# Proposal for a Master's thesis

Eva Richter

December 13, 2023

**Abstract**

Event-related potentials (ERPs) are crucial for language comprehension, however, there is no agreement on the cognitive processes which they index. Multi-stream models assume that, in line with traditional interpretations, the N400 indexes semantic integration, while Retrieval-Integration (RI) theory suggests that the N400 indexes lexical retrieval and the P600 indexes integration processes. Recently, Aurnhammer et al. (2023) tested the two hypotheses using a context manipulation design in which plausibility was varied at three levels and, in addition, a semantically attractive alternative was present in one condition. They found a graded P600 effect, which supports the interpretation of RI theory that the P600 indexes integrative processes and no N400 effects between conditions.

The aim of this thesis is to further examine the assumptions of multi-stream models and RI theory and provide support for one of the interpretations of the ERP components. For this study, the design of Aurnhammer et al. (2023) was modified in such a way that the three conditions vary in plausibility, but no longer regarding the presence of a semantically attractive alternative, which was eliminated. In this case, the predictions of multi-stream models and RI theory are even more contradictory: while multi-stream models predict a graded N400 effect, RI theory predicts a graded P600 effect. A self-paced reading study, which has already been conducted, shows that reading times are graded for target word plausibility, reflecting graded integration effort. Since behavioural measures such as reading times are not indicative of underlying cognitive processes and thus the assumptions of the conflicting hypotheses, an EEG study still needs to be conducted.

As the predictions of reading times and P600 amplitude are closely tied to plausibility ratings, another objective of this thesis is to assess the effectiveness of plausibility ratings as a predictor and investigate possible improvements by using plausibility ratings on a per-trial basis rather than relying on average plausibility ratings.

## 1 Introduction

Event-related potentials (ERPs) are important for studies of language comprehension since they capture the neural processes associated with cognitive functions, which are not reflected in behavioural studies. Traditionally, the N400 has been interpreted as an index of semantic integration (Hagoort et al., 1993) and the P600 has been thought to reflect syntactic violations (Osterhout and Holcomb, 1992). However, with the increasing number of observations of a 'semantic P600' (Kim and Osterhout, 2005; Berkum et al., 2005; Hoeks et al., 2004), complex multi-stream models emerged, attempting to reconcile the traditional views with the findings of the 'semantic P600' (Bornkessel-Schlesewsky and Schlesewsky, 2008; Kupferberg, 2007). In recent years, multi-stream models have been increasingly challenged by single-stream accounts that suggest a reinterpretation of the ERP components. One example is the single stream Retrieval-Integration (RI) account, which interprets the N400 as an index of lexical retrieval and the P600 as an index of integration processes (Brouwer et al., 2012; Brouwer and Crocker, 2017).

Aurnhammer et al. (2023) tested the conflicting assumptions of multi-stream models and RI theory using a context manipulation design, in which plausibility varied at three levels and a semantically attractive alternative was either available or not. The results revealed a graded P600 effect for plausibility, reflecting integration effort, and no N400 effect across conditions, contradicting the predictions

of multi-stream models. While multi-stream models can account for the P600 effect in the presence of an attractive alternative, they struggle to explain the presence of the P600 and absence of the N400 effect while no attractive alternative is available. Therefore, the primary goal of this study is to eliminate the attractive alternative while maintaining graded plausibility across conditions. If a self-paced reading study again shows graded reading times for plausibility, an EEG study will be conducted. Based on the modified stimuli, the hypotheses of the multi-stream account, which predicts a graded N400 effect, and the single-stream RI account, which predicts a graded P600 effect, can then be further tested.

The proposal is organised as follows: Chapter 2 provides an overview of the background and related work, Chapter 3 outlines the research questions, Chapter 4 addresses data and methodology, Chapter 5 describes the conducted and planned experiments and Chapter 6 concludes.

# 2 Background and related work

This section introduces theoretical concepts related to the ERP components, the N400 and the P600, which are associated with language processing in the brain, as well as multi-stream models and Retrieval-Integration theory, a single-stream account. It also describes the concepts of plausibility and surprisal and their relationship to behavioural and neurophysiological measures.

## 2.1 ERP components

Event-related potentials play an important role in electrophysiological studies of language comprehension since they offer detailed insights into how language comprehension unfolds in real-time. While behavioural studies such as reading times measure the overall processing effort, they cannot capture the mechanisms which are underlying and driving this effort. In contrast, ERP signals provide high temporal resolution, allowing the analysis of cognitive processes linked to language comprehension at the millisecond level.

The two most prominent ERP components in the context of language comprehension are the N400 and the P600. As indicated by the name, the N400 describes a negative-going voltage deflection that peaks approximately 400 ms after stimulus onset, although the negativity extends over a period of 250-500 ms after stimulus presentation (Kutas and Federmeier, 2009).

The N400 was first discovered in 1980 by Kutas and Hillyard (1980), who observed the component in response to sentences ending with unexpected words compared to expected words (e.g. "he spread his warm bread with **socks**" compared to "**butter**") (Kutas and Federmeier, 2011, for a review). Moreover, it was found that the N400 is not just sensitive to the expectancy of words, but rather to the expectancy of words in their immediate context, such that words which are semantically associated or related to the preceding context yield a semantic priming effect, reflected in a smaller N400 amplitude compared to the same words out of context (Kutas and Hillyard, 1984). Building on these findings, the N400 has for many years been widely regarded as an index of semantic integration processes (Brown and Hagoort, 1993; Kutas and Federmeier, 2011). Kutas and Federmeier (2000) showed that in addition to the immediate linguistic context, relationships stored in long-term memory also play an important role in language processing. In addition, the N400 has been shown to be sensitive to factors such as word frequency (Rugg, 1990) and repetition (Rugg, 1985; Petten and Kutas, 1991).

The P600, on the other hand, is an ERP component characterised by a positive-going deflection with an onset around 500-600 milliseconds ms after stimulus presentation. It was first documented by Osterhout and Holcomb (1992), who found that P600 effects occurred when the sentence structure deviated from the preferred sentence structure, and thus interpreted the P600 as an index of syntactic anomalies. Subsequent studies further analyzed the P600 as a response to syntactic violations (Hagoort et al., 1993; Gouvea et al., 2010; Kaan and Swaab, 2003). Hagoort et al. (1993) observed that a P600 amplitude relative to baseline is present in instances of agreement errors, as demonstrated by sentences like "the spoilt child **throw** the toys on the floor", where a grammatical error in number agreement between subject (child) and verb (throw) exists. This P600 response has also been identified in various

other forms of grammatical inconsistency, including tense, gender, number, and case, as reported Gouvea et al. (2010). Moreover, P600 effects have been found in the absence of grammatical errors, e.g. when reading garden-path sentences such as "the broker persuaded to sell the stock was sent to jail" (Osterhout and Holcomb, 1992), which initially leads the reader to a misleading interpretation of the sentence in which a noun is expected after "persuaded". Other interpretations assume that the P600 reflects processes of revision for grammatical but unpreferred sentence structures and repair for ungrammatical sentences when compared to grammatical and preferred structures (Kaan and Swaab, 2003).

These traditional interpretations of the N400 as an index of semantic integration and the P600 as an index of syntactic violations have been challenged since several studies (Kim and Osterhout, 2005; Hoeks et al., 2004; Berkum et al., 2005) found a P600 effect in grammatically correct but semantically anomalous sentences. These findings sparked controversy about the interpretation of the language-sensitive ERP components and ultimately resulted in multi-stream models (Kim and Osterhout, 2005; Bornkessel-Schlesewsky and Schlesewsky, 2008; Kupferberg, 2007) as well as single-stream models, such as Retrieval-Integration theory (Brouwer et al., 2012).

