participants' reaction time, how fast they decided on an object, was measured. After that, participants were told to indicate the expectedness of the stimulus using a 5-Point Likert Scale ("Please type a number between 1-5 to rate the following sentence: *"The presented object was like I expected after hearing the voice."* (1 = strongly disagree and 5 = strongly agree)). The experiment ended with an end scene saying "Thank you for your participation".

2.1.4 *Data Preparation*

As the contents of our study require the participants to be able to hear and understand English, we are using the indicated native language and hearing and vision as exclusion criteria prior to participation.

In addition, we excluded the data from any participant who answered the question with a reaction time larger than 4 seconds, as this can be interpreted that the answers might be falsified due to the time delay and are not strong enough for defining the effect of the experiment. Trials with a reaction time equal to zero were also excluded. If more than 3 stimulus trials from one participant were invalid or the participant spent in total more than 20 mins walking around in the environment, the whole data point from the participant was excluded. The cleaned data set has been used for statistical analysis.

3. Results

The data from 10 participants was collected. For one participant two trials had to be excluded as the response time was equal to zero (House CI, House IC). Four trials, two of which were completed by the same participant, were removed because the participant did not respond in 4 seconds (Tree II, Tree CI, House II, House CC). The participants' data still got used as the majority of the stimuli trials have been correct. In total, 6 stimuli trials got excluded from 4 participants.

The analysis has been done with Python (Version 3.7.14) in Jupyter Notebook (Google Collaboratory). All plots have been created using the libraries Seaborn and Matplotlib.

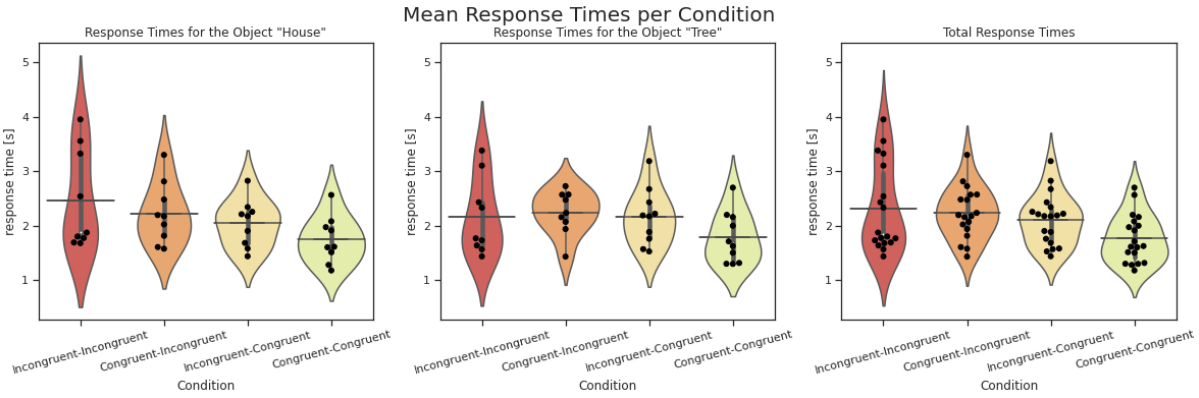

Fig. 5: Violin plot representing the mean response time for each condition. On the x-achses the condition is displayed, with the audio condition specified first (either congruent or incongruent) - and the visual condition specified second (either congruent or incongruent). The y-achses correspond to the response time in seconds. The left plot shows the response times for the object “House”, the middle plot for the object “Tree”, the plot on the right for both conditions together. Each black dot represents one participant response.The grey horizontal dash indicates the mean response time. The first state corresponds to the audio condition, the second to the visual condition (i.e. Congruent-Incongruent: Auditory Match - Visual Mismatch)

Fig. 5 gives an overview of the mean response times in each condition, displayed separately for the two possible stimulus objects (Fig. 5.1 & 5.2) as well as averaged over both objects (Fig. 5.3). The average response time for the “house” stimulus was 2.12 seconds and for the “tree” stimulus 2.08 seconds. Looking at each condition separately, the participants took on average 1.76 seconds to respond to auditory and visually congruent trials, 2.23 seconds for auditory congruent and visually incongruent trials, 2.10 seconds for auditory incongruent and visually congruent trials, and 2.31 seconds for auditory and visually incongruent trials.

