SAARLAND UNIVERSITY

FACULTY OF HUMANITIES

DEPARTMENT OF LANGUAGE SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Exploring the Impact of Offline and Online Plausibility Judgements on Reading Times

Author:
Eva RICHTER

Supervisors:
Prof. Dr. Matthew W. CROCKER
Dr. Francesca DELOGU

Advisor:
Dr. Christoph AURNHAMMER

August 14, 2024

UNIVERSITÄT
DES
SAARLANDES

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich versichere, dass die gedruckte und die elektronische Version der Masterarbeit inhaltlich übereinstimmen.

Statutory Declaration

I hereby declare that the thesis presented here is my own work and that no other sources or aids, other than those listed, have been used. I assure that the electronic version is identical in content to the printed version of the Master's thesis.

Signed:

Date: August 14, 2024

Abstract

Research on language comprehension often involves the manipulation of plausibility to investigate how individuals process and interpret language. Since different individuals vary in their knowledge and experiences, they may perceive the plausibility of a given situation or sentence differently. However, plausibility is usually assessed through offline pre-tests, in which plausibility ratings are collected and averaged over a group of participants. As a result, these aggregated ratings do not capture individual differences in how plausibility is perceived.

Assuming that single-trial plausibility ratings provide a more accurate estimate of each individual's perceived plausibility, the main goal of the present work is to assess whether single-trial plausibility ratings collected during an online experiment are a better predictor of processing effort than plausibility ratings collected offline. To explore this question, a self-paced reading study was conducted based on a context manipulation design developed by Aurnhammer et al. (2023), in which plausibility is varied across three levels. In addition to collecting offline plausibility ratings from a separate group of participants, the participants in the self-paced reading study were asked to provide plausibility ratings on each trial. Subsequently, the reading time (RT) data were re-estimated using a regression-based technique to compare the RT predictions of offline and online plausibility ratings.

Although the self-paced reading study revealed graded RTs for plausibility, the RTs for the medium plausible and implausible conditions were nearly identical. Post-hoc analyses showed that items rated as medium plausible, regardless of their condition, had higher RTs than those rated as plausible or implausible, suggesting that the rating task affected the RTs. This assumption was confirmed by a second self-paced reading study without a rating task, which revealed a more pronounced RT pattern. The regression-based analysis from the first self-paced reading study indicated that the offline plausibility ratings better captured the effects structure in the observed RT data compared to the plausibility ratings collected on each trial. Since the averaged single-trial ratings also yielded more accurate RT predictions than the single-trial ratings themselves, this suggests that the predictive power of offline plausibility ratings is higher due to their relative stability, irrespective of the effect of the rating task on the RTs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Psycholinguistic research on language comprehension investigates the processes involved in understanding spoken and written language. While behavioural measures such as reading times (RTs) provide an overall index of word-by-word processing effort, neural methods such as electroencephalography (EEG) aim to reveal the mechanisms driving this effort. According to Retrieval-Integration (RI) theory (Brouwer et al., 2012, 2017), the N400 component of the Event-Related brain Potentials (ERP) signal reflects the retrieval of word meaning from long-term memory, while the P600 ERP component is linked to the integration of this meaning into the unfolding utterance interpretation. Under this account, both the N400 and the P600 are sensitive to word expectancy, whereas only the P600 is sensitive to plausibility. The retrieval of implausible words is facilitated when they are contextually or lexically primed, while their integration into the unfolding utterance interpretation always requires higher cognitive effort, reflected in increased RTs and P600 amplitude. Aurnhammer et al. (2023) showed that this relationship between plausibility, RTs and P600 amplitude holds empirically and quantitatively, establishing the P600 (and RTs) as a continuous index of integration effort.

Studies manipulating plausibility, such as the one by Aurnhammer et al. (2023), usually rely on offline measures of plausibility that are averaged over a group of participants. These aggregated ratings, however, do not capture individual differences in how plausibility is perceived, which may vary based on each individual's world knowledge and experiences (Venhuizen et al., 2019) and be influenced by cultural (Hagoort et al., 2004) and other individual factors. Furthermore, offline plausibility ratings cannot be directly linked to individual RTs or ERPs, as they are usually obtained from a different group of participants than the measures of processing effort. In contrast, single-trial plausibility ratings collected from the same participants whose processing effort is being measured can capture each individual's perception of plausibility and can be related to their processing effort on each trial, potentially explaining the observed RT or ERP data better. Therefore, the current work aims to determine whether single-trial plausibility ratings collected during an online experiment can account for more variability in individual processing effort than offline plausibility ratings.

In order to address this question, a self-paced reading study is conducted based on a context manipulation design adapted from Aurnhammer et al. (2023), which manipulates plausibility in a graded manner. To compare the RT predictions of averaged (offline) and single-trial (online) plausibility ratings, plausibility ratings are collected from a separate group of participants in a pre-test and from the participants of the self-paced reading study, who are asked to provide a plausibility rating on each trial. Subsequently, both the offline and online plausibility ratings are used in a regression-based analysis to model RTs as a function of plausibility, allowing for a comparison between their RT predictions. Since the current study uses a modified version of the stimuli from Aurnhammer et al. (2023), while aiming to maintain graded plausibility, RTs are expected to show a graded pattern similar to that observed by Aurnhammer et al. (2023). Furthermore, the modified context manipulation design used in this thesis offers the opportunity to build on the study by Aurnhammer et al. (2023) and further test the predictions of multi-stream models and RI theory.

Chapter 2 describes related concepts and studies, while Chapter 3 lays out the research questions; Chapter 4 discusses the materials and methodology and Chapter 5 describes the procedures and results of the pre-tests; similarly, Chapters 6 and 7 describe procedures, results and discussions of the two self-paced reading studies, while Chapter 8 provides a general discussion; Chapter 9 concludes.

Chapter 2

Background and Related Work

In order to compare the RT predictions of offline and online plausibility ratings, the stimuli developed by Aurnhammer et al. (2023) are adapted in terms of expectancy and plausibility for this work. Therefore, the following sections provide a brief overview of theories of language comprehension and discuss the concepts and relevant literature on expectancy, plausibility, and individual variation.

2.1 Theories of Language Comprehension

Language comprehension in the brain can be studied through behavioural and electrophysiological measures. While behavioural measures such as RTs are relatively easy to collect, they only provide an approximation of the overall processing effort. In contrast, electrophysiological measures, specifically ERPs, are more time-consuming and costly to collect, but offer high temporal resolution and information about the cognitive processes that underlie language comprehension.

The two most prominent ERP components studied in the context of human language comprehension are the N400, a negative-going voltage deflection peaking around 400 ms after stimulus onset, and the P600, a positive-going deflection emerging around 500-600 ms after stimulus presentation. Traditionally, the N400 has been interpreted as an index of semantic integration processes (Brown and Hagoort, 1993), while the P600 has been linked to syntactic processing. However, these interpretations are incompatible with several studies that have found P600 effects in syntactically correct but semantically anomalous sentences (Kim and Osterhout, 2005; Hoeks et al., 2004; Nieuwland and van Berkum, 2005), which sparked controversy about the interpretation of the language-sensitive ERP components and resulted in multi-stream models (Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012, for reviews) and Retrieval-Integration (RI) theory (Brouwer et al., 2012, 2017), a single-stream account.

While multi-stream models adhere to the traditional interpretations of the ERP components, RI theory proposes that the N400 indexes the retrieval of a word's meaning from long-term memory. The more consistent the features of an incoming word are with the features pre-activated in memory, the easier the lexical retrieval of the word, which is reflected in a reduced N400 amplitude. In contrast, the P600 component is associated with the semantic integration of an incoming word into the unfolding utterance representation. The less the current interpretation of the unfolding utterance representation needs to be revised or reorganised due to the syntactic, semantic, and pragmatic information associated with an incoming word in order to become coherent, the lower the integration effort and thus the P600 amplitude (Brouwer et al., 2012, 2017). Importantly, according to this account, both the N400 and P600 components are sensitive to expectancy, as unexpected words are more difficult to retrieve and to integrate, which is supported by the findings of Aurnhammer et al. (2021). However, the retrieval of unexpected words is facilitated and N400 amplitude reduced when a word is sufficiently primed by the context or by lexical repetition. In contrast, only the P600 component shows sensitivity to plausibility: Words that render a sentence implausible, for example, by contradicting world knowledge lead to an increased P600 amplitude, even if the sentence is grammatically well-formed. However, the N400 component is indirectly sensitive to plausibility, as words that render a sentence implausible are usually also unexpected. Consequently, no N400 effect is observed for an implausible word relative to a plausible word when both are approximately equally primed by the preceding context or by lexical repetition (Brouwer et al., 2012, 2017).

To test the contrasting hypotheses of multi-stream models and the RI account, Aurnhammer et al. (2023) modified and extended the design of Nieuwland and van Berkum (2005) to a context manipulation design, in which a context paragraph is followed by a final sentence (see Figure 2.1). Since multi-stream models predict a P600 effect only if a semantic anomaly is repairable by the presence of a semantically attractive alternative, a semantically attractive alternative was made globally available in Condition B by raising the expectancy for a distractor word, which was never presented in target position ("the lady *weighed*" attracting the distractor word "suitcase" rather than the actually presented target word "tourist"). In contrast, in Conditions A and C no semantically attractive alternative was available, i.e. the expectancy of the target word was higher than the expectancy of the distractor word. Both target and distractor words were approximately equally primed by lexical repetition in the context paragraph. To test for graded integration difficulty, the main verb of the final sentence was varied across three levels, rendering the target word either plausible (Condition A, "the lady *dismissed* the tourist"), medium plausible (Condition B, "the lady *weighed* the tourist"), or implausible (Condition C, "the lady *signed* the tourist"), allowing to test for graded integration difficulty.

Aurnhammer et al. (2023) observed that both RTs and P600 amplitude pattern with the three plausibility levels of Conditions A, B and C, reflecting continuous integration effort. At the same time, no N400 effect was observed for the implausible Condition (Condition C) in the absence of a semantically attractive alternative. Despite relying on different premises, multi-stream models and RI theory both can account for the P600 effect observed in Condition B. According to the RI account, no N400 effect was observed due to the lexical repetition of the target (and distractor) word, which facilitates retrieval. Instead, a P600 effect was observed, which can be explained by the increased difficulty of integrating the meaning of the medium plausible target word into the unfolding utterance interpretation. Similarly, the increased P600 amplitude in Condition C is explained by the even less plausible target word, resulting in increased integration effort. Multi-stream models can explain the P600 effect in Condition B by the presence of a semantically attractive alternative that repairs the anomaly. However, the observed P600 amplitude in Condition C challenges the assumptions of multi-stream models, since no semantically attractive alternative was available. These findings provide evidence for RI theory since they (1) suggest a strong correlation between plausibility, RTs, and

P600 amplitude and (2) establish the P600 as a continuous, word-by-word index of integration effort.

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 tourist an extra fee. After that, the tourist opened his **suitcase** and threw several things out. Now, the **suitcase** of the ingenious tourist weighed less than the maximum of 30 kilograms.*

Condition A: Plausible, no attraction
Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.
Then dismissed the lady the tourist and afterwards he went to the gate.

Condition B: Less plausible, attraction
Dann wog die Dame den Touristen und danach ging er zum Gate.
Then weighed the lady the tourist and afterwards he went to the gate.

Condition C: Implausible, attraction
Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.
Then sign the lady the tourist and afterwards he went to the gate.

FIGURE 2.1: Experimental design of the study by Aurnhammer et al. (2023). The German word order is preserved for the English transliterations of the final sentences. Target words are underlined, and distractor words are highlighted in boldface.

2.2 Expectancy

Readers or listeners process language continuously and incrementally, more or less word by word (Tanenhaus et al., 1995). This involves the integration of syntactic, semantic and pragmatic information together with world knowledge (Hagoort et al., 2004) to construct an interpretation that reflects the meaning of the utterance. This interpretation leads to general expectations or predictions about upcoming words. The ease or difficulty of integrating the encountered words with the preceding context depends on the degree to which they align with these expectations (Kuperberg et al., 2020). Although the functions and even the names[1] referring to the concept of expectancy in language comprehension are not uncontroversial (see Van Petten and Luka, 2012; Kuperberg and Jaeger, 2016, for a discussion), both behavioural studies (Schwanenflugel and Shoben, 1985; Stanovich and West, 1983; Smith and Levy, 2013), showing slower RTs and ERP studies (Kutas and Hillyard, 1980, 1984), showing increased N400 amplitudes for unexpected words compared to expected words, provide evidence that unexpected words require more processing effort than expected words.

The most widely used method for assessing word expectancy based on human judgements is the cloze task (Taylor, 1953), in which participants are asked to fill in missing words given the context of the preceding sentence or text. The cloze probability of a word can be calculated by dividing the number of participants who provided the same completion for a given gap by the total number of participants. Thus, a higher cloze probability indicates that a higher proportion of participants provided the same word as a completion, reflecting a higher expectancy of that word in a given context. Previous studies that have used cloze probability to quantify expectancy have shown an inverse relationship between cognitive measures and cloze probability, i.e., faster RTs (e.g., Brothers and Kuperberg, 2021; Aurnhammer et al., 2021) and reduced N400 amplitudes (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011) for words with higher cloze probabilities.

Under information-theoretic accounts of language comprehension, the processing difficulty induced by a word can be measured in terms of surprisal, which corresponds to the negative logarithm of cloze probability. Surprisal estimates are computed by language models (LMs) and differ mainly from cloze probabilities in that their predictions capture only the statistics of the language, but not explicitly the meaning of a word with respect to world knowledge. However, Michaelov et al. (2022) point out that precisely for this reason LM surprisal is a more accurate method for assessing the extent to which linguistic input alone can predict measures of language comprehension because it isolates the specific effects of the linguistic input. Moreover, cloze probabilities are time-consuming and costly to collect, and provide unreliable estimates for low-probability words, since large sample sizes are needed to ensure that less expected continuations are produced at least once, i.e., don't result in a zero cloze probability despite not being unexpected continuations (Smith and Levy, 2013). In contrast, surprisal estimates generate probability distributions over

[1] For example Van Petten and Luka (2012) distinguish between the terms *expectation*, *prediction* and *anticipation*. However, in this thesis *prediction* and *expectation* or *expectancy* are used interchangeably.

the entire vocabulary, capturing high and low probability words to the same extent. Since in this thesis surprisal estimates are used to predict RTs, only surprisal will be discussed in more detail in the following section.