## 2.2 Multi-stream models

Multi-stream models have emerged primarily in response to the challenge of reconciling the traditional interpretations of the N400 as an index of semantic integration and the P600 as a syntactic component with later studies that found P600 effects for semantic anomalies (Kim and Osterhout, 2005; Hoeks et al., 2004; Berkum et al., 2005). Kim and Osterhout (2005) found so-called 'semantic P600' for sentences including attraction violations ("the hearty meal was **devouring**") compared to control sentences ("the hearty meal was **devoured**") instead of the expected N400 effect. Simultaneously, they found an N400 effect for non-attraction violations, i.e. when the initial noun phrases were implausible ("The dusty tabletops were **devouring**"). They hypothesized that this could be attributed to the perception of the syntactically error-free sentence as syntactically incorrect due to the strong semantic attraction between "devoured" and "the hearty meal" and concluded that strong semantic attraction can override syntax. However, to explain the increase in 'semantic P600' effects rather than N400 effects despite semantic anomalies, multi-stream models emerged. Regardless of the number and specific labels of the processing streams, which different multi-stream models apply to them, they share the idea that there are at least two processing streams: a processing stream of semantic nature and another algorithmically-driven processing stream which follows syntactic constraints (Kim and Osterhout, 2005; Bornkessel-Schlesewsky and Schlesewsky, 2008; Kupferberg, 2007). The semantic processing stream is controlled by the availability of a semantically attractive alternative interpretation, i.e. it arises when an interpretation which is semantically more attractive than the interpretation emerging from surface structure is available ("devoured" would be a semantically more attractive alternative compared to "devouring", even though it does not correspond to the surface structure). While the semantic processing stream does not detect a semantic anomaly based on surface structure, as the anomaly is repairable due to the presence of a semantically more attractive alternative, the algorithmically controlled processing stream detects the anomaly created by the semantic stream, which leads to a conflict between the two streams. Multi-stream models assume that this conflict manifests itself in a P600 effect, which can be seen as a revision of the interpretation constructed up to that point. This means that, according to multi-stream models, the occurrence of a P600 effect instead of an N400 effect depends on the repairability of a semantic anomaly, i.e. on the availability of a semantically more attractive interpretation than the one encountered in surface structure. Consequently, multi-stream models predict a P600 effect when such an attractive alternative interpretation is available and an N400 effect, representing the effort of integrating a semantically less plausible interpretation, when no semantically attractive alternative is available.

However, even these explanations of the language-sensitive ERP components by multi-stream models do not always explain the occurrence of P600 instead of N400 effects in response to semantic anomalies, for example with respect to the findings of Berkum et al. (2005), who observed a 'semantic P600' when a broader context was considered. Berkum et al. (2005) created a design in which a context

paragraph is either followed by continuations that are coherent due to a plausible target word (e.g. "Next, the woman told the **tourist**") or continuations that are incoherent due to an implausible target word (e.g. "Next, the woman told the **suitcase**"). Since the continuation with the implausible target word ("suitcase") contains a semantic anomaly and there is no semantically attractive alternative that could repair the semantic anomaly, there is also no conflict between the two processing streams and therefore multi-stream models predict an N400 effect for implausible target words compared plausible target words. Instead, Berkum et al. (2005) found a P600 effect for the continuation containing the implausible target word compared to the continuation containing the plausible target word and no N400 effect. It was argued that no P600 effect was observed due to the repetitions of both the plausible and implausible target words, making the plausible target word (e.g. "tourist") available as a semantically attractive alternative to the implausible word (e.g. "suitcase") in the larger context, i.e. globally (Brouwer et al., 2012). Kupferberg (2007) hypothesized that one explanation for this result might be that the content of the context paragraph is semantically associated with the anomalous word (e.g. "suitcase" in the context of the travel scenario described in the context), which causes a conflict between coherent semantic associations and violated thematic roles that are syntactically assigned by a verb to its arguments, resulting in a P600 effect.

A recent study by Aurnhammer et al. (2023), however, found a P600 effect for semantically implausible target words in the absence of a (global) semantically attractive alternative in a broader context. This contradicts the assumptions of multi-stream models, according to which an N400 effect should have been observed.

## 2.3 Retrieval-Integration theory

As discussed in 2.2, multi-stream models cannot explain all findings of 'semantic P600' and have therefore been increasingly challenged by newer ideas that propose a reinterpretation of the ERP components instead of complex multi-stream models (Kutas and Federmeier, 2011; Brouwer et al., 2012; Brouwer and Crocker, 2017). Brouwer et al. (2012) review various multi-stream models and conclude that none of them can account for all the findings of the 'semantic P600'. In particular the results of Berkum et al. (2005) discourse processing study cannot be explained by any of the models. Instead, they propose Retrieval-Integration (RI) theory, a single-stream architecture. According to RI theory, the N400 is reinterpreted as an index of lexical retrieval from long-term memory, a process in which all the features of an incoming word are retrieved. The ease of retrieval increases as the features of the incoming word align more closely with the features pre-activated in memory. In other words, context plays a crucial role. Based on the preceding words, features for subsequently expected words are pre-activated. The greater the disparity between the features of incoming words and those pre-activated in memory, the more effort is needed for retrieval, resulting in a higher N400 amplitude Apart from contextual priming, however, words can also be primed lexically, e.g. through repetitions, as well by semantic association and world knowledge ("he spread the warm bread with **socks**" goes against world knowledge compared to "he spread the warm bread with **butter**" which is in line with world knowledge and also contains the semantically associated words "bread" and "butter" (Kutas and Hillyard, 1980), resulting in facilitated retrieval). Simultaneously, Brouwer et al. (2012) suggest redefining the P600 as integration of the retrieved word features into the unfolding semantic representation of the utterance. Integration is more difficult, reflected in a higher P600 amplitude, when the unfolding representation requires more reorganization and updating to establish a coherent interpretation. Under this assumption, the language-sensitive ERP components are two successive phases in which the output of the retrieval stage serves as input for the integration stage. Thus, retrieval and integration can be seen as sub-processes of an overarching process function that maps an incoming word and its preceding context to an utterance interpretation of these words (Brouwer et al., 2021b). Consequently, Brouwer et al. (2012) consolidate the re-interpretation of the two ERP components, labelling them as lexical retrieval (N400) and integration (P600), into a unified framework known as the Retrieval-Integration (RI) theory.

Returning to the example in section 2.2, the 'semantic P600' which Kim and Osterhout (2005) observed, can be explained by RI theory. In the sentence "the hearty meal was **devouring**/**devoured**",

"devouring" and "devoured" receive the same contextual priming from "the hearty meal", so that no N400 effect appears for "devouring" compared to "devoured". However, if "devouring" is the verb, "the hearty meal" automatically becomes the subject of the sentence, which is more difficult to integrate than "devoured", making "hearty meal" the object of the sentence. This integration difficulty for "devouring" explains the observed P600 effect. The results of (Berkum et al., 2005) can also be explained by the fact that both the plausible ("tourist") and the implausible ("suitcase") target words are primed by multiple repetitions in the context paragraph, thus facilitating their retrieval from long-term memory, which explains why no N400 effect was observed. At the same time, the subsequent integration of "suitcase" into the unfolding sentence representation is more difficult because it contradicts world knowledge, which is reflected in a P600 effect for "suitcase" relative to "tourist". Since multi-stream models explain the findings of Berkum et al. (2005) with the global availability of a semantically attractive alternative due to the repetition of target words, both multi-stream models and RI theory predict a P600 effect for the implausible target word ("suitcase") relative to the plausible target word ("tourist").