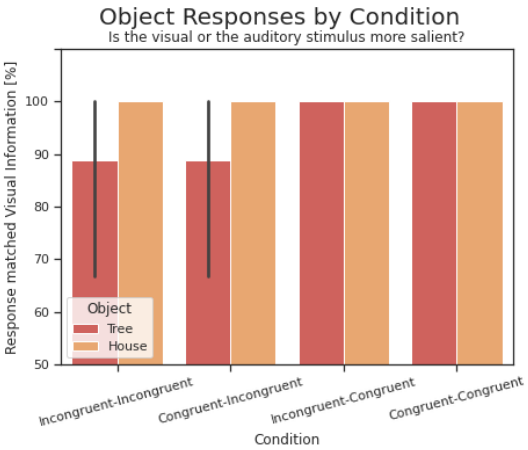

Fig. 6: Bar chart representing the classification of the object for each condition. The x-achses displays each condition combination and the y-achses the percentage of how many responses matched to the visual condition. The first state corresponds to the audio condition, the second to the visual condition (i.e. Congruent-Incongruent: Auditory Match - Visual Mismatch)

Data analysis for the object classification (Fig. 6), where the object could be classified as “house” or “tree”, reveals that perceiving the verbal audio information more saliently than the visual information was the exception. In two trials (Tree CI, Tree II), completed by two different subjects, the subject pressed a button that corresponded to the object verbally communicated. Notably, in both cases, the recorded response time (2.232s; 1.433s) was much lower than the conditions average time. In all other 74 stimuli trials, the response matched the stimulus that was presented on the screen.

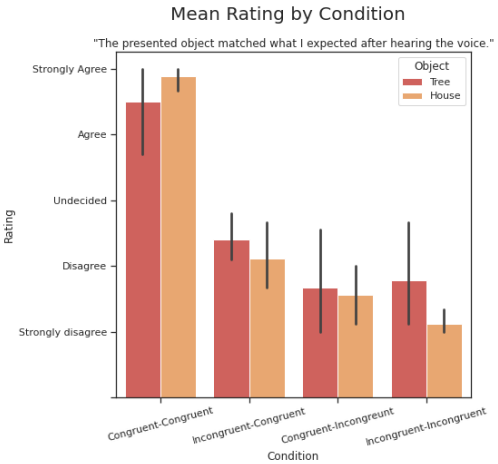

Fig. 7: Bar chart presenting the mean expectedness rating per condition. The four conditions (i.e. Congruent-Incongruent: Auditory Match - Visual Mismatch) are displayed on the x-axis and the rating on the y-axis.

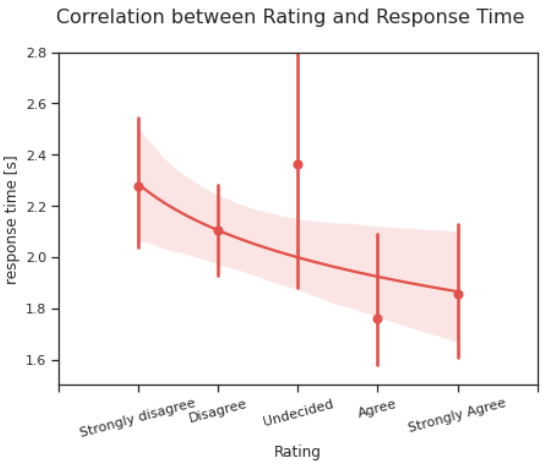

Fig. 8: Regression plot showing the correlation between the expectedness rating and the response time. Rating options are shown on the x-axis and response times on the y-axis. The individual points represent the mean response time per option.

Additionally to the response time, we have collected behavioural data to test the effectiveness of our stimulus manipulation to elicit unexpectedness. Fig. 7 shows that completely congruent stimuli were rated with high expectedness, meaning that the object matched what participants expected to see after being primed. The barplot suggests that subjects were more surprised when the presented object did not match the prime (Incongruent-Congruent) than when the object matched the prime but was presented inappropriately (Congruent-Incongruent). When the object did not match the prime and

additionally did not have the physically appropriate rotation, participants rated the stimulus to be the most unexpected. In the latter three cases, it appears that the tree object matched the participants' expectation better than the house object.

To validify the behavioural data, we correlate the expectedness rating with the response times (Fig. 8). We can see a negative correlation, meaning that the higher the response time, the lower the stimulus was rated in terms of expectedness.

4. Discussion

The aim of this study was to test the effects of auditory priming on the perception of semantic and syntactic inconsistencies in the visual domain in a VR environment. Participants were immersed in a virtual environment, a green landscape, in which they were following a path. The participant saw in total eight stimuli for which the audio priming was either congruent to the object (e.g. Audio: “Do you see the tree over there?”/ object visually displayed: tree) or incongruent (e.g. Audio: “Do you see the tree over there?”/ object visually displayed: house). The visual object was additionally either congruent to the laws of physics or incongruent (displayed up-side down).

In line with our prediction, the manipulations we made resulted in reliable response times and expectedness of the stimulus. This shows that the auditory priming has most likely an influence on the perception and processing of the stimulus object.