2.2.1 Surprisal

Surprisal theory is an expectation-based processing theory that draws on principles from information theory (Shannon, 1948) and has proven effective in explaining word-by-word processing difficulties (Hale, 2001; Levy, 2008). It is based on the assumption that each word carries a certain amount of information, which is predictive of the cognitive effort required to process the word. The processing effort is proportional to the surprisal of the word, which in turn is inversely proportional to the expectancy of a word. More formally, given a sequence of words $w_1, ..., w_t$ the surprisal of the upcoming word w_{t+1} is defined as the negative logarithm of the probability of the upcoming word given the preceding context:

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

While the amount of information carried by each word can be estimated from LMs, the amount of cognitive effort that is required to process a word can be observed through behavioural (Frank et al., 2015) and neural measures (Hale et al., 2018; Shain et al., 2020). In particular, RTs have been shown to be positively correlated with word surprisal (Monsalve et al., 2012; Smith and Levy, 2013; Roark et al., 2009; Fossum and Levy, 2012). Accordingly, words that carry more information, as indicated by higher surprisal values, are read more slowly than less informative words. Importantly, Monsalve et al. (2012) found that significant RT effects may be missed if spillover regions are not taken into account, i.e., if surprisal is only analysed in relation to the current item without considering its influence on the following item. In electrophysiological research, surprisal has been found to be predictive of N400 amplitude during reading (Frank et al., 2015; Aurnhammer and Frank, 2019; Merkx and Frank, 2021; Michaelov and Bergen, 2020), while the P600 ERP component has been less frequently studied in this context (De Varda et al., 2023; Krieger et al., 2024). A recent study by Krieger et al. (2024) found that LM surprisal can account for P600 effects elicited by violations of selectional restrictions, but fails to capture P600 effects from script knowledge violations and graded P600 modulations, which raises concerns about the reliability of LM surprisal in accurately representing the underlying mechanisms of the N400 and P600.

Finally, the accuracy of LMs in predicting RT or EEG data also depends on the architecture of the LM itself. Over time, several architectures for estimating surprisal values have been developed and studied, including Probabilistic Context-Free Grammars (PCFGs), N-gram models, and Simple Recurrent Networks (SRNs). The more recently developed Transformer model (Vaswani et al., 2017) has outperformed previous LMs in several NLP tasks and is increasingly being investigated as a model for human sentence processing (Ettinger, 2020; Wilcox et al., 2020; Merkx and Frank, 2021; Michaelov et al., 2021). Larger, more sophisticated Transformer-based LMs have been found to be more predictive of comprehension difficulty in terms of RTs (Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Wilcox et al., 2020) as well as both N400 (Michaelov et al., 2021, 2023; De Varda et al., 2023) and P600 (De Varda et al., 2023) amplitude.

However, contrasting results have been reported specifically for RTs. For example, Oh and Schuler (2022) found that surprisal estimates derived from variants of the pre-trained GPT-2 LM with more parameters and lower perplexity were less predictive of RTs obtained from naturalistic reading of texts. Further investigation showed that the degree of underprediction increases with model size, particularly for open-class words (Oh and Schuler, 2023), and that this relationship is most pronounced for the subset of the least frequent words (Oh et al., 2024). Consequently, Oh and Schuler (2023); Oh et al. (2024) argue that large LMs may be less suitable for cognitive modelling since they are trained with non-human learning goals on vast amounts of text that are not accessible to humans.

The operationalisation of expectancy as cloze probability has previously been found to be a better predictor of RTs than LM-derived surprisal estimates (Smith and Levy, 2011; Brothers and Kuperberg, 2021). More recently, however, sophisticated Transformer-based LMs such as GPT-3 have been found to be better predictors of RT and N400 data than cloze probabilities (Hofmann et al., 2022; Michaelov et al., 2022, 2023). However, the results of Michaelov et al. (2022) are not inconsistent with previous studies in which cloze probability was found to be a better predictor of processing difficulty, as they only found surprisal estimates of the more recent Transformer-based LMs to be a better predictor of processing difficulty, not the RNN-based surprisal estimates used in earlier work Smith and Levy (2011); Brothers and Kuperberg (2021). Furthermore, contrasting results between studies can often be attributed to differences in experimental design, stimuli, and participant demographics. It can therefore be concluded that there are different ways of operationalising the expectancy of a word in context, each of which has its strengths and weaknesses (Smith and Levy, 2011; Shain et al., 2020; Michaelov et al., 2022).

Regardless of their architecture, the LMs discussed in the previous sections all differ from human-derived operationalisations of expectancy, such as cloze probabilities, in that they do not reflect extralinguistic factors such as world knowledge, which have been shown to influence language processing beyond linguistic experience (e.g., Hagoort et al., 2004; Nieuwland et al., 2007). One exception is the model of language comprehension of Venhuizen et al. (2019), which instantiates a *comprehension-centric* notion of surprisal incorporating both linguistic experience and world knowledge by deriving a Distributed Situation-state Space (DSS). While other studies have found a correlation between purely linguistic surprisal and the amplitude of the N400 ERP component (Frank et al., 2015; Michaelov and Bergen, 2020; Michaelov et al., 2022; Merkx and Frank, 2021), *comprehension-centric* surprisal is predicted to reflect P600 amplitude (Venhuizen et al., 2019). In this framework, *comprehension-centric* surprisal reflects the likelihood of transitioning from one point in the situation-state space to the next and the P600 reflects the neurophysiological processing effort that is associated with that transition, i.e., longer transitions correspond to higher surprisal and lead to increased P600 amplitudes, reflecting increased processing effort. The word-by-word estimates from the neurcomputational model of incremental language comprehension of Brouwer et al. (2021) provide evidence that RTs and P600 amplitudes increase in response to implausible compared to plausible target words, supporting a qualitative link between RTs, P600, and *comprehension-centric* surprisal. The results of Aurnhammer et al. (2023) further show that this link holds empirically and quantitatively, as the observed RTs and the P600 - in line with a *comprehension-centric* notion of surprisal - continuously index integration effort.

2.3 Plausibility

Although plausibility[2] is a widely used concept, the effects of which have been studied in many cognitive contexts (Rayner et al., 2004; Warren et al., 2008; Matsuki et al., 2011), the exact nature of plausibility itself remains poorly defined. Instead, most studies describe plausibility in terms of plausibility ratings, which serve as a subjective assessment based on numerical values. However, a common assumption is that plausibility involves concept-coherence in the sense that "some concept, scenario, event or discourse is plausible if it is conceptually consistent with what is known to have occurred in the past" (Connell and Keane, 2004, p. 186). For example, considering the sentence "The bottle rolled off the shelf and smashed on the floor" compared to "The bottle rolled off the shelf and melted on the floor" (Connell and Keane, 2004, p. 186), the former is likely to be judged as more plausible because it is consistent with most people's experience that bottles break rather than melt when dropped. The perceived plausibility of a situation or statement is therefore closely linked to how well it aligns with someone's world knowledge, which is inherently subjective and can vary across cultures (Hagoort et al., 2004) and individuals.

Plausibility is typically operationalised as a rating task, in which participants are presented with experimental items and asked to make a plausibility judgement for each item. Typically, plausibility ratings are collected based on a Likert scale ranging from 1, indicating that a sentence is highly implausible, to 7, indicating that a sentence is highly plausible (Nieuwland et al., 2020; Haeuser and Kray, 2022; Delogu et al., 2021; Aurnhammer et al., 2023; Michaelov et al., 2023). Although Likert scales are among the most widely used response formats for measuring attitudes and opinions in psycholinguistics (and other fields), the assignment of response options such as "very plausible", "less plausible", "implausible" to numerical values is often criticised for implying equal differences between categories simply because the numerical differences are equal, when the actual differences between responses may not be equal (Knapp, 1990). More recently, LMs have been used to generate plausibility ratings on a scale of 1 to 7. Amouyal et al. (2024) found a high correlation between GPT-4 and human plausibility ratings and concluded that LM-generated plausibility ratings are as effective as human ratings for coarse-grained ratings, but less reliable for fine-grained ratings.

Plausibility ratings, similar to cloze responses, are usually collected in an offline task, allowing participants to take as much time as they need to fill in a gap or provide a rating, which is different from actual language comprehension scenarios that occur in real time. This rather unnatural setup allows for "conscious reflection and other strategic effects" (Smith and Levy, 2011, p. 1637) and may distort the responses

[2] Alternatively, terms such as "acceptability", "coherence" or "likelihood" appear in the psycholinguistic literature. While these terms overlap in meaning, they can carry distinct notions depending on the specific research focus.

or ratings. This is problematic because systematic biases in offline responses or ratings can introduce confounds when measuring or controlling for plausibility or expectancy. However, Goodall (2021) point out that the main difference between offline plausibility ratings and online methods such as RTs is the timing – i.e. when the measurements took place – rather than the extent to which the offline responses may be influenced by conscious reflection, as participants tend to move on to a new item within approximately five seconds. Since the offline collected plausibility ratings are averaged over a group of participants, they do not capture individual variation in perceived plausibility. The implications of this are discussed in section 2.4.

One challenge in studying plausibility is to distinguish its effects from those of predictability, since less plausible stimuli are generally less predictable (Nieuwland et al., 2020). For example, in the sentence "The bottle rolled off the shelf and smashed/melted on the floor", the verb "smashed" not only renders the sentence more plausible than "melted", but "smashed" is also a more predictable continuation than "melted". One way of thinking about the different notions of these related concepts is that predictability measures the likelihood of a word occurring at a particular position in the sentence based on the preceding context, whereas plausibility describes the likelihood of a sentence as a whole. As pointed out by Matsuki et al. (2011), the predictability of a critical word is therefore unaffected by any post-target continuation, which is usually included to capture spillover effects in RTs, whereas plausibility is not conditional in nature and can vary depending on the post-target continuation. However, most studies investigate a notion of conditional plausibility by measuring effects on specific words, which makes it difficult to distinguish between plausibility and predictability effects (Matsuki et al., 2011). Nevertheless, the correlation between plausibility and predictability can also be seen as an advantage in some contexts. Michaelov et al. (2023, p. 124) point out that given the limited discriminative capacity of cloze probabilities in the lower range, "plausibility ratings may serve as a proxy for their predictability", allowing for a more accurate distinction between low-probability and very low-probability words.

Given the difficulty of distinguishing between plausibility and predictability effects, most studies have examined their effects separately (Rayner et al., 2004; Warren et al., 2008). However, as Matsuki et al. (2011) point out, the mean cloze probability in these studies is usually not exactly zero and not identical for the plausible and implausible items. Some studies have also attempted to investigate predictability and plausibility effects within a single study by comparing responses to equally unpredictable plausible and implausible words (Haeuser and Kray, 2022; Brothers et al., 2020; DeLong et al., 2014). The results from a self-paced reading study by Haeuser and Kray (2022) showed an early-emerging effect (at the target word) for an unpredictable, medium plausible condition and a later-emerging effect (at the spillover regions) when both plausibility and predictability were violated. These results are consistent with those from ERP studies that found an earlier effect of predictability and a later effect of plausibility either in the N400 time window (Nieuwland et al., 2020) or in later-emerging post-N400 time windows (DeLong et al., 2014; Brothers et al., 2020). This aligns with the predictions of RI theory (Brouwer et al., 2012, 2017) and the findings of Brouwer et al. (2021), which suggest that the N400 component is primarily sensitive to expectancy and the later-emerging P600 component is sensitive to plausibility. According to this account, both RTs and the

P600 respond to plausibility in a graded manner (Brouwer et al., 2012, 2021). This has been confirmed by Aurnhammer et al. (2023), who found that slower RTs and higher P600 amplitudes were associated with less plausible words compared to plausible words, establishing them as continuous indices of integration effort.

2.4 Individual Variation

Most psycholinguistic studies that have investigated individual differences in language processing (see Boudewyn, 2015, for an overview) have examined individual differences in general cognitive processes, such as working memory (Nakano et al., 2010) and cognitive control (Boudewyn et al., 2012) or individual differences related to language proficiency (McLaughlin et al., 2004, 2010), age (Federmeier et al., 2010) or gender (Payne and Lynn, 2011). A study by Troyer and Kutas (2018) investigated how individual differences in domain-specific knowledge influence real-time sentence processing, using participants' knowledge of the Harry Potter (HP) universe for their investigation. They recorded ERPs while participants with different levels of HP knowledge read sentences related to HP and about general topics, ending with either contextually supported or unsupported words. N400 amplitudes were reduced to supported endings for both types of sentences, but varied as a function of participants' HP knowledge only for HP-related sentences, with larger effects observed in more knowledgeable individuals, indicating that N400 context effects vary as a function of individuals' knowledge levels. This study, along with the aforementioned studies, takes into account between-subject variability, which can be useful for identifying meaningful subgroups of participants, but still focuses on average differences and does not account for within-subject variability.

Focusing on average changes in behavioural or neural measures implies that experimental manipulations affecting cognitive processing remain stable throughout the duration of an experiment. According to this approach, any variation or fluctuation that may occur in repeated measurements from the same individual is considered "noise" - random variation that may obscure the true effects of the experimental manipulation (Payne and Federmeier, 2017). In contrast, single-trial analyses, i.e., methods that account for variation across individual trials, allow for the investigation of individual differences by quantifying effects within and between subjects (Pernet et al., 2011). In behavioural psycholinguistics, most studies taking into account trial-level variability have focused on how the effects of experimental manipulations affect the shape of the underlying RT distributions, demonstrating that the language processing system does not always respond consistently to linguistic difficulty across all trials of an experiment (Payne and Federmeier, 2017). Single-trial analyses can reveal effects and interactions that may be obscured in averaged data, for example by allowing for a systematic mapping between subjects' behavioural variability and neural responses. By coregistering self-paced reading times with ERPs, Payne and Federmeier (2017) investigated the extent to which neural indices of sentence processing vary based on trial-by-trial variability in behavioural measures in response to contextual constraints and found that within-subject variability in RTs modulated the degree of contextual facilitation on the N400 component. This indicates that examining the relationship between behavioural and neural responses at the trial level provides insights into the dynamics of sentence comprehension that are obscured by the averaging process.

Moreover, predictor variables can be operationalised at the trial level to obtain a more detailed understanding of how specific factors influence individuals' behavioural or neural responses on each trial. Returning to the study by Troyer and Kutas (2018), participants were categorised based on high or low HP knowledge. However, the study did not take into account which (and how many) facts each individual knew. Therefore, Troyer and Kutas (2018) could only assume that the N400 amplitudes of individuals with higher HP knowledge were reduced for supported words because they were likely to know more facts. To investigate whether the observed pattern was a result of the proportion of facts an individual knew or whether those with greater HP knowledge also knew more facts, and whether this was reflected in a higher proportion of larger versus smaller N400s in their averages, Troyer et al. (2020) conducted a similar experiment, in which participants had to report on each trial whether they had known a fact or not. Based on this single-trial design, Troyer et al. (2020) showed that the proportion of trials that participants knew was highly correlated with their HP domain knowledge and served as a strong predictor. Crucially, the results of Troyer et al. (2020) were consistent with those of Troyer and Kutas (2018), demonstrating that domain knowledge was correlated with a decrease in N400 amplitude for contextually supported sentences. This suggests that the N400 effects were not influenced by task effects based on participant reports. In addition, HP knowledge had the greatest effect when participants did not know a fact, suggesting that domain knowledge especially has an effect particularly when retrieval is difficult.