To test the predictions of multi-stream models against those made by RI theory, Aurnhammer et al. (2023) modified and extended the design of Berkum et al. (2005) to a context manipulation design in which a context paragraph is followed by a final sentence in three different conditions. These conditions vary in terms of the degree of plausibility (A: plausible, B: less plausible, C: implausible), more precisely in the plausibility of the main verb with respect to the target word. Secondly, the conditions differ regarding the global availability of a semantically attractive alternative (the distractor word) in the context of condition B. Since a graded P600 amplitude ($C > B > A$) was observed, Aurnhammer et al. (2023) concluded that, given the correlation between P600 and reading times and their modulation by plausibility, P600 serves as a continuous index of integration difficulty, as anticipated by RI theory. In addition, an early negativity (250-400 ms) was observed in condition B compared to conditions A and C, which does not indicate lexical retrieval, as no N400 effect was found and is instead interpreted by Aurnhammer et al. (2023) as a lexical mismatch between the encountered target word and the expected distractor word. Under the assumptions of multi-stream models, the P600 amplitude in condition B can be explained by the presence of a semantically attractive alternative which repairs the anomaly and hence leads to a conflict between the two processing streams. The P600 amplitude in condition C, however, contradicts the multi-stream account since the absence of a semantically attractive alternative renders the anomaly irrepairable, which should instead result in a N400 effect relative to the baseline. Thus, despite relying on different premises, multi-stream models and RI theory, both predict and explain the P600 effect in condition B. However, the observed P600 amplitude in Condition C challenges the assumption of multi-stream models, which posit that the P600 amplitude arises due to the presence of a semantically attractive alternative, since no such semantically attractive alternative was available in condition C. Instead, these results provide support for RI theory and demonstrate that P600 amplitude continuously predicts integration effort. Based on these findings, the aim of this is to further test the two hypotheses, as more evidence is needed to make well-founded statements about the assumptions of multi-stream models and RI theory.

## 2.4 Plausibility and Surprisal

In general terms, plausibility can be defined as the degree to which a statement is consistent with our understanding of the world (Brouwer et al., 2021a). Apart from some special cases (Kutas and Federmeier, 2011), N400 amplitude is generally predicted to be highly sensitive to plausibility according to the traditional view of the N400 as index of semantic integration processes. RI theory predicts the N400 to be relatively insensitive to plausibility, assuming that a plausible and implausible word would both not lead to a higher N400 amplitude if they are equally primed by the preceding context (Brouwer et al., 2012). RI theory predicts that, instead, P600 amplitude is sensitive to plausibility (Brouwer et al., 2012). Words that render the sentence in which they appear, for instance by violating world knowledge, implausible, are associated with greater difficulty in updating the mental representation of the unfolding sentence with the features of the incoming word (Brouwer et al., 2012).

Based on a design that manipulates plausibility at three levels (plausible, less plausible, implau-

sible), Aurnhammer et al. (2023) found that both reading times and P600 amplitude, pattern with plausibility and that P600 amplitude is not just a binary response to syntactic or semantic violations, but is rather modulated as a function of integration difficulty by each word, i.e. continuously. The relationship between plausibility, reading times and P600 confirms that P600 is furthermore consistent with comprehension-centric surprisal, as previously assumed by Brouwer et al. (2021b).

Expectancy is similar to the concept of plausibility and indicates how much a word is expected given its preceding context. However, while an unexpected word can still be a plausible continuation, implausible continuations are always unexpected. Expectancy is often assessed using cloze tests, which indicate the percentage of readers who have completed the sentence with a specific word. While these expectancy judgements are performed by humans, it is also possible to use surprisal values (which can also be calculated from cloze probabilities as their negative logarithm) calculated by language models as an estimate of word expectancy. Surprisal originates from information theory (Shannon, 1948) and has been increasingly adapted to studies of language comprehension. Surprisal theory states that the cognitive effort incurred by a word is directly proportional to its expectation in context:

$$Surprisal(w_{t+1}) = -log_2 P(w_{t+1}|w_1 ... w_t).$$

In human language processing, an interpretation of a sentence is constructed incrementally. Since each word conveys a certain amount of information, surprisal quantifies the amount information gained when encountering a specific word, considering the preceding context (Hale, 2001; Levy, 2008). While the amount of information gained for a word can be calculated from language models, the cognitive effort involved in processing the word can be observed through measures such as reading times (Frank et al., 2015). While it is generally recognised that behavioral measures such as reading times are positively correlated with word surprisal (e.g. Smith and Levy, 2013), there is uncertainty regarding the relationship between surprisal and the ERP components. Some studies have found that N400 amplitude is predicted by surprisal (Frank et al., 2015; Frank and Willems, 2017), however, there are no such findings regarding the P600 yet. However, most of the studies examining the relationship between surprisal and the ERP components correlate surprisal values with ERP amplitudes on a word-by-word basis, which only captures the notion of linguistic experience (i.e. the linguistic input fed to the model). In contrast, Venhuizen et al. (2019) developed a model that calculates more human-like online surprisal by considering not only linguistic experience, but also world knowledge. In this model, online surprisal is reflected in a change from one point in a Distributed Situation-state Space (DSS) to another during word-by-word processing. In line with the assumption of RI theory that P600 amplitude represents the effort of integrating word meaning with the unfolding utterance representation, Venhuizen et al. (2019) suggest that this comprehension-centric notion of surprisal is linked to P600 amplitude. Combining this comprehension-centric model of surprisal (Venhuizen et al., 2019) and the neurocomputational model of the RI account (Brouwer and Crocker, 2017), Brouwer et al. (2021b) derive a model which produces N400, P600, as well as online surprisal estimates. They demonstrate that the model predictions are consistent with the obtained empirical electrophysiological results and reading times (Delogu et al., 2019) and conclude that surprisal reflects both reading times and P600 amplitude, providing evidence for the P600 as an index of comprehension-centric surprisal. The findings from Aurnhammer et al. (2023), as described earlier, further support a link between plausibility, reading times and P600, which aligns with comprehension-centric surprisal.

# 3 Research questions

As described in more detail in Sections 2.2 and 2.3, multi-stream models and RI theory differ in their assumptions about which of the ERP components, the N400 or the P600, indexes integration processes. The aim of this thesis is to further test these hypotheses and to gather evidence to support the assumptions of multi-stream models or RI theory. In summary, Aurnhammer et al. (2023) tested the predictions of multi-stream models and RI theory based on a context manipulation design in which the plausibility of the target words in a final sentence varied on three levels (A: plausible, B: less

plausible, C: implausible) and in condition B the expectancy of the distractor word was higher than that of the target word, i.e. a semantically attractive alternative was present. A study of self-paced reading first showed graded reading times for plausibility, and the subsequent ERP study confirmed the predictions of RI theory that P600 amplitude is graded for plausibility. This relationship between plausibility, reading times and P600 also confirms the assumption (Brouwer et al., 2012) that P600 is a graded index of integration difficulty on a word basis. While multi-stream models can account for the P600 effect observed in condition B by the existence of a semantically attractive alternative, they cannot explain the observed P600 effect in condition C, where no such attractive alternative is present. Given the absence of an attractive alternative in condition C, only an N400 effect would have confirmed the predictions of multi-stream models.