The results indicate, as hypothesised, that congruent stimuli were processed more quickly than incongruent stimuli. Furthermore, a mismatch between the verbal information and the actual object led to longer response times than objects that matched the audio information, but were syntacilly deviant from the subjects expectation. In combination with the results from the object classification task, one can conclude that the visual input is more significant for perceiving and processing the information in our world, but it is important to note that the verbal inputs still have an influence on the process and should not be underestimated and neglected. However, the results may have been different, if the object would have been presented before the audio clip, as the auditory information would have been more present in working memory. This leads to the assumption that the audio priming makes a difference in the brain activity during the perception process.
This is also visible in Fig. 7, which represents the results from the 5-point likert scale used to check on the expectedness of the stimulus. In our opinion this is an essential measure to assume a N400 and

P300 component. However, it is important to point out that this is just an assumption and further research that focuses on this question would be needed to ascertain whether this is actually the case.

The results indicate that congruent conditions matched better with the expectation of the participants. However, against our expectations the influence of the visual condition on the expectedness was larger if it was incongruent, meaning the stimulus was turned up-side down, compared to the congruent condition, when a completely different object was represented. From this it can be concluded that either no mental rotation has taken place, or it was not fast enough. Besides the effect of audio priming, we can assume that the degree of lifelikeness of the object influences the expectedness rating as well.
The correlation between the expectedness rating and the response time (Fig. 8) indicates that an unexpected stimulus is connected to a longer response time. Following up on the top-down process (von Stein, Chiang & König, 2000), it can be assumed that an object, which does not fit into our expectation, takes longer to process because we cannot use the structures of previously learned categories in our brain. And it therefore takes longer to filter and process the unexpected input.

Our study is a good follow up on the language processing study from Trump et al. and offers a possibility for probable assumptions, although there is still space for improvement.

Firstly, the free motion in the environment offers self determination and real life experience, but it also limits the control of the movement from the participant. In a follow up study we would recommend to include invisible boundaries on both sides of the path which will tell the participant to stay on the path. Free movement would still be given, but at the same time the risk of participants not following the path would be reduced.

Secondly, it cannot be said with certainty whether an habituation effect occurred or not as the experiment design repeats a pattern. The participants may no longer be surprised to see an upside-down house or tree after several stimuli, which would falsify the expectedness rating. However, such a bias was not observed in the data.

Moreover, our results only support that the auditory priming has an influence on the response time and the expectedness of the object in both cases, from which we could indirectly assume that the stimuli elicit a N400 and P300 component. As mentioned previously to find significant results on the topic and to make sure that the semantically incongruent objects in our experiment will actually elicit a N400 component and the syntactically incongruent objects a P300 component, further research with a broader experiment setup including EEG would be needed. With the suggestions made in this section, the same experiment could be implemented with the use of EEG. The participants will sit in a room

wearing VR goggles and additionally an EEG cap. Measuring the electrophysiological brain activity and analysing the data with the help of the Matlab EEGlab toolbox. In addition a greater number of participants would be beneficial to decrease the influence of outliers and other sources of error and to ensure more reliable results.

References

Connolly, Byrne & Dywan. (1994). *Assessing adult receptive vocabulary with event- related potentials: An investigation of cross-modal and cross-form priming.* Taylor & Francis Online. Retrieved from https://doi.org/10.1080/01688639508405145. (Retrieved 2022, September 16)

Holcomb & Neville. (1990). *Auditory and Visual Semantic Priming in Lexical Decision: A Comparison Using Event-related Brain Potentials*. Taylor & Francis Online. Retrieved from https://doi.org/10.1080/01690969008407065. (Retrieved 2022, September 17)

Kutas & Hillyard. (1981). *Event-related brain potentials to semantically inappropriate and surprisingly large words.* Elsevier. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/0301051180900460. (Retrieved 2022, September 18).

Tromp et al. (2017). *The combined use of virtual reality and EEG to study language processing in naturalistic environments.* Springer Link. Retrieved from https://doi.org/10.3758/s13428-017-0911-9. (Retrieved 2022, September 18).

von Stein, Chiang & König. (2000). *Top-down processing mediated by interareal synchronization.* PNAS. Retrieved from https://doi.org/10.1073/pnas.97.26.1474. (Retrieved 2022, September 20).

Work schedule:

Name	Workload
Johannes	
Luna	- Paper research for finding the topic - Unity functions + Collecting Data - Data Analysis (python) for the results + visualisation (plots) - Paper: • Introduction + Hypotheses • Results
Nele	- Paper research for finding the topic - Paper: • Abstract • part of the discussion - Proofreading
Eva	- Unity Environment design + some functions - Paper: • Methods sections • Hypotheses • Discussion • References - GitHub repository