A third study by Troyer and Kutas (2020) investigated whether individuals with greater domain knowledge make use of richer information when processing incoming words in sentences by employing a related anomaly paradigm with sentences describing HP-sentences ending in (a) contextually supported, (b) related but unsupported or (c) unrelated and unsupported words. In contrast to Troyer et al. (2020), participants' reports of whether a fact was known or unknown were collected offline after the study by presenting participants with the same items again. Single-trial analyses revealed that participants' reports (known/unknown) again influenced the N400 responses to contextually supported words. Specifically, N400s to contextually supported words were lower for individuals with greater HP knowledge (even when they reported not knowing), whereas N400s to unsupported words did not vary based on participants' reports, suggesting that domain knowledge influences the information brought to mind during language processing in a broader sense. The results of both Troyer et al. (2020) and Troyer and Kutas (2020) show that single-trial analyses are useful for establishing a direct link between individual participants' by-trial reports and their neural responses, revealing nuances of how language is processed at the individual level and providing a richer understanding of the underlying mechanisms that would be missed in aggregated data analyses.

Although plausibility is a commonly manipulated variable in many studies, no study has examined the effects of plausibility at the trial level. As discussed in sections 2.2 and 2.3, human judgement tasks used to assess, for example, the expectancy or plausibility of items are typically based on the responses of larger groups of participants and do not necessarily correspond to the perceptions of any individual. Since different participants may respond differently to the same item, the range of variation increases. There is, however, disagreement on whether (and to what extent) this variation is simply noise that should be disregarded by averaging the judgements or if it contains valuable information that can explain individual differences in language comprehension. Featherston (2007) argues that variation in judgements stems primarily from the inherent noise in individuals' judgements rather than from systematic differences (reflecting individual variation in grammars). Thus, comparing individual judgements increases the error variance because each individual introduces their own noise, and the variability in each judgement can differ, even in opposite directions. According to Featherston (2007), the mean judgements of a group of individuals should therefore be considered, as "the errors cancel each other out and the judgements cluster around a mean, which we can take to be the 'underlying' value, free of the noise factor" (Featherston, 2007, p. 284). An alternative perspective is that "variability is structured rather than random" (Foulkes, 2006, p. 654). Verhagen et al. (2019) examined variation not only across participants but also across items, time and methods using seven-point Likert scale or Magnitude Estimation scale judgement data and concluded that variation in metalinguistic judgements is rarely mere noise, but rather an interpretable source of information. They argue that what might be considered noise – such as unnoticed typos or participants assigning random ratings to finish quickly – are actually no real judgements. In contrast, any variation in actual judgements is attributed to characteristics of language use and linguistic representations and constitutes valuable information rather than just noise. However, a major challenge lies in distinguishing between functionally significant variability and noise. Verhagen et al. (2019) point out that taking into account individual variation does not necessarily mean that the unexplained variance in the data is reduced or eliminated, but rather that analysing this variance may reveal meaningful information.

This suggests that the decision of taking into account individual variation in judgement (or any other) data should be guided by the specific goals of the study. If the primary objective is to identify differences in language comprehension and the factors driving these differences, individual judgements may be particularly valuable due to their granularity. However, it remains to be investigated whether single-trial judgements or averaged judgements are a more suitable operationalisation of the predictor variable in terms of minimising unexplained variance in the data. On the one hand, single-trial judgments provide a finer-grained assessment of how each instance of a stimulus is perceived. On the other hand, averaged judgments offer more stable data, as they smooth out random (and systematic) variations.

Chapter 3

Research Questions

Several studies have investigated how plausibility manipulations affect measures of processing effort (Haeuser and Kray, 2022; Matsuki et al., 2011; Brothers et al., 2020; DeLong et al., 2014; Aurnhammer et al., 2023). Crucially, these studies rely on offline plausibility ratings collected from a different group of participants than the measures of processing effort. Since these offline plausibility ratings are averaged across all participants, they do not correspond to each individual's perceived plausibility, which may vary due to cultural, contextual or linguistic factors. In contrast, plausibility ratings collected on a trial-by-trial basis provide an estimate of each individual's perceived plausibility on every trial. This raises the question of how accurately single-trial plausibility ratings, compared to offline plausibility ratings, can predict processing effort measures, such as RTs. Thus, the first research question is:

(1) Are online plausibility ratings collected on each trial during a self-paced reading experiment a better predictor of reading times than offline plausibility ratings?

In previous studies using single-trial analyses, trial-level responses collected during (Troyer et al., 2020) or after (Troyer and Kutas, 2020) EEG experiments were strong predictors of participants' ERPs. Similarly, this thesis aims to collect participants' plausibility ratings on a trial-by-trial basis. However, the main goal is to determine whether this trial-level operationalisation of plausibility yields more accurate estimates of RTs than offline plausibility ratings averaged over a group of participants. Since single-trial plausibility ratings account for individual differences in perceived plausibility and are collected from the participants of the self-paced reading study, they may lead to more accurate RT predictions. At the same time, offline plausibility ratings, which are less influenced by systematic and random variability, provide a more general and stable measure of plausibility, which may enhance their ability to capture RT effects.

To explore this question, the stimuli from the study by Aurnhammer et al. (2023) were adapted for the current study. In the original design, target word plausibility was varied across three levels $(A > B > C)$ and the expectancy of the distractor word was either low (A, C) or high (B). Since the present study aims to investigate only the effects of plausibility, the expectancy of the distractor word in Condition B is reduced, while maintaining graded plausibility, by modifying the main verb in the final sentence. This leads to the second research question:

(2) Are reading times (still) graded for plausibility after modifying the main verb in Condition B to achieve lower distractor word expectancy?

If the main verb in Condition B is successfully manipulated, i.e. if it renders the target word in the final sentence less plausible than in Condition A and more plausible than in Condition C, RTs are expected to be graded, reflecting differential integration effort for varying levels of plausibility.

In addition, two Transformer-based LMs, GPT-2 and LeoLM, which differ primarily in the number of their parameters and the size of their training data, are used to compute surprisal values to assess whether the expectancy of the distractor in Condition B (as well as in Conditions A and C) is lower than the expectancy of the target word. Although this is not related to a specific research question, it may be interesting to test whether and how using surprisal values from different LMs as measures of expectancy affects the RT estimates. Given that Aurnhammer et al. (2023) found no significant RT modulation based on distractor word cloze probability, despite high distractor word expectancy in Condition B, this suggests that distractor word surprisal may not significantly predict the RTs obtained in the current study either. However, RT predictions may differ depending on the choice of expectancy metric (cloze probability or LM surprisal) or, in terms of surprisal, depending on the characteristics of the LM.

Furthermore, the current design provides an opportunity to investigate two additional research questions that build on the work by Aurnhammer et al. (2023) and are discussed in Chapter 8.

Chapter 4

Materials and Methodology

This chapter describes the structure of the stimuli, which are assessed in the pre-tests and used in the self-paced reading experiments, as well as the architecture of the LMs used to compute surprisal values for assessing the expectancy of the target and distractor words. Moreover, it describes the linear mixed effects regression re-estimation technique used to analyse the RT data.

4.1 Experimental Stimuli

The stimuli used in the self-paced reading study are based on the stimuli from Aurnhammer et al. (2023), who developed a total of 96 items by translating and adapting the stimuli from Nieuwland and van Berkum (2005). The 60 best items chosen by Aurnhammer et al. (2023) based on the results of a cloze task were also selected for the current study after slight modification (see Appendix A for the full list of German stimuli).

Each item consists of a context paragraph followed by a final sentence in which the plausibility of the target is manipulated in a graded manner. The context paragraph repeats the target and distractor words three or four times each to prime the meaning of the target word when presented in the target position. Whether the target or distractor word is mentioned last in the context paragraph varies by item and is approximately equally distributed across all items. According to RI theory, priming the target and distractor words should facilitate retrieval and thus no N400 effect should be observed across conditions (Brouwer et al., 2012, 2017), however, in the current work no EEG data was collected. The main verb of the final sentence is varied across three levels, rendering the target word in the given context plausible (Condition A: "the lady *dismissed* the tourist"), medium plausible (Condition B: "the lady *weighed* the tourist") or implausible (Condition C: "the lady *signed* the tourist"). The target word was kept the same across conditions to minimise potential effects due to word length or word frequency. To ensure that the entire main verb can be integrated with the preceding context before reading the target word, no separable verbs were used (e.g., "*Dann* ***bereitete*** *der Mann das Essen* ***zu***"). Furthermore, reflexive verbs were avoided, as they change the position of the target word[1]. Finally, the verbs were chosen in such a way that the implausibility arises only when reading the

[1] Item 40 is an exception as it contains a reflexive verb in Condition C ("*Dann* ***schminkte sich*** *der Minister mit dem Präsidenten*").

target word and not already from the combination of the preceding main verb and the agent. However, this could not always be fully achieved, as the main verb itself often introduces some degree of implausibility in the less plausible conditions. In the final sentence, the target word is always followed by an additional clause ("[...] and then he went to the gate") to capture spillover effects in RTs.

In the study by Aurnhammer et al. (2023), the main verb in the final sentence of Condition B was additionally chosen in such a way that the expectancy of the distractor word was higher than the expectancy of the target word. That is, in the final sentence "the lady *weighed* the tourist", the distractor word **"suitcase"** was globally available as a semantically attractive alternative, although it never appeared in the target position. As the aim of the current work is to test the predictions of single-trial plausibility ratings compared to averaged plausibility ratings, the availability of a semantically attractive alternative is not required. Therefore, the main verb in Condition B was changed to remove the ambiguity, i.e., to lower the expectancy of the distractor word, while maintaining graded plausibility. After changing, for example, "the lady *weighed* the tourist" to "the lady *welcomed* the tourist", the expectancy of the distractor word (**"suitcase"**) is lower than the expectancy of the target word ("tourist"), while the sentence remains less plausible than in Condition A and more plausible than in Condition C, given the preceding context. Thus, the stimuli used in the current study differ from those used by Aurnhammer et al. (2023) only with respect to the main verb in Condition B. Figure 4.1 shows an item used in the current study compared to an item used by Aurnhammer et al. (2023) and Figure 4.2 shows three example items in Conditions A, B and C for both the target and distractor words.

If the three levels of plausibility for the target word are effectively maintained, the self-paced reading study should replicate the graded RT effect observed by Aurnhammer et al. (2023), with RTs increasing as plausibility decreases, reflecting greater integration effort. To assess whether the manipulation in Condition B has been successful – i.e., whether distractor word expectancy has been reduced while maintaining graded plausibility – two norming studies are conducted prior to the self-paced reading study.

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 tourist an extra fee. After that, the tourist opened his **suitcase** and threw several things out. Now, the **suitcase** of the ingenious tourist weighed less than the maximum of 30 kilograms.*

Present study

Condition A: Plausible & no attraction
Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.
Then dismissed the lady the tourist and afterwards he went to the gate.

Condition B: Less lausible & no attraction
Dann begrüßte die Dame den Touristen und danach ging er zum Gate.
Then welcomed the lady the tourist and afterwards he went to the gate.

Condition A: Implausible & no attraction
Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.
Then signed the lady the tourist and afterwards he went to the gate.

Design by Aurnhammer et al. (2023)

Condition A: Plausible & no attraction
Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.
Then dismissed the lady the tourist and afterwards he went to the gate.

Condition B: Less lausible & attraction
Dann wog die Dame den Touristen und danach ging er zum Gate.
Then weighed the lady the tourist and afterwards he went to the gate.

Condition A: Implausible & no attraction
Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.
Then signed the lady the tourist and afterwards he went to the gate.

FIGURE 4.1: Experimental design of the current study compared to the study by Aurnhammer et al. (2023). The German word order is preserved in the English transliterations of the final sentences. Target words are underlined, and distractor words are highlighted in boldface.

Item 12
A paparazzi set up his big **camera** and waited for a famous actress...

Target
A: Then threatened the paparazzi the actress ...
B: Then recognised the paparazzi the actress ...
C: Then coloured the paparazzi the actress ...

Distractor
A: Then threatened the paparazzi the **camera** ...
B: Then recognised the paparazzi the **camera** ...
C: Then coloured the paparazzi the **camera** ...

Item 31
The guests stood excitedly in the church and listened to the priest's moving sermon. The bride could hardly wait for the moment when she would say "I do" to the groom and receive the **ring** ...

Target
A: Happily kissed the bride the groom ...
B: Happily left the bride the groom ...
C: Happily simplified the bride the groom ...

Distractor
A: Happily kissed the bride the **ring** ...
B: Happily left the bride the **ring** ...
C: Happily simplified the bride the **ring** ...

Item 10
A curator at a museum was in the process of organising a new exhibition. As it was about sculptural art, the curator had borrowed a **sculpture** from a gallerist...

Target
A: Then hugged the curator the gallerist ...
B: Then booked the curator the gallerist ...
C: Then collected the curator the gallerist ...

Distractor
A: Then hugged the curator the **sculpture** ...
B: Then booked the curator the **sculpture** ...
C: Then collected the curator the **sculpture** ...

FIGURE 4.2: Three example items, transliterated from German. Target words are underlined and distractor words are highlighted in boldface.

4.2 Language Model Architectures

In order to assess whether the expectancy of the distractor word has been successfully lowered in Condition B and whether the expectancy of the distractor word is generally lower than the expectancy of the target word across conditions, two Transformer-based LMs were used to calculate surprisal values. Both LMs, a German GPT-2 version (Schweter, 2020) and LeoLM (Plüster, 2023), are decoder-only architectures that are based only on the decoder component of the Transformer architecture. They use a causal attention mechanism that allows each token in the generated sequence to attend only to the previous and the current tokens, but not to future tokens. This feature is essential for computing surprisal values, as it ensures that the predictions are based only on the preceding context up to the current word, reflecting the sequential nature of language processing observed in humans.

GPT-2 and LeoLM primarily differ in the number of parameters and the size of the datasets on which they are trained. The first LM is a pre-trained German GPT-2 model (Schweter, 2020), which has the same architectural features as the original GPT-2 model (Radford et al., 2019) and belongs to the small GPT-2 version trained with 124 million parameters. The authors used the same training data as for a German BERT model[2], which consists of Wikipedia articles[3], the EU Bookshop corpus (Skadiņš et al., 2014), Open Subtitles (Lison and Tiedemann, 2016), ParaCrawl (Bañón et al., 2020), NewsCrawl (Ngo et al., 2021) and CommonCrawl[4]. This results in a training dataset of approximately 16 gigabytes of data and 2.3 million tokens. The second model, LeoLM (Plüster, 2023), is an open source German Foundation LM built on Llama-2, which is larger than GPT-2 in terms of the number of parameters and data it was trained on. LLama-2 is a family of LLMs ranging from 7 billion to 70 billion parameters that are pre-trained on approximately 2 trillion tokens of mostly English texts (Touvron et al., 2023). For the current study, the LeoLM version with 13 billion parameters was chosen. In order to improve its proficiency in the German language, LeoLM was initialised with Llama-2 weights and further trained on a large German text corpus containing 65 billion tokens of filtered web texts from the OSCAR corpus (Ortiz Suárez et al., 2019). LeoLM was also trained on two smaller datasets, consisting of Wikipedia[5] and Tagesschau[6] (news) articles, resulting in a training dataset of approximately 600 gigabytes.