Based on these findings, the semantically attractive alternative is removed in condition B in the present study, ensuring that the three conditions differ only in terms of plausibility. In this case, the predictions of multi-stream models and RI theory diverge even further: multi-stream models predict a graded N400 effect for plausibility, while RI theory predicts a graded P600 effect for plausibility (Table 1).

| | **Multi-stream** | | **Retrieval-Integration (a)** | |
|---|---|---|---|---|
| | N400 | P600 | N400 | P600 |
| **A**: Plausible & no attraction | - | - | - | - |
| **B**: Less plausible & **attraction** | - | + | - | + |
| **C**: implausible & no attraction | + | - | - | ++ |
| | **Multi-stream** | | **Retrieval-Integration (b)** | |
| | N400 | P600 | N400 | P600 |
| **A**: Plausible & no attraction | - | - | - | - |
| **B**: Less plausible & **no attraction** | + | - | - | + |
| **C**: implausible & no attraction | + | - | - | ++ |

Table 1: Predictions of multi-stream models, Retrieval-Integration theory in the study of Aurnhammer et al. (2023) (a) and Retrieval-Integration theory in the present study (b) for the N400 and P600 ERP components.

As reading times are a behavioural measure of integration difficulty, the slightly modified design of (Aurnhammer et al., 2023) will first be validated in a self-paced reading study, which leads to the first research question of the present study:

1. *Do reading times pattern with plausibility* ($A > B > C$) *in the sense that lower target word plausibility is reflected in higher (slower) reading times on average?*

Ideally, the manipulation of the final sentences with respect to the context should lead to a gradual scaling of reading times with plausibility, indicating graded integration difficulty. Since the semantically attractive alternative in condition B was eliminated, it can be predicted that no significant reading time modulation due to distractor word surprisal will occur, if reading times are sensitive to unfulfilled expectations at all. If the self-paced reading study reveals graded reading times for plausibility, an ERP study can be conducted to test the predictions of multi-stream models and RI theory, as a self-paced reading time study alone does not provide information about which ERP component ultimately indexes integrative processing effort. Consequently, the second research question is:

2. *Will plausibility modulate the presence of N400 and/or P600 effects in the ERP signal? Will there be an N400 or P600 effect in condition B and C relative to the baseline?*

This allows to test the predictions of multi-stream models and RI theory and also for a direct comparison between behavioural and neurophysiological indices of integrative processing effort. Since multi-stream models lack computational specification, they do not allow for quantitative predictions.

Instead, they typically make binary predictions about when to expect an N400 or a P600 effect. Consequently, these models can only predict an N400 effect for condition B and C compared to condition A in the present design, regardless of the varying levels of plausibility. One exception is a computationally specified model developed by Rabovsky et al. (2018); Rabovsky and McClelland (2019). Their *Sentence Gestalt* model simulates N400 amplitude, which they assume to reflect automatic interpretation of a probabilistic representation of a situation or event described by a sentence. The model posits that incoming stimuli function as cues to meaning, automatically changing an activation pattern that represents estimates of the conditional probabilities of all aspects of meaning regarding the situation/event the sentence describes. Thus, a greater change in a probabilistic representation of meaning, introduced, for instance by a less plausible word, (equivalent to higher prediction error) is reflected in higher N400 amplitude. Under the assumptions of this computationally specified model, a graded N400 effect (++) would be expected for conditions C and B relative to condition A.

Aurnhammer et al. (2023) found an early negativity (250-400 ms after stimulus onset) in condition B compared to conditions A and C. They hypothesised that this negativity occurs in the context of unfulfilled expectations, i.e. due to a lexical mismatch between the encountered target word and the expected distractor word, as it was the case in condition B. In the context of the present study this raises the following question:

3. *Will the early negativity (250-400 ms) observed in Aurnhammer et al. (2023) disappear after the attractive alternative was eliminated, i.e., after the distractor word expectancy was lowered in Condition B?*

If the assumption that the early negativity occurs in the context of unfulfilled expectations is correct, then it should disappear in the context of the present study.

As the three research questions outlined above rely strongly on plausibility as a predictor of reading times and P600 amplitude, it is worth investigating the validity of these plausibility ratings and possible improvements. A common practice to assess plausibility involves collecting plausibility ratings in a pre-study, which can be used to predict, for example, reading times in subsequent statistical analyses. Given that these ratings are not obtained from the participants in the main experiments, only the average ratings per item and condition can be used as a predictor. Instead, using individual ratings collected online during self-paced reading or in EEG studies could improve predictions.

A series of studies by Troyer and Kutas (2018); Troyer et al. (2019); Troyer and Kutas (2020) revealed evidence that variability in world knowledge impacts the access to information during real-time word processing. Particularly, in Troyer et al. (2019) participants with varying degrees of Harry Potter (HP) knowledge read contextually supported sentences and reported whether they had known the information presented in the sentence or not, while an EEG was recorded. Based on single-trial EEG analysis they revealed that the influence of HP domain knowledge on the ERPs extends beyond offering a proportion of information known by an individual. Instead, participants' single-trial reports showed a strong correlation with HP domain knowledge and predicted ERPs to supported HP sentence endings. Similarly, one goal of this thesis is to obtain online plausibility ratings during a self-paced reading and during an EGG study to examine whether these per-trial ratings serve as a more effective predictor of reading times and ERPs compared to average plausibility ratings collected in a pre-study. Given that per-trial plausibility ratings allow capturing individual differences in how participants perceive plausibility, it can be assumed that they will serve as a more accurate predictor of reading times and ERPs compared to average plausibility. However, it might also be the case that average plausibility ratings provide a more stable estimate of reading times and ERPs. In this study, both average ratings and per-trial ratings will be collected in order to address the following question:

4. *Which is a better predictor of reading times and N400/P600 amplitude: average plausibility ratings collected in a pre-study or per-trial plausibility ratings collected during self-paced reading and ERP studies?*

# 4 Data and methodology

In this section, the materials required for conducting the two pre-studies, as well as the self-paced reading study and EEG study, are described, along with the methodology to ultimately answer the research questions outlined in the preceding section.

## 4.1 Experimental Stimuli

The stimuli used for the following experiments are based on the stimuli of Aurnhammer et al. (2023) who translated and adapted items from Berkum et al. (2005) and partially developed new items. Similar to Aurnhammer et al. (2023), a context manipulation design is used. The 60 chosen stimuli all consist of a context paragraph followed by a subsequent final sentence manipulated on three levels in terms of its plausibility in relation to the context paragraph. To prime the target and distractor words, both are repeated three or four times in the context paragraph. Which word, i.e. whether the target or the distractor word was mentioned last in the context paragraph was distributed approximately evenly across all items. Under the framework of RI theory, the repeated mention of the words should ease their retrieval, resulting in the absence of an N400 effect across all conditions, irrespective of their plausibility. While the context paragraph is the same across all conditions within each item, the final sentence varies across conditions regarding the main verb which renders the sentence either plausible (Condition A, "the lady dismissed the tourist"), less plausible (Condition B, "the lady welcomed the toruist"), or implausible (Condition C, "the lady signed the tourist") (Table 2). In the final sentence, the target word is followed by an additional clause ("[...] and then he went to the gate.") to capture spillover effects in reading times.

In contrast to the study conducted by Aurnhammer et al. (2023), the final sentences differ only with respect to the plausibility level, but not with respect to the availability of a semantically attractive alternative. Since the attractive alternative was removed from the previously ambiguous final sentences in condition B, they are now expected to be unambiguous in each condition. This manipulation of the final sentences should shed more light on whether the assumptions of multi-stream models or RI theory hold true, as the two approaches now differ not only in their predictions for condition C, as in Aurnhammer et al. (2023), but also for condition B. While multi-stream models predict an N400 effect for conditions B and C relative to condition A due to the absence of a semantically attractive alternative, RI theory predicts a P600 effect for conditions B and C relative to condition A. Moreover, if the assumption is correct that the early negativity (250-400 ms) observed in Aurnhammer et al. (2023) in condition B relative to conditions A and C is due to a lexical mismatch between the encountered target word and the expected distractor word, it should disappear after removing the ambiguity in condition B. An example of design of the study of Aurnhammer et al. (2023) (left) and the present study (right) is shown below. For the final sentences the German word order is maintained in the English transliterations. The target words are underlined and **distractor words** are highlighted in boldface.

*Context*
Ein Tourist wollte seinen riesigen **Koffer** mit in das Flugzeug nehmen. Der **Koffer** war allerdings so schwer, dass die Dame am Check-in entschied, dem Touristen eine extra Gebühr zu berechnen. Daraufhin öffnete der Tourist seinen **Koffer** und warf einige Sachen hinaus. Somit wog der **Koffer** des einfallsreichen Touristen weniger als das Maximum von 30 Kilogramm.

*A tourist wanted to take his huge **suitcase** onto the airplane. The **suitcase** was however so heavy that the woman at the check-in decided to charge the underlinetourist an extra fee. After that, the underlinetourist opened his **suitcase** and threw several things out. Now, the **suitcase** of the ingenious underlinetourist weighed less than the maximum of 30 kilograms.*

| Design by Aurnhammer et al. (2023) | Present study |
|---|---|
| *Condition A: Plausible & no attraction*<br>Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.<br>*Then dismissed the lady the tourist and afterwards he went to the gate.* | *Condition A: Plausible & no attraction*<br>Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.<br>*Then dismissed the lady the tourist and afterwards he went to the gate.* |
| *Condition B: Less lausible & **attraction***<br>Dann **wog** die Dame den Touristen und danach ging er zum Gate.<br>*Then **weighted** the lady the tourist and afterwards he went to the gate.* | *Condition B: Less lausible & **no attraction***<br>Dann **begrüßte** die Dame den Touristen und danach ging er zum Gate.<br>*Then **welcomed** the lady the tourist and afterwards he went to the gate.* |
| *Condition A: Implausible & no attraction*<br>Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.<br>*Then signed the lady the tourist and afterwards he went to the gate.* | *Condition A: Implausible & no attraction*<br>Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.<br>*Then signed the lady the tourist and afterwards he went to the gate.* |

## 4.2 Linear Mixed-Effects Regression Re-estimation

After log-transforming reading times to normalise their distribution, the transformed reading times were analysed using a linear mixed-effects regression re-estimation method (see Aurnhammer et al. (2021)). Separate reading time models were fitted for each region (pre-critical, critical, spillover and post-spillover), which makes it possible to determine the influence, significance and residuals [1] of the predictors separately for each region. The predictors which were used in the linear mixed effects models, that were subsequently used to re-estimate the reading time data, are target word plausibility and distractor word surprisal. Target plausibility was used as a continuous predictor to quantify integration difficulty and distractor surprisal as a predictor to account for additional effort which might have been caused by the availability of a semantically attractive alternative. If distractor expectancy has been eliminated, as indicated by the surprisal values, this predictor should not have have a significant impact. In the model specification below $\beta_0$ represents the fixed-effect intercept and $\beta_1, \beta_2$ the fixed-effect coefficients (slopes) of target plausibility and distractor surprisal, while $S_0, I_0$ represent the random-effect intercepts for subjects and items and $S_1, S_2, I_1, I_2$ the random-effect coefficients (slopes) for subjects and items. The residual error term $\epsilon$ captures the variability in the dependent variable (reading times) which is not explained by the predictors.

$$Y = \beta_0 + S_0 + I_0 + (\beta_1 + S_1 + I_1)PlausTar + (\beta_2 + S_2 + I_2)SurprisalDist + \epsilon$$

The predictors were standardised, which centered their average value around zero and expressed them on the (same) scale of standard deviations. This ensures that the coefficients represent the change in the response variable corresponding to a one-standard-deviation change in the predictor. Additionally, standardising the predictors, results in the intercept being equal to the mean of the data to which the model is fitted. Finally, plausibility was inverted, as it can be predicted that lower plausibility ratings will be reflected in higher (slower) reading times. A separate analysis was performed for each of the four regions, treating each region as an independent family of hypotheses. Therefore it was not necessary to correct for multiple comparisons across regions.

[1] the difference between the observed reading times and the estimated reading times

Similar to the analysis of reading times, $r$ERPs (Smith and Kutas, 2015), an ERP re-estimation technique, will be used to analyze the EEG data, which, for reasons of computational efficiency, is based on linear regression instead of linear mixed-effects regression. While the traditional ERP method calculates condition averages based on the average of individual subjects, the $r$ERP technique allows for a distinct regression model to be computed for each subject on each electrode and time sample. Again, target plausibility and distractor surprisal will be used as predictors. The selection of linear regression over linear mixed-effects regression yields the following model specification:

$$Y = \beta_0 + \beta_1 PlausTar + \beta_2 SurprisalDist + \epsilon$$

where $\beta_0$ is the intercept, $\beta_1$ is the coefficient of target plausibility, $\beta_2$ is the coefficient of distractor surprisal and $\epsilon$ is the residual error. To obtain individual t-values and p-values for each subject per electrode and time-sample, the same model is also calculated across subjects.

# 5 Experiments

In the following sections the pre-studies on plausibility and surprisal and the two main studies are described. Results from the pre-studies and the self-paced reading study are reported, whereas the EEG study still has to be conducted.

## 5.1 Pre-tests: Plausibility and Surprisal

Plausibility ratings were collected to assess whether the plausibility manipulation was successful, i.e. whether a graded plausibility effect was achieved. The ratings were collected in a web-based experiment using the software PCIbex (Zehr and Schwarz, 2018). In this study participants rated the plausibility of the final sentence of each item in light of its preceding context paragraph on a seven-point Likert scale, with 1 indicating an implausible sentence and 7 indicating a plausible sentence. Plausibility ratings were collected for both target words and distractor words in all three conditions to assess if the main verbs were chosen in a way that resulted in high plausibility for condition A, moderate plausibility for condition B, and low plausibility for condition C. In contrast to the study conducted by Aurnhammer et al. (2023), the continuation of the last sentence ("and after that he left the store") was not excluded because plausibility ratings were also collected in the subsequent reading time study and ERP study, in which the continuation is crucial to capture spillover effects. In addition to the 60 critical items, each participant read 12 filler items that could be either plausible, less plausible, or implausible, analogous to the critical items. The purpose of the filler items was to make sure that participants read the texts carefully. Therefore, filler items contained instructions in the middle of the paragraph asking participants to rate the trial either 1 or 7, regardless of the actual plausibility of the sentence. If more than two of these 12 attention checks were failed, the participant's data were excluded from further analyses.

A total of 66 participants were recruited for this purpose on Prolific Academic Ltd and paid £4.95. Four participants were excluded due to exceptionally fast completion of the study (more than 3 standard deviations below the mean) or more than 2 out of 12 failed attention checks. Two additional participants were randomly excluded to keep the number of participants per list equal. On average, participants rated 99% (mean = 99.16%, SD = 2.52, range = 91.66%-100%) of the items that contained attention checks correctly, i.e., with the number indicated in the context paragraph. Table 2 displays the statistics of the collected plausibility ratings for both target and distractor words across different conditions.