As the parameter overview of the two LMs in Table 4.1 shows, LeoLM is a more powerful model. Since the general capabilities of the two LMs were not evaluated using a metric such as perplexity, their performance cannot be compared in this context. However, given the significant difference in size, LeoLM would most likely perform better (i.e., achieve a lower perplexity) than GPT-2 in a performance test.

One issue that should be addressed when using LM surprisal estimates in psycholinguistic research is related to the tokenisation of the input text. While GPT-2 uses Byte-Pair-Encoding (BPE; Sennrich et al., 2016) as a subword-tokenisation

[2] https://huggingface.co/dbmdz/bert-base-german-cased. [Accessed: 2024-04-16].
[3] https://dumps.wikimedia.org. [Accessed: 2024-04-16].
[4] https://commoncrawl.org. [Accessed: 2024-04-16].
[5] https://dumps.wikimedia.org. [Accessed: 2024-04-16].
[6] https://huggingface.co/datasets/bjoernp/tagesschau-2018-2023. [Accessed: 2024-04-16].

	GPT-2	LeoLM
Parameters	124M	13B
Vocabulary size	50,257	32,000
Context size	1024	4096
Embedding dimension	768	5120
Decoder/Hidden layers	12	32
Attention Heads	12	32

TABLE 4.1: Parameter overview for the two LM architectures used to compute surprisal values in a pre-test: GPT-2 and LeoLM.

algorithm, LeoLM uses the Llama tokeniser, which is based on SentencePiece (Kudo and Richardson, 2018). BPE works by iteratively merging the most frequent pairs of consecutive characters in the input text, while SentencePiece is a language-independent subword tokeniser that implements BPE and unigram LMs as subword segmentation algorithms. Unlike BPE, SentencePiece treats the entire input text as a single sequence without relying on whitespaces to define word boundaries, using a probabilistic model to segment text and generate the vocabulary. Both are commonly used tokenisation models that reduce the size of the model's vocabulary while maintaining its expressive power and enabling the model to make accurate predictions for rare or Out-of-Vocabulary (OOV) words.

From a cognitive modelling perspective, the problem is that the above mentioned subword tokenisation models rely on frequency-based approaches to build the subword vocabulary, which differs from the morphological subword decomposition in human processing (Nair and Resnik, 2023). Crucially, the surprisal values for orthographic words are calculated as the sum of the surprisal values of their subwords ($P(w) = P(sw_1) + ... + P(sw_n)$). Subwords are assigned the same token id, regardless of whether they occur as independent words or as a part of a word, as there is no way to distinguish between them. Since the surprisal of a word is derived from the surprisal values of its constituent subwords, words segmented into multiple subwords receive higher surprisal estimates by default than subword units that exist as independent words, even if the expectancy of these segmented words is higher. To avoid this problem, some studies included only items that were not split by the tokeniser (Michaelov et al., 2023). However, this approach is not appropriate for the current study as the items should be consistent with those used in Aurnhammer et al. (2023) for reasons of comparability.

4.3 Data Analysis

The distributions of the RTs that were collected in the two self-paced reading studies are right-skewed, indicating that a small proportion of the RT data consists of longer RTs than a larger proportion of the data. To normalise their distributions, the RT data were log-transformed. The log-transformed RTs were then analysed using the same linear mixed effects regression re-estimation technique used by Aurnhammer et al. (2021, 2023). Using this technique, a separate linear mixed effects model was fitted for each of the four critical regions to test the influence and significance of the two predictors in each region. The four critical regions of the final sentence include the determiner preceding the target word (the *Pre-critical region*), the target word (the *Critical region*) and the two words following the target word (the *Spillover region* and the *Post-spillover region*). The Spillover region usually consists of a conjunction ("und" / "*and*") and the Post-spillover region of an adverb (e.g., "danach" / "*afterwards*"), both belonging to the category of closed class words. In the following example, the critical regions are underlined: "Dann verabschiedete die Dame den (*Pre-critical*) Touristen (*Critical*) und (*Spillover*) danach (*Post-spillover*) ging er zum Gate".

The predictors included in the models are target word plausibility and distractor word surprisal, which are independent of each other as shown in Table 5.2 of Chapter 5. Target word plausibility serves as a continuous predictor to quantify the difficulty of integrating the target word with the preceding context and distractor word surprisal serves as a predictor to explain additional variability in RTs due to distractor word expectancy. The predicted values represent the model's best estimates of the RTs given the values of the predictor variables. In other words, the observed RTs (y) refer to the actual RTs measured during the experiments, while the estimated RTs ($\hat{y}$) are predicted by the linear mixed effects regression models. How well these estimates reflect the observed RTs can be assessed by the residual error, which is calculated by subtracting the observed RTs from the predicted RTs ($y - \hat{y}$).

Before fitting the models, both predictors were standardised by dividing the difference between each data point and the mean of the respective predictor by its standard deviation. This ensures, firstly, that the mean value of each predictor variable is 0 and the standard deviation is 1 and, secondly, that the intercept is equal to the mean of the outcome variable when all predictors are set to zero, thus facilitating the interpretation and comparison of the coefficients in the regression models. Thus, the coefficients represent the change in the outcome variable (RTs) associated with a 1 standard deviation change in the predictor variables (plausibility and surprisal, respectively). To further simplify the interpretation of the regression coefficients, the plausibility predictor was multiplied by -1, as higher RTs are expected for lower plausibility. By inverting this predictor variable, the coefficients for plausibility should be positive, indicating that as plausibility decreases, RTs increase.

The data from the two self-paced reading studies were then re-estimated using a separate linear mixed effects regression model for each of the four critical regions. Linear mixed effects models allow for the estimation of fixed effects, which capture the average effect of the predictors on the outcome, and random effects, accounting for differences within groups, e.g., subjects and items (Jaeger, 2008). As plausibility and surprisal are the predictors of interest, they were included in the model as fixed effects. Random slopes and intercepts were added for subjects and items to account for individual differences in RTs that are not explained by the fixed-effect predictors. For example, subjects may vary in terms of their reading speed or comprehension ability and items may introduce different levels of difficulty. Thus, the full model specification is

$$Y = \beta_0 + S_0 + I_0 + (\beta_1 + S_1 + I_1) PlausTar + (\beta_2 + S_2 + I_2) SurprisalDist + \epsilon \quad (1)$$

where β_0 is the fixed-effect intercept term, representing the value of the outcome variable (Y) when the predictor variables are zero. β_1 and β_2 represent the fixed-effect coefficients of plausibility and surprisal, respectively, indicating the average change in RTs for a one-unit change in the predictor variables. S_0 and I_0 represent random intercepts and S_1, I_1, S_2, I_2 random slopes for both subjects and items. The error term ϵ represents the random variability in the outcome variable that is not explained by the fixed-effect predictors or the random effects, capturing the difference between the observed and the predicted values.

First, single-trial target word plausibility and GPT-2 distractor word surprisal were included in the models as predictors (see Table 4.2). The same models were then fitted with LeoLM distractor surprisal. Next, averaged target word plausibility collected in a pre-test was included as a predictor, first along with GPT-2 distractor word surprisal and then along with LeoLM distractor word surprisal. This should reveal whether online plausibility ratings collected during the self-paced reading study or offline plausibility ratings collected in a pre-test are more predictive of the RT data. Moreover, this allows for a comparison between the predictions using surprisal estimates from different LMs. Coefficients, z-values and p-values were reported for all models. As separate analyses were performed for the four critical regions, treating each of them as a distinct set of hypotheses, it was not necessary to correct p-values for multiple comparisons. Figure 4.3 provides an overview of the study design and the predictors used to re-estimate the RT data.

	Plausibility		**Surprisal**	
	Online	Offline	GPT-2	LeoLM
Predictor Combination 1	+	-	+	-
Predictor Combination 2	+	-	-	+
Predictor Combination 3	-	+	+	-
Predictor Combination 4	-	+	-	+

TABLE 4.2: Predictor combinations used for fitting the linear mixed effects models: online or offline target word Plausibility combined with GPT-2 or LeoLM distractor word Surprisal.

Finally, a likelihood ratio test (LRT) was performed to objectively compare the goodness of fit of a simple model, including only the plausibility predictor that explains more variance in the outcome variable (together with distractor word surprisal), and a complex model, including both single-trial and averaged plausibility (and distractor word surprisal). The LRT determines whether the complex model (corresponding to the alternative hypothesis) significantly improves the model fit compared to the simple model (corresponding to the null hypothesis). Based on the number of degrees of freedom, which is equal to the number of additional parameters in the complex model, the LRT determines a critical value from the chi-squared distribution and compares it to the observed likelihood ratio statistic. If the likelihood ratio statistic is greater than the critical value, the null hypothesis is rejected, suggesting that the predictor that is included in the complex model, but not in the simple model, provides additional information that improves the model's ability to explain the observed data.

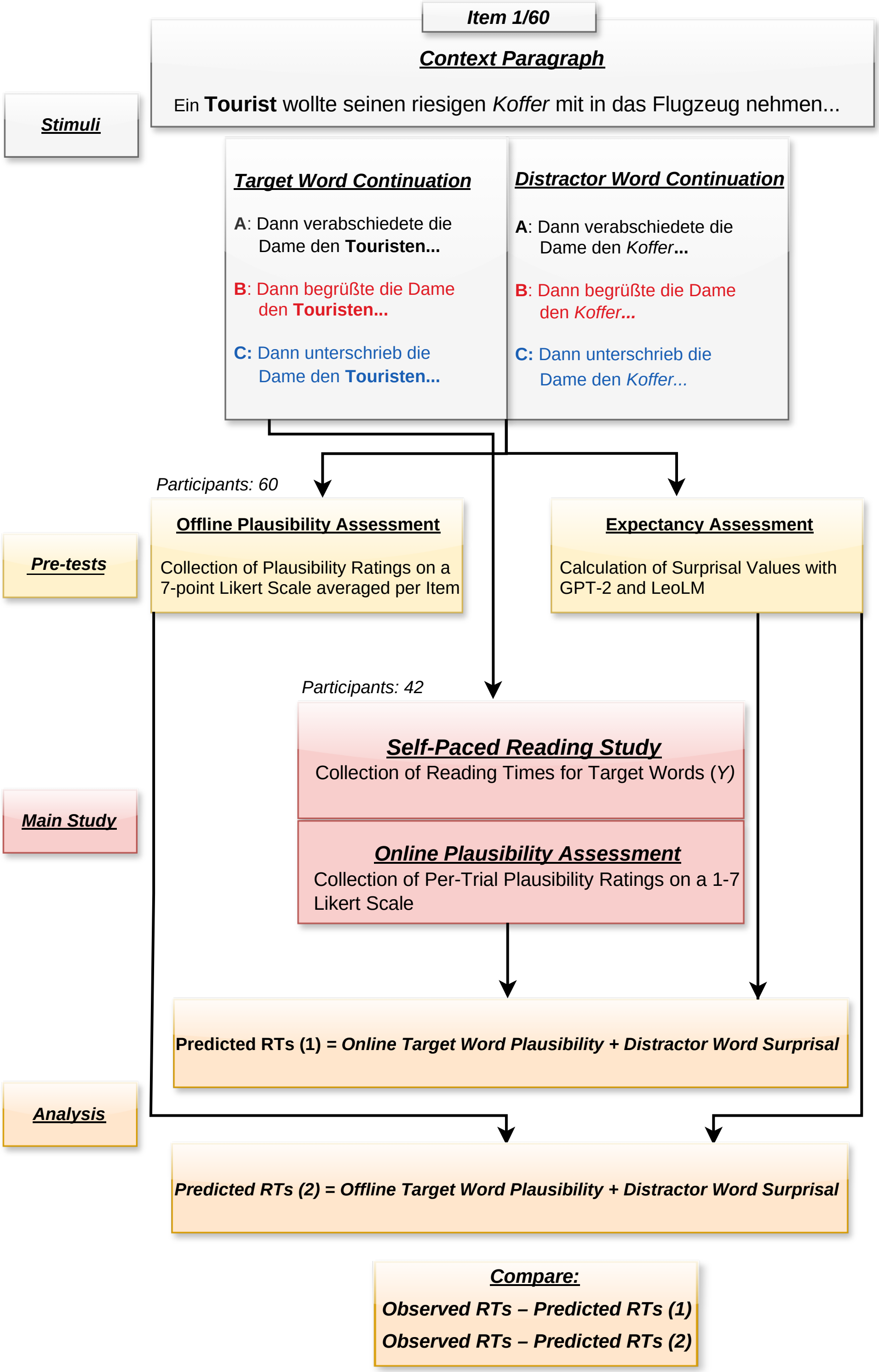
Item 1/60
Context Paragraph
Ein Tourist wollte seinen riesigen Koffer mit in das Flugzeug nehmen...
Stimuli
Target Word Continuation
A: Dann verabschiedete die Dame den Touristen...
B: Dann begrüßte die Dame den Touristen...
C: Dann unterschrieb die Dame den Touristen...
Distractor Word Continuation
A: Dann verabschiedete die Dame den Koffer...
B: Dann begrüßte die Dame den Koffer...
C: Dann unterschrieb die Dame den Koffer...
Participants: 60
Pre-tests
Offline Plausibility Assessment
Collection of Plausibility Ratings on a 7-point Likert Scale averaged per Item
Expectancy Assessment
Calculation of Surprisal Values with GPT-2 and LeoLM
Participants: 42
Main Study
Self-Paced Reading Study
Collection of Reading Times for Target Words (Y)
Online Plausibility Assessment
Collection of Per-Trial Plausibility Ratings on a 1-7 Likert Scale
Analysis
Predicted RTs (1) = Online Target Word Plausibility + Distractor Word Surprisal
Predicted RTs (2) = Offline Target Word Plausibility + Distractor Word Surprisal
Compare:
Observed RTs – Predicted RTs (1)
Observed RTs – Predicted RTs (2)

FIGURE 4.3: Study overview: Preparation of the stimuli, assessment of the manipulations in two pre-tests, implementation of a self-paced-reading study and linear mixed effects regression analysis using offline and online plausibility together with distractor word surprisal to predict the observed RT data.

Chapter 5

Pre-Tests

Prior to the self-paced reading experiment, two norming studies were conducted to test whether the expectancy and plausibility manipulations were successful. First, a plausibility rating study was conducted to ensure that plausibility is graded across conditions ($A > B > C$). Second, surprisal values were computed using GPT-2 and LeoLM to assess whether the expectancy of the distractor word is lower than the expectancy of the target word across conditions. The plausibility norming study and the self-paced reading experiments, were conducted as web-based experiments using the PCIbex software (Zehr and Schwarz, 2018).

5.1 Plausibility

5.1.1 Procedure

In the first norming study, plausibility ratings were collected to assess whether the items have been successfully manipulated in terms of plausibility. Participants were asked to rate the plausibility of the final sentence of each item in the context of the preceding context paragraph on a seven-point Likert scale, with 7 representing the highest plausibility and 1 the lowest plausibility. The context paragraph was presented together with the final sentence and the seven-point Likert scale. In contrast to the plausibility rating study by Aurnhammer et al. (2023), the continuation of the final sentence ("and after that he left the store") was not excluded in the current study, as the final sentence continuation is also included in the case of the single-trial plausibility ratings collected during the self-paced reading study in order to capture spillover effects in RTs. This is crucial for consistency reasons, as the post-target material may alter the plausibility ratings. Plausibility ratings were collected for both target and distractor words in all three conditions in order to assess whether plausibility is still graded after modifying the main verb in Condition B. Consequently, each item was assessed in six different conditions: A (*target*), B (*target*), C (*target*), A (*distractor*), B (*distractor*), and C (*distractor*), resulting in six variations of the final sentence.

To ensure that each participant read each item in only one condition and that all conditions received the same number of ratings from each participant, participants were assigned to six different lists. Participants in each list read the same items, which were different from those in the other lists. As each participant rated a total of 60 items, 10 in each of the six conditions, a total of 3600 ratings were collected. In addition to the 60 critical items, each participant was presented with 12 filler items, in which plausibility was also varied across three levels. Since the purpose of the filler items was to ensure that participants read the texts carefully, they contained instructions in the middle of the context paragraph that asked participants to assign a plausibility rating of either 1 or 7 to the item, regardless of its actual plausibility. If more than 2 out of 12 attention checks were failed, the participant's data were excluded from further analyses. In addition, participants were presented with three practice items, so that each participant encountered a total of 75 items, including critical items, filler items, and practice items.

A total of 66 participants were recruited through Prolific Academic Ltd.[1], an online platform for research recruitment. Each participant was paid £4.95 and agreed to participate in the study by approving a consent form. The data from six participants were excluded due to exceptionally fast completion of the study (more than three standard deviations below the mean) or more than 2 out of 12 failed attention checks. All participants were native speakers of German, aged between 18 and 32 years, with no language-related disorders or literacy difficulties.

5.1.2 Results

On average, participants rated 99% (mean = 99.16%, SD = 2.52, range = 91.66%-100%) of the items containing attention checks correctly, i.e., with the number that was indicated in the context paragraph. Table 5.1 shows the mean, standard deviation and range of the collected plausibility ratings for both target and distractor words in the three conditions.

The mean plausibility ratings show that target word plausibility is graded $(A > B > C)$, indicating that Condition B was successfully modified so that participants on average rated it as medium plausible. Additionally, the results align with those of Aurnhammer et al. (2023), who found graded plausibility across conditions. The larger standard deviation and range in Condition B, compared to Conditions A and C, reflect increased variability in participants' plausibility ratings due to the difficulty in judging items of medium plausibility compared to items that are clearly plausible or implausible. For the distractor word, the mean plausibility ratings are also graded $(A > B > C)$. However, the differences between the mean plausibility ratings per condition are small, as they all fall in the lower (implausible) range of the scale. This doesn't seem surprising, since replacing the target word with the distractor word in the final sentence renders even Conditions A and B less plausible ("Then *dismissed* (A)/*welcomed* (B) the lady the **suitcase**"). Although the expectancy of the target and distractor words was assessed in a separate norming study, the lower plausibility of the distractor word compared to the target word in Condition B suggests that the expectancy of the distractor word was successfully reduced, as less plausible stimuli are generally less expected. In this context, it is worth noting that while the plausibility of the sentence containing the target word is higher than the plausibility of the sentence containing the distractor word in Conditions A and B, the opposite is true for Condition C. This could be due to the combination of the implausibility of Condition C and the generally lower expectancy of the distractor compared to the target word, which, at least in some cases, renders the final sentence slightly more plausible than in the target condition. For example, the sentence "Then *signed* the

[1] `https://www.prolific.com/`. [Accessed: 2024-01-27].

woman the **suitcase** (distractor)" seems slightly more plausible than "Then *signed* the woman the tourist (target)", because signing objects is generally more plausible (and expected) than signing people (see also the example items in Table 5.2). However, this should not affect the subsequent analyses, as the main goal was to achieve a graded effect for target word plausibility. Moreover, both target and distractor words received similarly low plausibility ratings on average in Condition C.

Figure 5.1 shows the distributions of the average plausibility ratings per item in Conditions A, B and C for both the target and the distractor words. The top left density plot shows the distribution of plausibility ratings for the target word. Conditions A and C show a unimodal distribution, peaking at plausibility levels of 6.5 and 1.5 respectively. The distribution of Condition C is slightly skewed to the right and the distribution of Condition A is slightly skewed to the left. This indicates that participants assigned mostly low ratings to the items in Condition C and high ratings to the items in Condition A. In contrast, Condition B shows a bimodal distribution, with less pronounced peaks around 3 and 4.5, indicating that most items received ratings corresponding to a medium level of plausibility. However, the averaged ratings are more spread out across the entire spectrum in Condition B than in Conditions A and C. As noted above, this is not surprising, since judging nuances of medium plausibility is inherently more difficult than judging clearly plausible or implausible items. On the top right, Figure 5.1 shows the distribution of plausibility ratings for the distractor word. On average, plausibility ratings for Condition C are more concentrated at the implausible end of the scale compared to Conditions A and B. In contrast, Conditions A and B show broader distributions, indicating greater variability in participants' ratings, although most items were also rated as implausible.

Based on the results of this plausibility study and those of Aurnhammer et al. (2023), it is predicted that the single-trial plausibility ratings that will be collected during a self-paced reading study will follow the same pattern. In addition, the self-paced reading study is predicted to show graded RTs for target word plausibility, with implausible items being read more slowly compared to plausible items, $(A < B < C)$, reflecting increased integration effort.

5.2 Surprisal

5.2.1 Procedure

A second norming study was conducted to assess whether the expectancy of the target word is higher than the expectancy of the distractor word across conditions. Specifically, the goal was to assess whether the expectancy of the target word is higher than the expectancy of the distractor word in Condition B, since the manipulation of the main verb aimed to eliminate the ambiguity in Condition B by reducing the expectancy of the distractor word. Since Conditions A and C were adopted from Aurnhammer et al. (2023) without modification, the expectancy of the target words is predicted to be higher than the expectancy of the distractor words in this study as well, even though a different metric was used to assess their expectancy. Aurnhammer et al. (2023) determined the expectancy of the target and distractor words based on cloze probabilities, a human-based operationalisation of expectancy. In the current study, surprisal, an LM-derived operationalisation of expectancy, is used to estimate the expectancy of target and distractor words across conditions. If the expectancy of the distractor word was successfully lowered, this should be reflected in lower surprisal values for the target than for the distractor word in Condition B (as well as in the unmodified Conditions A and C).

Two different Transformer-based LMs were used to compute the surprisal values for the target and distractor words: a pre-trained German GPT-2 model (Schweter, 2020) and LeoLM (Plüster, 2023), a German Foundation LM built on Llama-2 (see Chapter 4.2). The sentence materials used as input to the LMs are the same as in the plausibility rating study.[2] First, the stimuli were preprocessed using a regular expression that inserted a whitespace between all instances of an alphanumeric character adjacent to a non-alphanumeric character. For example, a whitespace was inserted between the letter "*e*" and the full stop at the end of the following sentence: "*Der Urlauber freute sich über den Flyer und dankte dem Guide* **.**". This ensures that the tokeniser recognises non-alphanumeric characters as separate tokens rather than treating them as part of the preceding word. In a second preprocessing step, the final sentence of each item was truncated after the target/distractor word, excluding the continuation of the final sentence, since only the surprisal of the target and distractor word given the preceding context is relevant. For example, the stimulus "Then *dismissed* the lady the tourist/**suitcase** and afterwards he went to the gate" was truncated after tourist/**suitcase**.

The models then processed the input sequence and generated logits, which were transformed into probabilities using the softmax function. Each probability score represents the likelihood of the corresponding token being the next token in the sequence. Surprisal values for the tokens were then computed by applying the $-log_2$ function to the probability estimates, measuring how surprising the token is given the preceding sequence. Finally, the tokens, i.e., the subword units created during tokenisation, were recombined to form the original stimuli based on the different word encodings. The surprisal values of the recombined tokens were summed to obtain a single surprisal value for each word (see also Oh and Schuler, 2022; De Varda et al., 2023).

5.2.2 Results

The average surprisal values computed with GPT-2 are higher for the distractor words than for the target words across all conditions (see Table 5.1). Given that surprisal is inversely proportional to expectancy, this shows that, on average, the expectancy of the target word is higher than the expectancy of the distractor word in all conditions. This indicates that the distractor word expectancy in Condition B was successfully lowered while maintaining a medium level of plausibility. Similar to distractor word plausibility, the expectancy of the distractor word in Condition C is slightly higher than the expectancy of the target word when considering LeoLM surprisal. One potential explanation, that was discussed in the previous section, is that the combination of the less expected distractor word and the implausible Condition C renders some items more plausible and expected compared to items in which the target word appears in the context of Condition C (see also Figure 4.3).

[2] Except for the filler items, for which no surprisal values were required.

Furthermore, the surprisal values of both GPT-2 and LeoLM show a graded pattern across conditions for the target word $(C > B > A)$, indicating that the target word in Condition C is on average less expected than the target word in Conditions A and B, which is consistent with the target word plausibility levels. The average surprisal values calculated by GPT-2 for the distractor word differ from the three expectancy levels of Conditions A, B and C computed for the target word. Although Condition C has the highest average surprisal value, Condition A has slightly higher surprisal than Condition B $(C > A > B)$. LeoLM's average surprisal values for the distractor word deviate even further from the target word pattern $(A > B > C)$.

Finally, a direct comparison of the two LMs shows that the average surprisal values computed by the larger LM, LeoLM, are slightly lower than those computed by GPT-2 across all conditions, except for the distractor word in Condition A.

		Plausibility			**Surprisal (GPT-2)**			**Surprisal (LeoLM)**		
	Cond.	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Target	A	6.03	0.71	4.40-7.00	2.36	2.33	0.06-10.52	0.74	0.99	0.01-5.37
	B	3.79	1.20	1.70-6.80	3.95	3.58	0.03-16.76	3.36	3.24	0.16-16.77
	C	1.91	0.57	1.00-3.30	6.61	4.70	0.13-18.71	5.48	3.93	0.46-17.90
Distractor	A	2.97	1.48	1.20-6.80	6.79	4.98	0.24-21.67	9.04	4.79	0.97-24.00
	B	2.92	1.41	1.10-6-40	6.55	4.41	0.15-20.90	5.56	4.02	0.35-19.43
	C	2.11	0.83	1.00-4.70	7.05	4.74	0.12-19.07	5.30	3.66	0.47-15.49

TABLE 5.1: Averages, standard deviations and ranges for the results of the two pre-tests that collected seven-point scale plausibility ratings and surprisal values for the target and distractor words.

Figure 5.1 shows the distributions of the surprisal values computed by GPT-2 and LeoLM for the target and distractor words. As some words are highly unexpected and therefore have very high surprisal values, the densities of the surprisal values are strongly right-skewed. Although the distributions of the surprisal values in the different conditions overlap more than in the case of plausibility, the pattern C > B > A can be seen from the densities and the dashed lines representing the mean surprisal values. The high density of surprisal values observed for Condition A in the lower range aligns with the summary statistics in Table 5.1, confirming that the LMs, particularly LeoLM, predominantly predict low surprisal values, reflecting high expectancy. Since the target words in Conditions B and C are less predictable than those in Condition A, this is reflected in the wider spread of density curves in Conditions B and C. The spread of the surprisal values appears to be even wider for the distractor word due to its low expectancy compared to the target word. In the case of GPT-2, the density curves for the different conditions overlap almost completely, whereas the surprisal values computed by LeoLM, differ mainly in the density of high surprisal values for Condition A compared to Conditions B and C.

Table 5.2 shows correlations between plausibility, GPT-2 surprisal and LeoLM surprisal for target and distractor words. Target word GPT-2 surprisal and target word LeoLM surprisal show the strongest positive correlation ($r = 0.60$) among all variables, followed by distractor word GPT-2 surprisal and distractor word LeoLM surprisal ($r = 0.56$). Furthermore, a moderate negative correlation can be observed

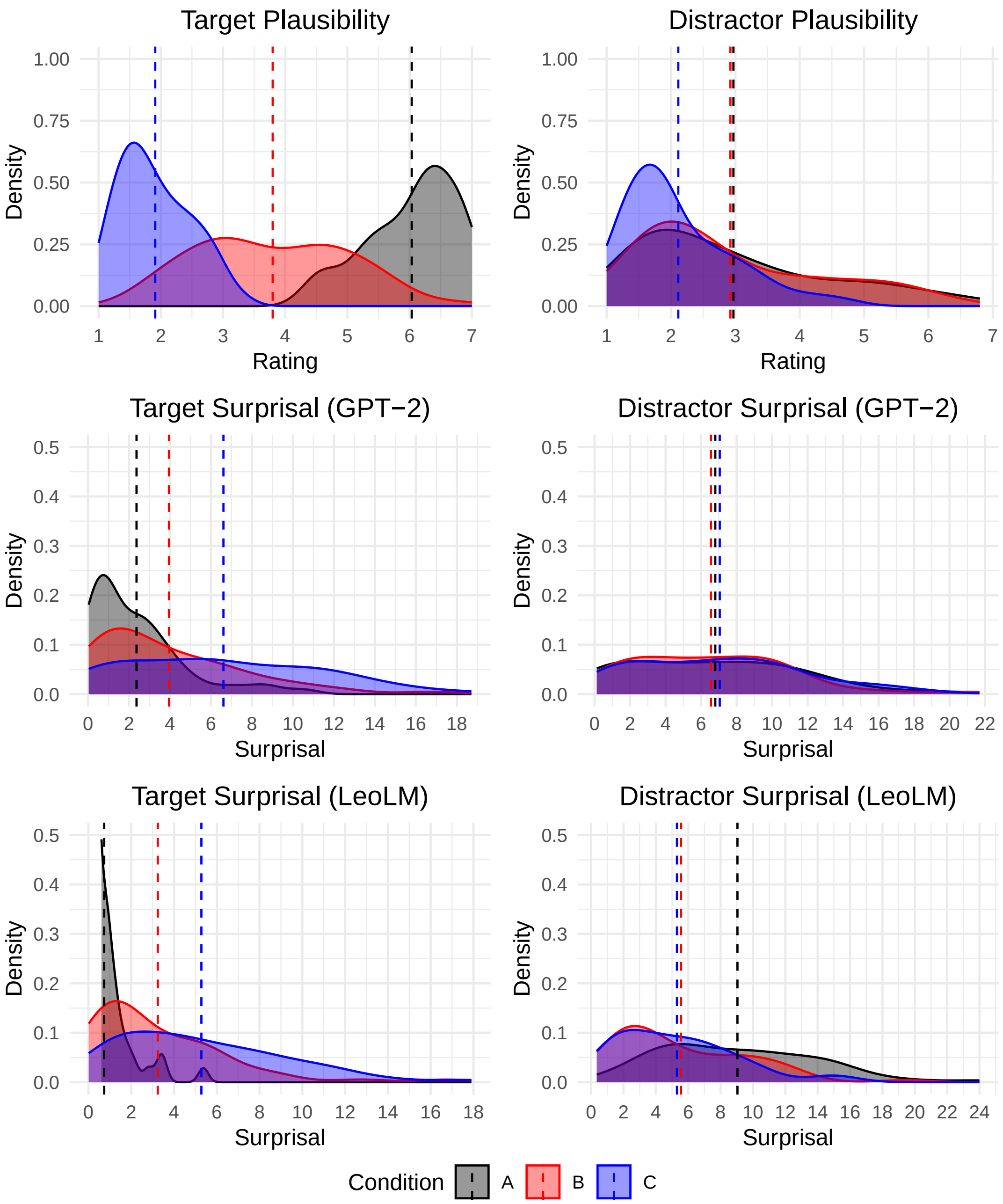

FIGURE 5.1: Densities for the results of the plausibility rating study that collected seven-point scale plausibility ratings and the surprisal values computed by GPT-2 and LeoLM for the target and distractor words.