The plausibility ratings show that average target word plausibility is stepped ($A > B > C$) across conditions, indicating that the main verbs in the final sentences were successfully selected such that A was considered to be highly plausible on average, B was considered to be medium plausible on average, and C was considered to be implausible on average. As for the distractor word, the average plausibility ratings across conditions show the same gradation pattern ($A > B > C$), that is, condition A was rated with the highest average plausibility, condition B with medium average plausibility and

condition C with the lowest average plausibility. However, the average ratings per condition for the distractor word are significantly closer to each other compared to the target word. The fact that distractor word plausibility was rated lower on average for Condition B than for Condition A indicates that, in contrast to Aurnhammer et al. (2023), the distractor words no longer represent a semantically attractive alternative to the target words in Condition B and that the ambiguity in Condition B has thus been successfully eliminated.

| | | **Plausibility** | | | **Surprisal** | | |
|---|---|---|---|---|---|---|---|
| | Cond. | Mean | SD | Range | Mean | SD | Range |
| Target | A | 6.03 | 0.71 | 4.40-7.00 | 2.36 | 2.33 | 0.06-10.52 |
| | B | 3.79 | 1.20 | 1.70-6.80 | 3.95 | 3.58 | 0.03-16.76 |
| | C | 1.91 | 0.57 | 1.00-3.30 | 6.61 | 4.70 | 0.13-18.71 |
| Distractor | A | 2.97 | 1.48 | 1.20-6.80 | 6.79 | 4.98 | 0.24-21.67 |
| | B | 2.92 | 1.41 | 1.10-6-40 | 6.55 | 4.41 | 0.15-20.90 |
| | C | 2.11 | 0.83 | 1.00-4.70 | 7.05 | 4.74 | 0.12-19.07 |

Table 2: Means, standard deviations and ranges for the two pre-studies that collected plausibility ratings and surprisal values for target and distractor words.

To assess the expectancy of target and distractor words across conditions, the GPT-2 language model was used to calculate surprisal values. Since surprisal is inversely proportional to expectancy, higher surprisal values are associated with a lower expectancy of a word in its context. Especially in condition B, where the goal was to remove the ambiguity by changing the main verb in such a way that the distractor word no longer is an attractive alternative to the target word, a lower average surprisal value in the target condition than in the distractor condition should prove that the ambiguity of the main verb with respect to the target word has been removed. Hence, the distractor words in the context of the final sentences and preceding paragraphs should no longer represent a semantically attractive alternative to the target words in conditions B.

For the calculation of the surprisal values, the Hugging Face Transformer library (Wolf et al., 2020) was used to deploy the German version of GPT-2 (Schweter, 2020). GPT-2 is a transformer-based model, which is is designed for efficient processing of sequential data. The sentence materials used as input to the model are identical to those in the plausibility rating study, involving target and distractor words in conditions A, B, C for 60 items, amounting to 360 different items in total. However, the items were truncated after the target/distractor word, omitting the continuation of the final sentences since the focus is only on the surprisal values associated with the target/distractor words. No filler items were used because the values were ultimately calculated by a language model instead of humans.

The results show that the average surprisal value for the target words in each condition is lower than the average surprisal value for the distractor words in the corresponding condition. This shows that the target words across all conditions have on average a higher expectancy than the distractor words. In particular, this also means that the item manipulation was successful and distractor words in condition B no longer represent a semantically more attractive alternative to target words. The resulting surprisal values for target and distractor in conditions A, B and C are shown in Table 2.

Correlation between plausibility and surprisal for target words and distractor words is not strong for any combination. The highest correlation coefficient observed is -0.36 between target plausibility and target surprisal, which shows that overall the pattern of target surprisal corresponds to that of target plausibility and thus that the plausibility manipulation was successful. However, target surprisal didn't enter the analysis. Instead, the analysis will focus on target word plausibility and distractor word suprisal. Crucially, the correlation between these two is nearly zero ($r = 0.01$), indicating an independent relationship between these predictors.

## 5.2 Self-Paced Reading Study

The first main experiment was a web-based self-paced reading study, some aspects of which are explained below in terms of procedure, analysis and results.

### 5.2.1 Procedure

Forty-five participants were recruited from Prolific Academic Ltd. for a web-based self-paced reading study. The data of three participants were excluded from the subsequent statistical analysis due to low accuracy on the comprehension questions ($< 0.70$). The remaining 42 participants (mean age 26.26 years; SD 3.7; age range 19-32 years; 17 male, 25 female) were all native German speakers and had no language-related disorders or literacy difficulty. All participants consented to the study by means of a consent form and received a payment of £9.00 for their participation. Participants were assigned to six different lists and were presented with three practice items before starting the main experiment. Subsequently, each participant saw 105 items distributed across three blocks, i.e. 35 items per block, of which 20 were critical items and 15 were filler items. In about half of the trials, participants were presented with comprehension questions that they could answer with *Yes* or *No* (assigned to the keys *D* and *K*). Based on their answers, participants were shown an accuracy score at the end of each block to encourage attentive reading. In addition, the participants were asked to rate the plausibility of the sentences on a scale from 1 to 7, analogous to the plausibility rating study described in 4.3.

### 5.2.2 Analysis

If reading times on one of the critical regions, i.e. the pre-critical region (the article "the" which is preceeding the target word), the critical region (the target word "tourist"), the spillover region ("and") or the post-spillover region ("afterwards") were lower than 50 ms or higher than 2500 ms and if reaction time on the task (comprehension question) was lower than 50 ms or higher than 10,000 ms, all trials for this item of the corresponding subject were excluded. As participants were also asked to provide plausibility ratings in each trial, not only the reading times, but also the plausibility ratings of the corresponding items were excluded. Based on this criterion, 7 out of 2520 trials were excluded (= 0.28%) from the subsequent statistical analysis. All results reported below are obtained after removing the outliers. The statistical analysis is based on a linear mixed-effects regression re-estimation technique, which is described in 4.2.

### 5.2.3 Results

Due to the limited scope of the proposal, the following sections provide only a rough summary of the most important findings so far.

**Comprehension Questions** On 46% of the trials, i.e. half of the critical trials and two-fifths of the fillers, participants were asked to answer comprehension questions that could relate to both the contextual paragraph and the final sentence, with *Yes* or *No*. Descriptive measures were calculated for accuracy and reaction times across subjects: Average accuracy was 95.2% (SD = 5.5, range = 80.0%-100.0%) and average reaction time was 2929 ms (SD = 627.3, range = 1758ms-4473ms).

**Plausibility Ratings** In addition to the average plausibility ratings collected in a pre-study, participants were also asked to provide plausibility ratings for the sentences they read as part of the self-paced reading study. As reading times were not measured for the distractor words, only plausibility ratings for the target words were collected. Both the per-trial and average plausibility ratings, are used and compared as a predictor of reading times in the subsequent statistical analysis. Table 3 shows the average plausibility ratings per condition for both per-trial ratings (a), obtained in the self-paced reading study and average plausibility ratings (b), obtained in a pre-study. Like the plausibility ratings during the pre-study, the plausibility ratings collected as part of the self-paced reading study show that plausibility is stepped across the three conditions ($A < B < C$), with the average values

for the conditions being slightly closer to each other. This confirms that plausibility was successfully manipulated to be graded across three levels and that the expectancy of distractor words in Condition B was lowered.

| | | **Plausibility (a)** | | | **Plausibility (b)** | | |
|---|---|---|---|---|---|---|---|
| | Cond. | Mean | SD | Range | Mean | SD | Range |
| Target | A | 5.83 | 0.78 | 4.00-6.93 | 6.03 | 0.71 | 4.40-7.00 |
| | B | 3.92 | 1.04 | 1.79-6.71 | 3.79 | 1.20 | 1.70-6.80 |
| | C | 2.20 | 0.68 | 1.07-4.29 | 1.91 | 0.57 | 1.00-3.30 |

Table 3: Means, standard deviations and ranges for the plausibility ratings collected in the self-paced reading study (a) and the plausibility ratings collected in the pre-study (b).

**Reading Times** Figure 1 illustrates the average reading times for each condition and critical region. Although the anticipated reading time pattern $A < B < C$ can be seen across all regions, the average reading times for condition B and C are much more similar compared to condition A, particularly in the post-spillover region. The graded pattern is most pronounced in the spillover region.

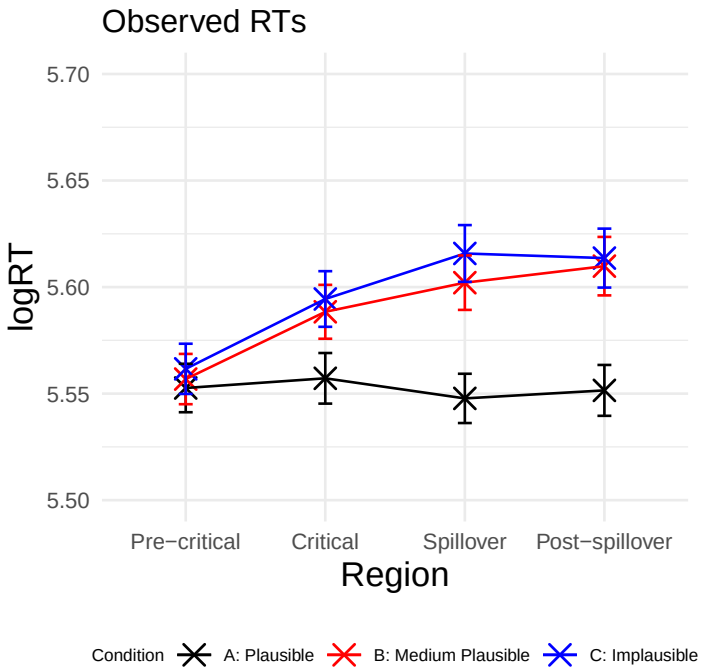


Figure 1: Log reading times per condition on the Pre-critical, Critical, Spillover, and Post-spillover region

Subsequently, a separate model with target plausibility, obtained from the self-paced reading study, and distractor surprisal as predictors was fit for each critical region. These models were then used to predict the average reading times for conditions A, B and C per region. Figure 2 shows the estimated reading times of these models and Figure 3 the residual error. Apparently, the model captures the overall structure of the observed reading times in the sense that the graded pattern ($A < B < C$) is visible across all regions, especially in the spillover region. However, the visible differences between the observed and estimated reading times, which are also reflected in the residual error plot (figure 3), indicate that the model did not accurately capture all patterns in the data.

For comparison, the same models were fitted with the average plausibility ratings from the pre-study as predictor instead of the per-trial plausibility ratings. Due to the limited scope of this proposal, no visualisations are included, but the more accurate predictions and lower residuals seem to indicate that the average plausibility ratings are a better predictor. In order to find out whether average plausibility is actually a better predictor of reading times or whether the ratings collected in the pre-study study only capture reading times better by chance, the same model was fit, but this time using average plausibility ratings from the self-paced reading study. It appears that the accuracy of the predictions is roughly between the average ratings from the pre-study and the individual ratings from the self-paced reading study, indicating that the average ratings from the pre-study, and thus average ratings in general are a better predictor of reading times.

The coefficients for target plausibility (per trial and average) are positive across all regions, indicating that lower plausibility predicts slower reading. In contrast, the coefficient for distractor surprisal is negative across all regions, which means that lower surprisal predicts higher reading times. The reason for this might be that the distractor surprisal values calculated by the language model were almost identical across the three conditions and even higher in condition A than in condition B, in contrast to the graded pattern calculated for target surprisal ($C > B > A$). The p-values show that per-trial plausibility is a significant predictor of reading times in the spillover and post-spillover regions, but not in the pre-critical and critical regions, whereas average plausibility is a significant predictor in the critical, spillover and post-spillover regions. Although the z-values for distractor surprisal are larger than the z-values for plausibility in the region where plausibility is not a significant predictor (pre-critical and pre-critical/critical), distractor surprisal is not a significant predictor in any region. This shows that no significant modulation of reading times can be attributed to distractor surprisal, i.e. distractor words are no longer an attractive alternative in condition B. However, reading times might just not be sensitive to unfulfilled expectations in general, as hypothesized by Aurnhammer et al. (2023) who found no significant distractor cloze modulation of reading times despite the presence of an attractive alternative.

**Model comparison** A likelihood ratio test was performed to compare the model fit with the per-trial plausibility ratings to the model fit with the average plausibility ratings objectively. For this purpose, a model containing only the average plausibility ratings from the pre-study (and distractor surprisal) as a predictor (simple model) was fitted and compared to a model containing both the average plausibility ratings from the pre-study and the per-trial plausibility ratings as predictors (complex model). The obtained chi-squared values and p-values (in all regions $< 0.05$) indicate that the complex model provides a better fit to the data than the simple model. However, when using the complex model to predict reading times per region and condition, the model fit is slightly worse than the fit of the simple model. This shows that the complex model improves the overall model fit to the data (for each region), but it does not improve the model's ability to make predictions about the reading times per region. Instead, the simple model predicts the reading time data per condition more accurately than the complex model. Apparently, the per-trial plausibility predictor can explain additional variability in reading times, which is reflected in an enhanced model fit when using the complex model (containing per-trial and average plausibility ratings) compared to the simple model. However, per-trial plausibility does not appear to adequately capture the reading time differences between conditions, resulting in slightly worse predictions of reading times for the three conditions.

## 5.3 Electroencephalography

Since reading times were graded for target plausibility, reflecting graded integration difficulty, the next step would be to conduct an EEG study to investigate whether this graded integration difficulty will be reflected in a graded N400 or P600 effect and to test the different interpretations of multi-stream models and RI theory of ERP components as an index of integrative processes. According to multi-stream models, an EEG study should reveal a graded N400 effect, while RI theory predicts a graded P600 effect. This would require finding around 30 participants who are right-handed, native German speakers and have normal or corrected-to-normal vision. The materials that will be used are the same