between target word plausibility and target word LeoLM surprisal (r = -0.51) and a weaker negative correlation between target word plausibility and target word GPT-2 surprisal (r = -0.36). The negative sign indicates that as target word plausibility increases, target word surprisal decreases, reflecting higher expectancy due to the inverse relationship between surprisal and expectancy. The correlations indicate that LeoLM surprisal aligns more closely with human plausibility judgments than GPT-2 surprisal, suggesting that larger LMs show more human-like understanding and reasoning capacities compared to smaller variants. There is a weak positive correlation between target word plausibility and LeoLM distractor surprisal (*r* = 0.28). In contrast, there is virtually no linear relationship between target word plausibility and distractor word GPT-2 surprisal (*r* = -0.01), indicating that the two variables are independent of each other. Since both target word plausibility and distractor word surprisal are used as predictors in the subsequent RT analysis to explore graded effects of plausibility and (no) effects of semantic attraction, the independence of the predictors is crucial to ensure accurate coefficient estimates. Thus, the correlation between target word plausibility and distractor word LeoLM surprisal, which are used as predictors for RTs, should ideally be closer to zero. However, the relationship is still not excessively high and should not necessarily be problematic in terms of multicollinearity, for example.

		Plausibility		**Surprisal** (GPT-2)		**Surprisal** (LeoLM)	
		Target	Distractor	Target	Distractor	Target	Distractor
Plausibility	Target	1.00	0.35	- 0.36	- 0.01	- 0.51	0.28
	Distractor	0.35	1.00	0.08	- 0.32	- 0.02	- 0.34
Surprisal **(GPT-2)**	Target	- 0.36	0.08	1.00	- 0.34	0.60	- 0.25
	Distractor	- 0.01	- 0.32	- 0.34	1.00	- 0.12	0.56
Surprisal **(LeoLM)**	Target	- 0.51	- 0.02	0.60	- 0.12	1.00	- 0.13
	Distractor	0.28	- 0.34	- 0.25	0.57	- 0.13	1.00

Table 5.2: Correlations between offline plausibility ratings and (GPT-2 and LeoLM) surprisal of the target and distractor words.

Plausibility and expectancy are less aligned in the current study compared to the study by Aurnhammer et al. (2023). These differences could either be due to the human-derived nature of cloze probabilities, which may be more closely aligned with plausibility judgments compared to LM-based surprisal estimates, which are not unproblematic, for example, in terms of subword creation during tokenisation (see Nair and Resnik, 2023, for a discussion), or to other factors.

Ultimately, distractor word surprisal may not impact the analyses of the data from the self-paced reading study at all. In line with previous research (Rich and Harris, 2021), Aurnhammer et al. (2023) found no significant RT modulations due to distractor word cloze probability despite the high distractor word expectancy in Condition B. This suggests that behavioural measures such as RTs might not be sensitive to unfulfilled expectations. Since the surprisal pre-test has shown that the expectancy of the distractor word is low across conditions, including Condition B, distractor word surprisal should certainly not modulate RTs in the current study. In other words, given that the distractor word expectancy in Condition B has been reduced, no significant RT modulation due to distractor word surprisal should be observed, even if RTs were to be sensitive to unfulfilled expectations.

Chapter 6

Self-Paced Reading Study I

In the first experiment, a self-paced reading study was conducted to (1) investigate whether RTs are graded for plausibility, with plausible items being read faster on average than medium plausible and especially implausible items, and (2) determine whether single-trial plausibility ratings collected online during self-paced reading are a better predictor of the RT data than averaged plausibility ratings collected in a pre-test. The materials for the rating task were identical to those used in the plausibility pre-test. However, in the self-paced reading study, RTs were recorded only for the target word, but not the distractor word.

6.1 Participants

A total of forty-five participants were recruited via the platform Prolific Academic Ltd. to take part in the web-based self-paced reading study, which, similar to the plausibility rating study, was conducted using the experiment platform PCIbex (Zehr and Schwarz, 2018). The data of three participants were excluded from the subsequent statistical analyses due to inattentive reading, as demonstrated by low response accuracy on the comprehension questions (less than 80% correct). The remaining 42 participants (mean age 26.26; SD 3.7; age range 19-32; 17 male, 25 female) were all native German speakers (including three early bilinguals) who did not report any language-related disorders or literacy difficulties. To ensure that participants had no previous exposure to the study materials, individuals who had participated in the plausibility rating study of this thesis or any of the equivalent studies in Aurnhammer et al. (2023) were excluded from the subsequent self-paced reading study. Prior to their participation, participants consented to the study by agreeing to a consent form. Each participant was paid £8.20 for taking part in the study.

6.2 Procedure

Similar to the plausibility rating study, the self-paced reading study was conducted as a web-based experiment. Seven out of forty-two participants were assigned to one of six different lists and read different materials depending on their assignment, ensuring that each participant read each of the 60 items in only one condition (e.g., 1A, 2B, 3C) and simultaneously that all items were read by an equal number of participants. Each list consisted of three blocks, with each block containing 20 critical items and 15 filler items, resulting in 35 items per block and a total of 105 items per list, of which 60 were critical and 45 were filler items. As only half of the materials from the plausibility rating study – those containing the target word – were used in the self-paced reading study, there were a total of 180 item variations in total. Therefore, only three of the six lists contained unique items, while the other three lists contained the same items arranged in a different order. Specifically, in half of the lists, the order of the blocks was reversed compared to the other half, and the items were randomised within each of the three blocks. Consequently, each condition of an item was read by exactly 14 different participants – seven from a forward presented list and seven from a backward presented list.

After the participants confirmed their participation by agreeing to a consent form and provided demographic information (languages spoken, age, gender and handedness), they were presented with three practice items to familiarise them with the task before the start of the experiment. To start a trial, participants had to press the *Enter* key, after which only the context paragraph of an item appeared on the screen. Upon pressing the *Enter* key again, a hash sign appeared in the middle of a blank page, indicating the position in which the words of the final sentence would be subsequently presented. After that, participants read the final sentence word by word by pressing the *Space* key after each word to move to the next word. In this way, the time it took participants to read each word was measured, which is what is referred to as reading time. After reading the last word of the final sentence, participants were presented with a seven-point Likert scale, with 7 indicating a very plausible sentence and 1 indicating an implausible sentence, based on which they were asked to rate the plausibility of the final sentence given the context paragraph they had just read. This structure differs from the structure of the plausibility pre-test, where the seven-point Likert scale was presented on the same page as the context paragraph and the final sentence, allowing participants to reread the entire item as many times as necessary before providing a rating. However, since the self-paced reading design does not allow for the Likert scale to be presented on the same page as the context paragraph or the final sentence, participants have to more or less remember the content in order to give a rating. Although participants could reread individual parts of the item, such as the context paragraph or each word of the final sentence before moving on to the next word, they could not see the entire item or even the entire final sentence, including the rating scale, at once.

In 46% of the trials – half of the experimental trials and two-fifths of the filler items – the plausibility rating task was followed by a comprehension question that could refer to the context paragraph or the final sentence, within which it could focus on the manipulated region or the final sentence continuation. Participants could respond to the questions by pressing either the *D* key (corresponding to *Yes*) or by pressing the *K* key (corresponding to *No*), each of which was the correct answer for 50% of all questions. After the practice items and between each of the three blocks, participants received general feedback on their response accuracy to the comprehension questions (low, medium, high) to encourage attentive reading. As participants' response accuracy on the comprehension questions is presumably reflective of their attention during reading, this was used as a criterion to exclude the data of participants with too low overall response accuracy (below 80%) from all statistical analyses. In addition, participants were encouraged to take a short break between each of the three blocks.

6.3 Analysis

The items were analysed using a linear mixed effects regression re-estimation method (see also Aurnhammer et al., 2021, 2023). The analysis and all data pre-processing steps are described in detail in Chapter 4.3. Prior to the statistical analysis, trials were excluded if the reading time on any of the four critical regions was lower than 50 ms or higher than 2500 ms and if the reaction time on the task, i.e., on the comprehension question (in case there was one), was lower than 50 ms or higher than 10,000 ms. Based on these criteria, 7 out of 2520 trials (0.28%) were excluded.

6.4 Results

The results of the comprehension questions, the single-trial plausibility ratings collected online, as well as the observed RTs and their statistical analyses are described in the following sections.

6.4.1 Comprehension Questions

All participants answered comprehension questions on approximately half of the experimental items (46% of all trials) and two-fifths of all filler items. The descriptive statistics for response accuracy and reaction time on the comprehension questions were calculated across subjects. The mean accuracy was 95.2% (SD = 5.5, range = 80% - 100%). The mean reaction time on the comprehension questions was 2929 ms (SD = 627, range = 1757 ms - 4472 ms). The mean response accuracies and reaction times per condition are presented in Table 6.1. Condition A has the highest mean accuracy (96.4%), followed by Condition C (95.0%) and then Condition B (94.3%), suggesting that it was slightly easier for participants to answer questions about plausible items compared to items of low or medium plausibility correctly. Interestingly, the average reaction times are highest in Condition A (2947 ms), closely followed by Condition B (2941 ms) and then Condition C (2903 ms), indicating that, on average, participants processed and answered questions about plausible items more slowly than questions about medium plausible and especially implausible items. However, these differences are relatively small and should not be over-interpreted, especially when considering the higher standard deviation for mean accuracy in Condition B and for mean reaction time in Condition C, indicating greater variability in participants' accuracy and reaction time to the comprehension questions.

	Accuracy			Reaction Time		
Condition	Mean	SD	Range	Mean	SD	Range
A	96.4%	6.9	70.0% - 100.0%	2947 ms	660	1819 ms - 5009 ms
B	94.3%	8.6	70.0% - 100.0%	2941 ms	676	1620 ms - 5059 ms
C	95.0%	8.4	60.0% - 100.0%	2903 ms	723	1698 ms - 5106 ms

TABLE 6.1: Task performance on the comprehension questions in the first self-paced reading study. Accuracy and reaction times were computed across subjects.

6.4.2 Online Plausibility Ratings

During the self-paced reading study, plausibility ratings were collected on each trial based on the preceding context paragraph and the word-by-word presented final sentence. However, the number of plausibility ratings collected during the self-paced reading study differed from the number of ratings collected in the pre-test due to the different number of participants (60 in the pre-test and 42 in the self-paced reading study). In addition, in the self-paced reading study, plausibility ratings were only collected for the target word, resulting in 2520 trials (60 items read by 42 participants), whereas in the plausibility rating study conducted as a pre-test plausibility ratings were collected for both target and distractor words, resulting in 3600 trials (60 items read by 60 people), 1800 of which contained plausibility ratings for the target word. However, as the ratings from the plausibility pre-test were averaged per item across participants, this resulted in only 180 plausibility ratings, one per item and condition.

Another difference to the plausibility pre-test is that in the self-paced reading study, trials, including the respective plausibility ratings, were excluded if the RTs or reaction times on the task were too low or too high (see section 6.3). In contrast, during the plausibility rating pre-test, only the entire data of a participant could be discarded based on multiple failed attention checks, but it was not possible to exclude individual trials based on reading or reaction times, as these metrics were not recorded. The descriptive statistics of the plausibility ratings collected during the self-paced reading study are shown in Table 6.2 (right) and the densities of the plausibility ratings averaged across subjects and items are shown in Figure 6.1 (right). The plausibility ratings for the target word collected during the self-paced reading study follow a graded pattern across conditions $(A > B > C)$, similar to the results of the plausibility pre-test. However, the differences between the condition averages are smaller in the case of the online plausibility ratings than in the case of the offline plausibility ratings.

		Averaged Plausibility			Single-trial Plausibility		
	Condition	Mean	SD	Range	Mean	SD	Range
Target	A	6.03	0.71	4.40-7.00	5.84	0.78	4.00-6.90
	B	3.79	1.20	1.70-6.80	3.93	1.04	1.80-6.70
	C	1.91	0.57	1.00-3-30	2.20	0.69	1.10-4.30

TABLE 6.2: Averages, standard deviations, and ranges for the results of two studies that collected plausibility ratings offline in a pre-test (left) and online during the self-paced reading study (right) on a seven-point scale for the target word.

The correlations between single-trial and averaged pre-test target word plausibility and GPT-2 and LeoLM surprisal for target and distractor words are reported in Table 6.3. Distractor word plausibility is not included in the table, as it was not assessed during the self-paced reading study. The strongest observed correlation is between single-trial target word plausibility and averaged pre-test target word plausibility (r = 0.72). The correlation between the averaged pre-test plausibility and the single-trial plausibility ratings averaged across subjects per item is even higher (r = 0.92), although it is not explicitly shown in the table. This indicates a strong relationship between the plausibility ratings collected in both studies.

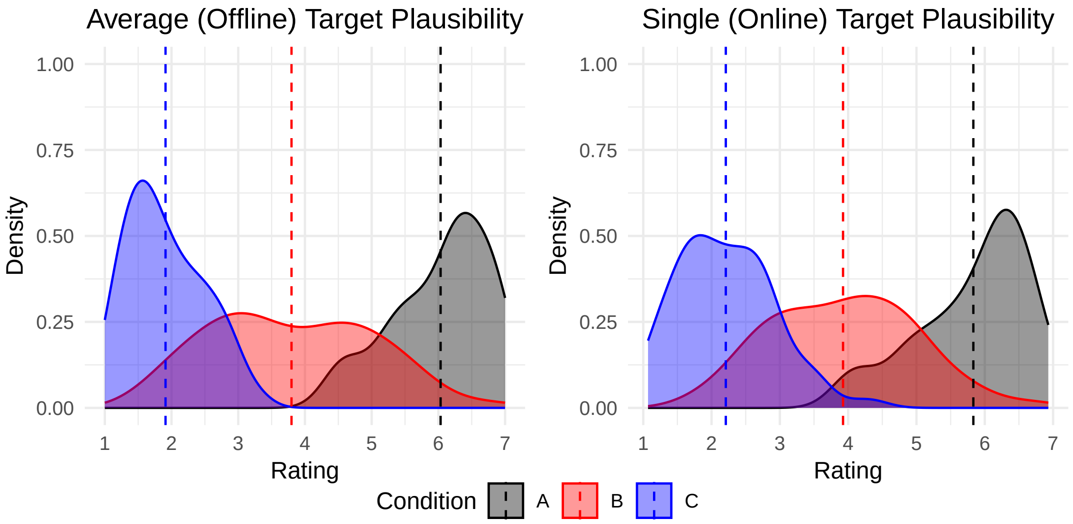

FIGURE 6.1: Densities for the results of two studies that collected plausibility ratings offline in a pre-test (left) and online during the self-paced reading study (right) on a seven-point scale for the target word.