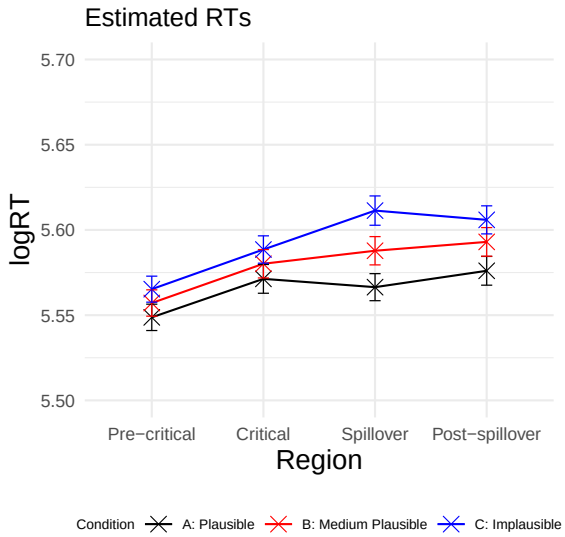


Figure 2: Estimated log-Reading Times per condition on the four regions

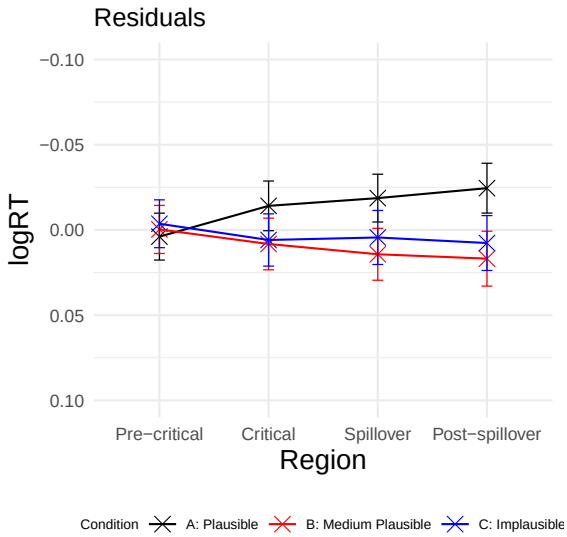


Figure 3: Residual errors per condition on the four regions

as in the self-paced reading study, i.e. each person reads a total of 105 items, divided into three blocks of 35 items each. Immediately after completing the EEG study, participants will be asked to read the items again and rate them as in the other plausibility rating studies. The obtained target word plausibility ratings will be used as predictor in the subsequent analysis. Similar to the self-paced reading study, both the per-trial plausibility ratings from the EEG study and the average plausibility ratings from the pre-study will be used to fit linear regression models and determine which one is a better predictor. The $r$ERP method described in 4.2 will be used for statistical analysis.

# 6 Conclusion

Multi-stream models and RI theory make divergent predictions about which cognitive processes the language-sensitive ERP components, the N400 and the P600, index. Aurnhammer et al. (2023) found a graded P600 effect and thus support for the predictions of RI theory. One goal of this thesis is to further test the hypotheses of the two approaches based on the design of Aurnhammer et al. (2023). For this purpose the context manipulation design has been modified in such a way that graded plausibility ($A > B > C$) is preserved, but the semantically attractive alternative, which was previously present in one of the conditions (B), has been eliminated. Based on this current design, multi-stream models and RI theory diverge even further in their predictions: multi-stream models predict a graded N400 effect due to the irreparable anomaly in two conditions relative to the baseline, whereas RI theory predicts a graded P600 effect. As the EEG study has not yet been conducted, no conclusions can be drawn regarding the research question of whether, and if so, which ERP effect is modulated in the ERP signal. Similarly, the question of whether the early negativity observed for the ambiguous condition in Aurnhammer et al. (2023) is sensitive to unfulfilled expectations and disappears after removing the attractive alternative cannot yet be answered. One question that can be answered instead relates to whether reading times pattern with the three levels of plausibility. It appears that reading times pattern with plausibility in the sense that higher plausibility is associated with higher reading times on average in all regions, i.e. the highest reading times are observed for condition C, then B and the lowest reading times for condition A in all regions. On the other hand, the pattern does not really seem to be evenly graded, as the reading times of condition B are quite slow, similar to the reading times of condition C. Whether this pattern is also reflected in a graded amplitude for one of

the ERP components remains to be seen. Another aim was to investigate whether average plausibility ratings, collected in a pre-study, or per-trial plausibility ratings, collected during the self-paced reading study, are a better predictor of reading times and ERP amplitude. Statistical analysis showed that the average plausibility ratings accounted for more variance in reading times and appeared to be a better predictor than per-trial plausibility ratings. A subsequent likelihood ratio test revealed that per-trial plausibility accounts for additional variability in reading times, which, however, does not capture the differences between conditions. These findings can be further validated by comparing the average and per-trial ratings which will be obtained in the EEG study.

# 7 Workplan

The two preliminary studies (plausibility study and calculation of surprisal values) have already been conducted. Also, the reading time study as well as the corresponding statistical analysis have been completed. In addition, if possible, an EEG study should be conducted, preferably starting in January, in order to find participants before the semester break. Otherwise, it is difficult to set an exact timetable, as this depends on whether I carry out EEG study, and if so, how quickly I can find participants. I don't have any strict deadlines regarding the submission date, but I would still like to register the Master's thesis this month and then submit it approximately in June.

# References


Aurnhammer, C., Delogu, F., Brouwer, H., and Crocker, M. W. (2023). The p600 as a continuous index of integration effort. *Psychophysiology*, 60(1-28):1236–1239.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., and Crocker, M. W. (2021). Retrieval (n400) and integration (p600) in expectation-based comprehension. *PLOS ONE*, 16(9):1–31.

Berkum, J. J. V., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443–467.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2008). An alternative perspective on "semantic p600 effects in language comprehension. *Brain Research Reviews*, 59(1):55–73.

Brouwer, H. and Crocker, M. W. (2017). On the proper treatment of the n400 and p600 in language comprehension. *Frontiers in Psychology*, 8(1327).

Brouwer, H., Delogu, F., and Crocker, M. W. (2021a). Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *European Journal of Neuroscience*, 53:974–995.

Brouwer, H., Delogu, F., Venhuizen, N. J., and Crocker, M. W. (2021b). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12(615538).

Brouwer, H., Fritz, H., and Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the p600 in language comprehension. *Brain Research*, 1446:127–143.

Brown, C. and Hagoort, P. (1993). The processing nature of the n400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, 5(1):34–44.

Delogu, F., Brouwer, H., and Crocker, M. W. (2019). Event-related potentials index lexical retrieval (n400) and integration (p600) during language comprehension. *Brain and Cognition*, 135(103569).

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

Frank, S. and Willems, R. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Cognition and Neuroscience*, 32(9):1192–1203.

Gouvea, A., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the p600. *Language, Cognition and Neuroscience*, 25(2):149–188.

Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, 8(4):439–483.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of NAACL*, 2.

Hoeks, J. C. J., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1):59–73.

Kaan, E. and Swaab, T. (2003). Repair, revision, and complexity in syntactic analysis: an electrophysiological differentiation. *Journal of Cognitive Neuroscience*, 15(1):98–110.

Kim, A. and Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Languag*, 52(2):205–225.

Kupferberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49.

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

Kutas, M. and Federmeier, K. D. (2009). N400. *Scholarpedia*, 4(10):7790.

Kutas, M. and 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.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785–806.

Petten, C. V. and Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In Simpson, G. B., editor, *Understanding Word and Sentence*, pages 129–174. North-Holland Press.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilis- tic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Rabovsky, M. and McClelland, J. L. (2019). Quasi-compositional mapping from form to meaning: A neural network- based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190313.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, 22(6):642–647.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, 18(4):367–379.

Schweter, S. (2020). German gpt-2 model (version 1.0.0). *Zenodo*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):378–423.

Smith, N. J. and Kutas, M. (2015). Regression-based estimation of erp waveforms: I. the rerp framework. *Psychophysiology*, 52(2):157–168.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Troyer, M. and Kutas, M. (2018). Harry potter and the chamber of *What?*: The impact of what individuals know on word processing during reading. *Language, Cognition and Neuroscience*, 35(5):641–657.

Troyer, M. and Kutas, M. (2020). To catch a snitch: Brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. *Journal of Memory and Language*, 113(2):104111.

Troyer, M., Urbach, T., and Kutas, M. (2019). Lumos!: Electrophysiological tracking of (wizarding) world knowledge use during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3):476–486.

Venhuizen, N., Crocker, M. W., and Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discource Processes*, 56:229–255.

Wolf, T., Debut, L., , Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davidson, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., and Rush, S. G. . A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing.

Zehr, J. and Schwarz, F. (2018). Penncontroller for internet based experiments (ibex).