		Plausibility (Tar)		**Surprisal** (Tar)		**Surprisal** (Dist)	
		Averaged	Single	GPT-2	LeoLM	GPT-2	LeoLM
Plausibility (Tar)	Averaged	1.00	0.72	- 0.36	- 0.51	- 0.01	0.28
	Single	0.72	1.00	- 0.25	- 0.35	- 0.05	0.16
Surprisal (Tar)	GPT-2	- 0.36	- 0.25	1.00	0.60	- 0.35	- 0.25
	LeoLM	- 0.51	- 0.35	0.60	1.00	- 0.12	- 0.13
Surprisal (Dist)	GPT-2	- 0.01	- 0.05	- 0.35	- 0.12	1.00	0.57
	LeoLM	0.28	0.16	- 0.25	- 0.13	0.57	1.00

TABLE 6.3: Correlations between averaged plausibility ratings collected in a pre-test, single-trial plausibility ratings collected online during the self-paced reading study for the target word and GPT-2 and LeoLM surprisal for the target and distractor word.

In terms of target word surprisal, the strongest (negative) correlation is observed between LeoLM surprisal and averaged target word plausibility, followed by GPT-2 surprisal and averaged target word plausibility. Conversely, the negative correlations between target word surprisal and single-trial target word plausibility are lower, particularly between single-trial plausibility and GPT-2 surprisal. Crucially, the negative correlations between GPT-2 distractor surprisal and averaged target word plausibility (r = - 0.01) and between GPT-2 distractor surprisal and single-trial target word plausibility (r = - 0.05) are close to zero, indicating that these variables are virtually independent of each other. The correlations between LeoLM distractor word surprisal and single-trial target word plausibility (r = 0.16) and especially between LeoLM distractor word surprisal and averaged target word plausibility (r = 0.28) are higher. Although the latter can still be considered as rather weak correlations, the correlation between the predictor variables should ideally be zero or close to zero. Higher correlations between predictor variables are more likely to cause multicollinearity, which makes it difficult to determine the individual effect of each predictor variable because their effects are confounded with each other. Furthermore, the correlations between LeoLM distractor word surprisal and both single-trial and averaged target word plausibility are positive, indicating that surprisal increases as plausibility increases. This is the case because the surprisal values calculated by LeoLM for the distractor word follow, perhaps unexpectedly, the same pattern as averaged and single-trial plausibility (A > B > C). Possible reasons for this are discussed in Chapter 5.2.2.

6.4.3 Reading Times

The observed log-transformed RTs per condition in the Pre-critical region (the determiner of the target word, e.g., "den" / "*the*"), the Critical region (the target noun, e.g., "Touristen" / "*tourist*"), the Spillover region (a conjunction that introduces the final sentence continuation, usually "und" / "*and*") and the Post-spillover region (usually an adverb, e.g., "danach" / "*afterwards*") are shown in Figure 6.2.

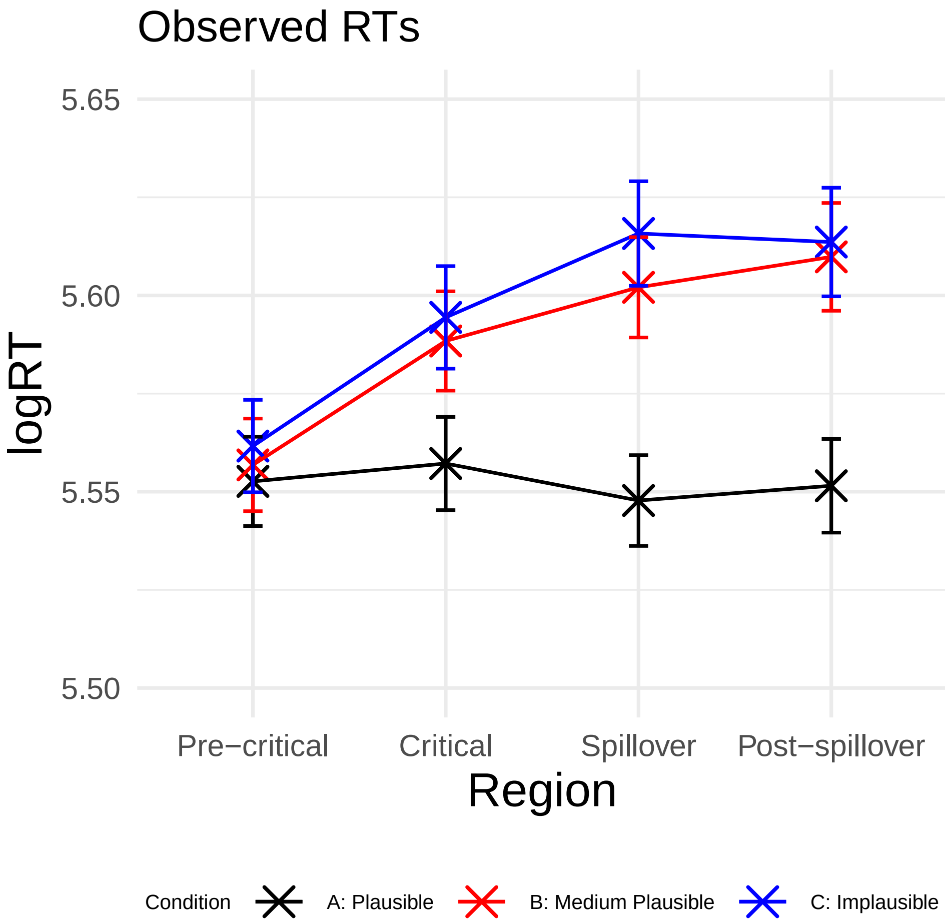

FIGURE 6.2: Log reading times per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions in the first self-paced reading experiment. The error bars show the standard error calculated from the per-subject per-condition averages.

The observed RT data indicates that, on average, the RTs in the Pre-critical region are lower than the RTs in the other regions and almost identical across conditions. However, upon closer inspection, they already seem to follow the pattern $A < B < C$ in the Pre-critical region, corresponding to the three plausibility levels associated with Conditions A, B and C. In the critical region, i.e., upon reading the target word, RTs increase in all conditions, but to a greater extent in Conditions B and C than in Condition A. In the subsequent Spillover and Post-spillover regions, RTs diverge even further, as they continue to increase in Conditions B and C but decrease in Condition A. Generally, the average RTs are graded for plausibility in all regions: Items in Condition A, corresponding to a high level of plausibility, were read the fastest on average, while items in Condition C, corresponding to a low level of plausibility, were read the slowest on average. However, this gradation is not very pronounced, as the average RTs for Conditions B and C are very similar in all four critical regions, especially in the Post-spillover region. While the pattern is the most pronounced in the Spillover region, the difference in RTs between Conditions B and C is still less distinct compared to the difference in RTs between Conditions A and B. This suggests that items in Condition B, representing a medium level of plausibility, were frequently rated as somewhat implausible, or that items in Condition C, reflecting a low level of plausibility, received more medium or high plausibility ratings than expected. The average RTs per region are similar to the RTs observed in the self-paced reading study conducted by Aurnhammer et al. (2023) for the Pre-critical region. However, they do not increase as much in the Spillover and Post-spillover regions and the difference between the RTs in Conditions B and C is small compared to the RTs observed by Aurnhammer et al. (2023).

A linear mixed effects model was fitted separately for each critical region in order to isolate the influence of each predictor variable on the RTs in each critical region. Both averaged target word plausibility collected in a pre-test and single-trial target word plausibility collected during the self-paced reading study were used in combination with GPT-2 distractor word surprisal or LeoLM distractor word surprisal, respectively, resulting in four different predictor combinations. The estimated RTs from these models are shown in Figure 6.3 and the corresponding residuals, i.e., the differences between the observed RT data and the predicted RTs are shown in Figure 6.4.

The relatively small residuals indicate that the models overall capture the effects structure in the observed RT data. However, the extent to which they capture this structure varies depending on the predictor combinations and the conditions and regions analysed. The models including averaged target word plausibility and either GPT-2 distractor word surprisal or LeoLM distractor word surprisal appear to capture the effects structure in the observed RT data better than the two models that used single-trial target word plausibility, as indicated by the smaller residual errors across regions and conditions. The residual errors in the models fitted with averaged target word plausibility are particularly small for Conditions A and C, but larger for Condition B. As both models underestimate the RTs in Condition B to a greater extent, the gradation of the estimated RTs appears more pronounced than in the observed RTs. Whether GPT-2 distractor word surprisal or LeoLM distractor word surprisal is used in combination with averaged target word plausibility does not seem to affect prediction accuracy. Inspection of the residual errors shows that the models incorporating GPT-2 distractor word surprisal yield slightly more

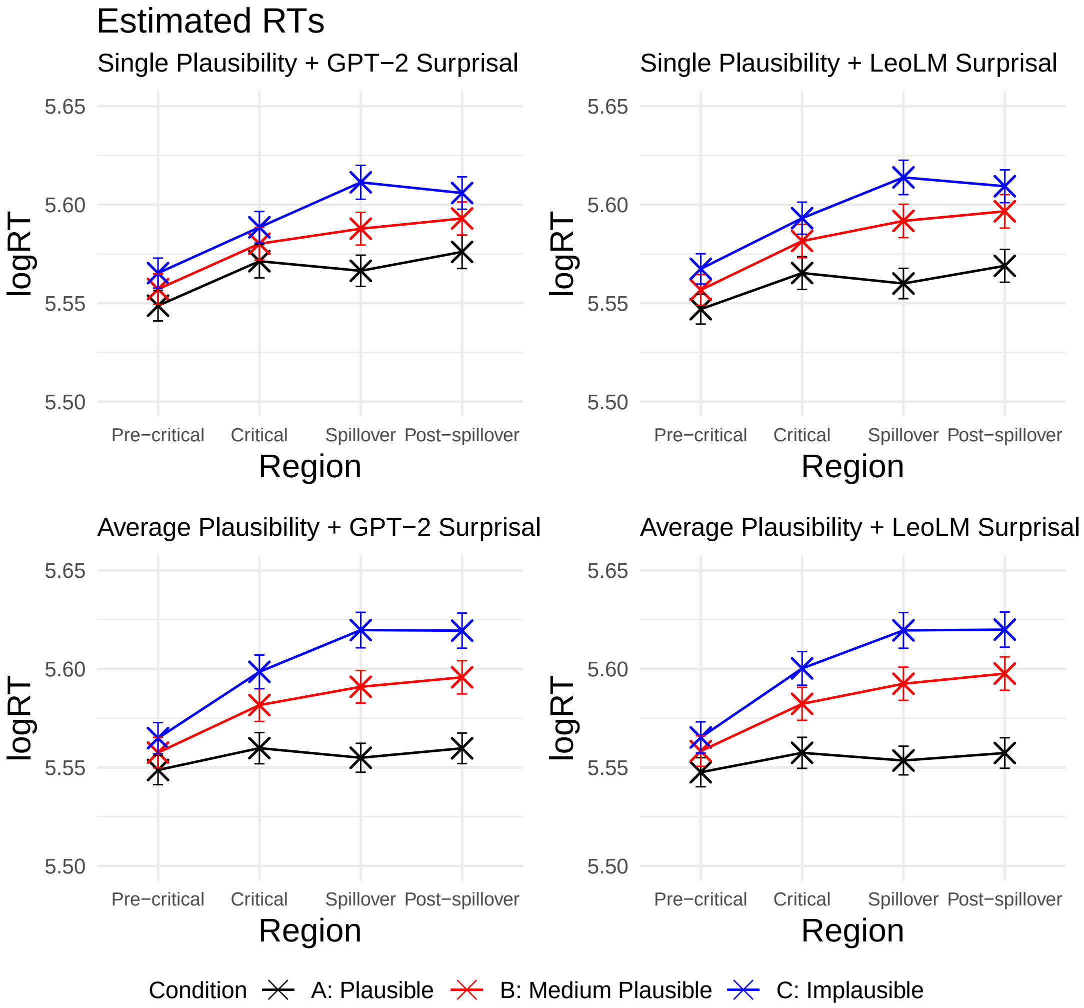

FIGURE 6.3: Estimated log reading times using the predictors single-trial plausibility and GPT-2 surprisal (top left), single-trial plausibility and LeoLM surprisal (top right), averaged plausibility and GPT-2 surprisal (bottom left) and averaged plausibility and LeoLM surprisal (bottom right) per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

accurate predictions for RTs in Condition C, whereas the models incorporating LeoLM distractor word surprisal explain the RT data in Conditions A and B slightly better. In contrast, the two models including single-trial target word plausibility seem to capture the effects structure in the observed RT data less well, as evidenced by a larger residual error. Specifically, the model including single-trial target word plausibility and GPT-2 surprisal overestimates the RTs in Condition A, as shown by the relatively large residual error in the negative range. When LeoLM surprisal is combined with single-trial target word plausibility, the estimated RTs and residuals for Conditions B and C are similar, but the prediction of the RTs in Condition A is improved.

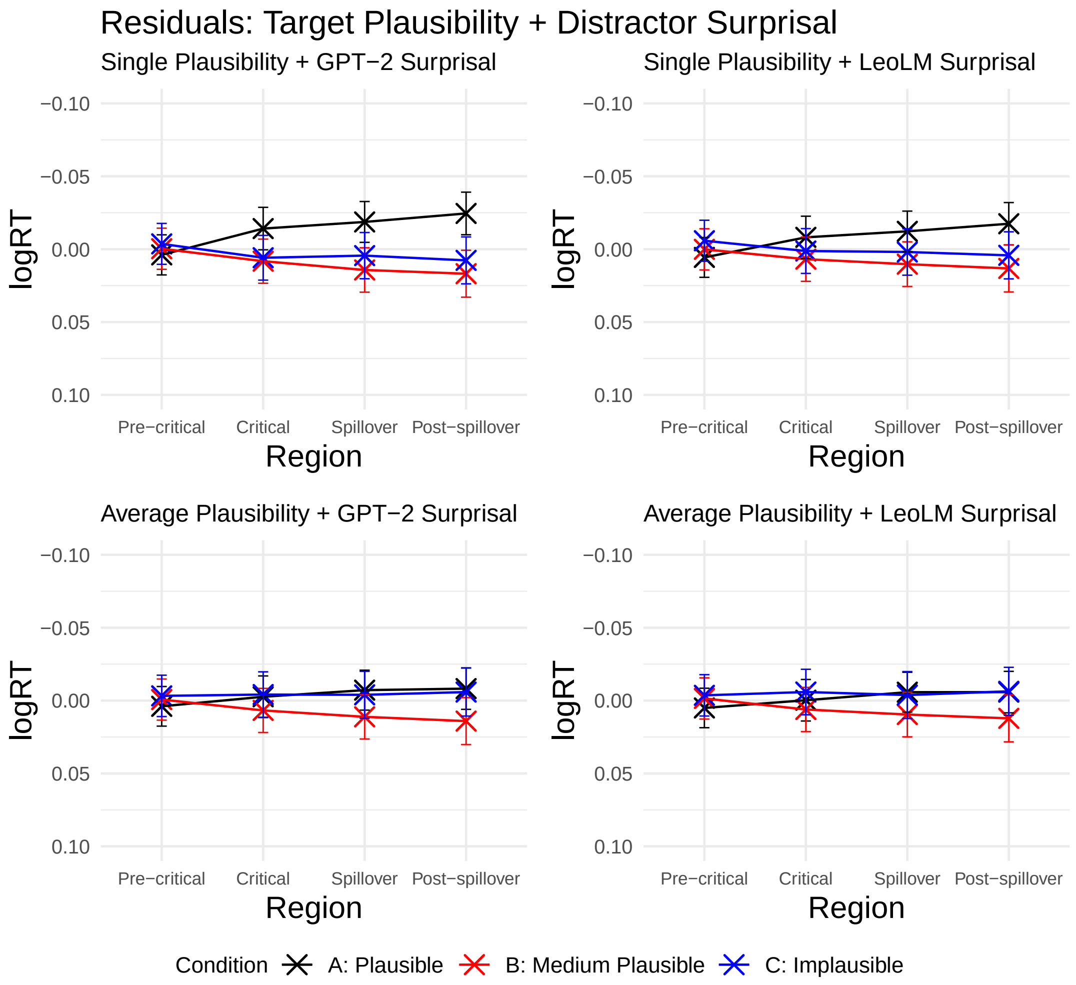

FIGURE 6.4: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

Finally, the same models were fitted using the online plausibility ratings averaged across items and subjects (see Appendix B.1), similar to the offline plausibility ratings. The resulting residuals are smaller than those of the models fitted with single-trial plausibility, but larger, especially in Condition A, than those of the models fitted with averaged pre-test plausibility. This suggests that averaging single-trial plausibility ratings over subjects makes them better predictors of RTs than the original single-trial ratings. However, even when averaged, they appear to capture the effect structure in the observed RTs less effectively than the averaged pre-test plausibility ratings.

Figure 6.5 shows the model coefficients, added to their intercept for single-trial and averaged target word plausibility as well as for GPT-2 and LeoLM distractor word surprisal. The coefficients for both single-trial target word plausibility and averaged target word plausibility are positive in all regions, indicating that lower plausibility predicts slower reading. In contrast, the coefficients for GPT-2 distractor word surprisal and LeoLM distractor word surprisal are negative in all regions, suggesting that less surprising (more predictable) words are read slower. Negative surprisal coefficients may seem counterintuitive, as lower surprisal (higher

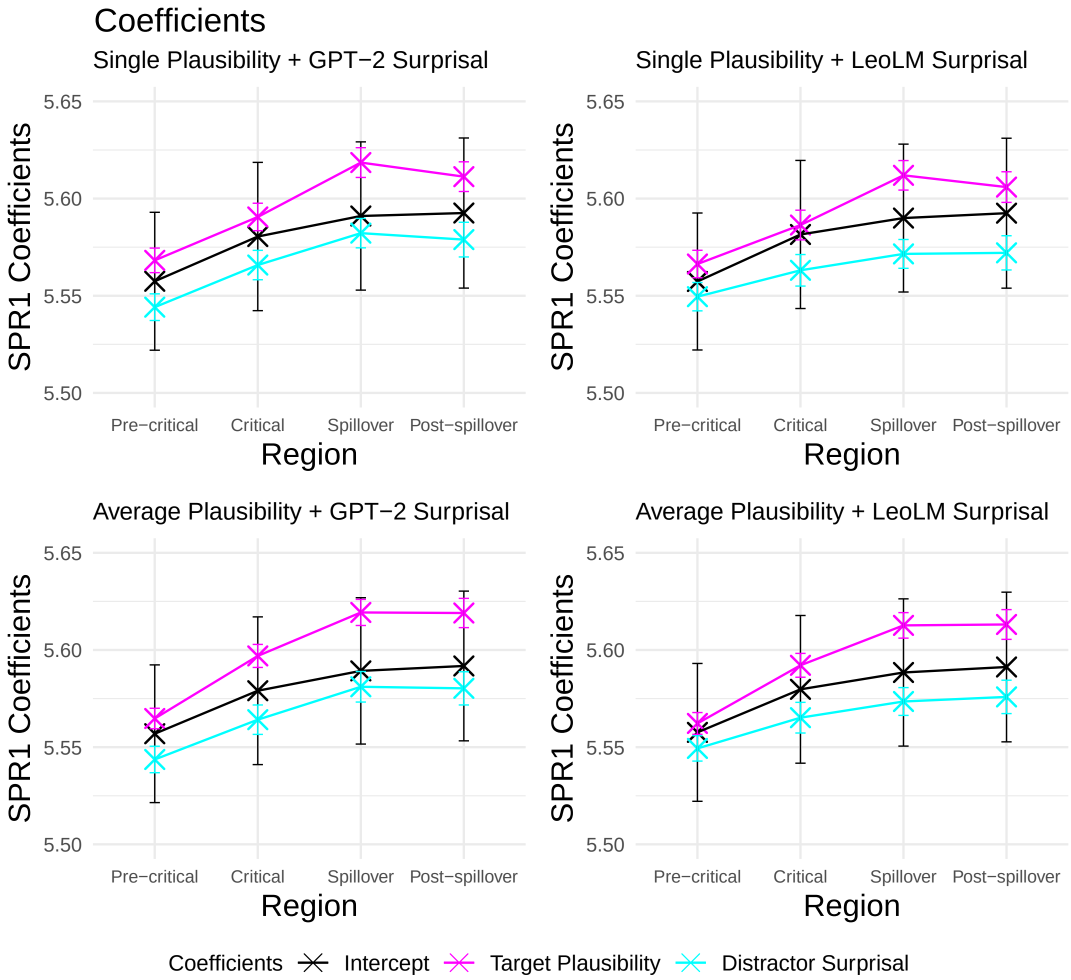

FIGURE 6.5: Coefficients, added to their intercept, from the models fitted with four different predictor combinations. Error bars indicate the standard error of the coefficients in the fitted statistical models.

predictability) would typically be expected to result in faster RTs. If the expectancy of the distractor word was high in Condition B, as it was the case in the study by Aurnhammer et al. (2023), or in any other condition, the increased processing cost of an expected but not presented word could indeed lead to slower reading. However, as distractor word expectancy was low in all conditions, it was expected that lower surprisal would lead to faster reading. The results can be explained by the patterns of average surprisal per condition ($C > A > B$ for GPT-2 and $A > B > C$ for LeoLM; see chapter 5.2.2), which indicate that items in Condition A are on average less expected (although they are on average more plausible and lower plausibility is usually associated with lower expectancy) than items in Conditions B and C. Thus, both lower plausibility and lower surprisal (higher predictability) simultaneously predict slower RTs.

Figure 6.6 shows the z-values associated with each fixed effect coefficient and whether or not a predictor was significant in that region. The p-values are presented in Table B.1. The z-values indicate how many standard deviations the estimated

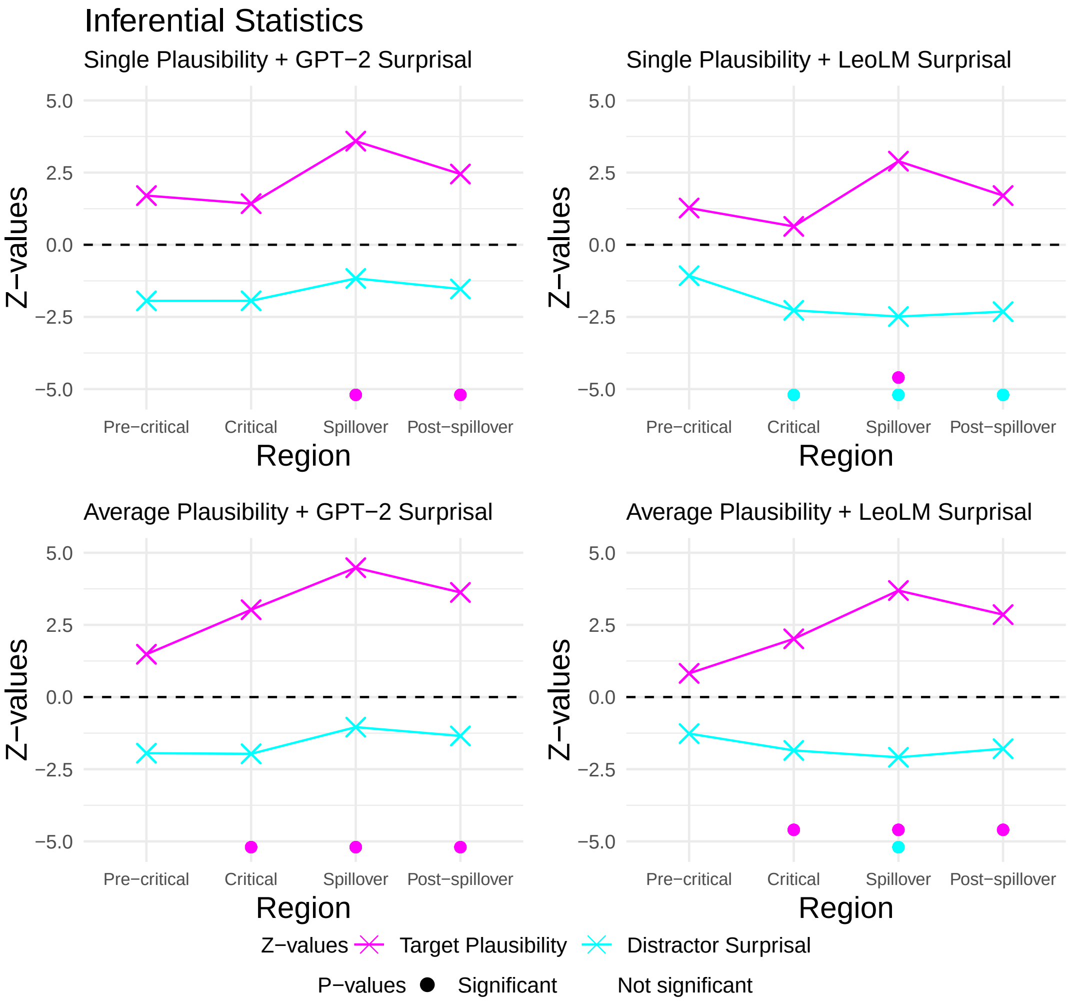

FIGURE 6.6: Effect sizes (z-values) and p-values from the models fitted with four different predictor combinations.

coefficients deviate from zero, i.e., from the mean, serving as a measure of the statistical significance of the effect of a predictor variable on the dependent variable (RTs in this case). Essentially, a larger (either positive or negative) absolute z-score indicates stronger evidence against the null hypothesis (that the coefficient is zero), potentially indicating statistical significance at the chosen significance level, which in the case of this study is 0.05[1].

As the z-values and p-values associated with the coefficients for single-trial target word plausibility and GPT-2 distractor word surprisal show, target word plausibility is a significant predictor of RTs in the Spillover and Post-spillover regions, whereas it is not significant in the Pre-critical and Critical regions. Distractor word surprisal does not significantly predict RTs in any region. In contrast, in the model fitted with single-trial target word plausibility and LeoLM distractor word

[1] In fact, the significance level is 0.025 for each tail given that both the positive and negative range of the distribution is taken into account.

surprisal, target word plausibility only significantly predicts RT in the Spillover region, whereas distractor word surprisal is significant in the Critical, Spillover and Post-spillover regions. In the two models including averaged target word plausibility as a predictor, target word plausibility is a significant predictor of RTs in the Critical, Spillover and Post-spillover regions. Conversely, GPT-2 distractor word surprisal is again not significant in any region, whereas LeoLM distractor word surprisal significantly influences RTs in the Spillover region, although the respective p-value of approximately 0.04 is relatively close to the chosen significance level of 0.05 (see Table B.1).

Figure 6.2 shows that the observed RTs in Condition C are, on average, already higher than the RTs in Condition B and especially in Condition A in the Pre-critical region. This is probably due to the processing of the preceding main verb, which may already introduce varying levels of plausibility before encountering the target word. To determine whether the RTs in Condition C in the Critical region (i.e., on the target word) and in the following regions differ due to the plausibility of the target word itself or due to variations in the contexts leading up to it, the same models were fitted, including Pre-critical RT as a third predictor (see Aurnhammer et al., 2023). By including Pre-critical RT in the analysis, the effects of target word plausibility and distractor word expectancy can be isolated from the influence of the preceding context. Now, the plausibility and surprisal predictors account for any additional RT variation beyond the RT differences in the Pre-critical region due to differences in the main verbs. Before fitting the models, the Pre-critical RT predictor was standardised but not log-transformed to ensure that it remained distinct from the log-transformed dependent variable, which also included the Pre-critical RT values.

The resulting coefficients are shown in Figure 6.7. As shown by the positive coefficient for Pre-critical RT across all regions, words that were read slower in the Pre-critical region were also read slower in all subsequent regions. Furthermore, the z- and p-values in Figure 6.8 show that Pre-critical RT is a significant predictor of RTs in all predictor combinations, i.e., when single-trial or averaged target word plausibility is used together with GPT-2 or LeoLM distractor word surprisal. In the model including single-trial plausibility and GPT-2 surprisal, plausibility remains a significant predictor of RTs in the Spillover and Post-spillover regions, over and above what is explained by the Pre-critical RT predictor, while surprisal is still not significant in any region. Using single-trial plausibility and LeoLM surprisal as predictors, plausibility still significantly predicts RTs in the Spillover region, but no longer in the Post-spillover region, while surprisal is still significant in the Spillover and Post-spillover regions, but no longer in the Critical region. In the two models including averaged target word plausibility, plausibility is still significant in the same regions as before, which includes the Critical, Spillover and Post-spillover regions. However, in the model including LeoLM surprisal, surprisal is no longer significant in the Spillover region, which was previously the only region where it had a significant influence on RTs. However, it should be noted that p-values were not corrected for multiple comparisons, which increases the likelihood of a false positive result (type I error), i.e. the likelihood of observing a significant effect or difference when there is none.

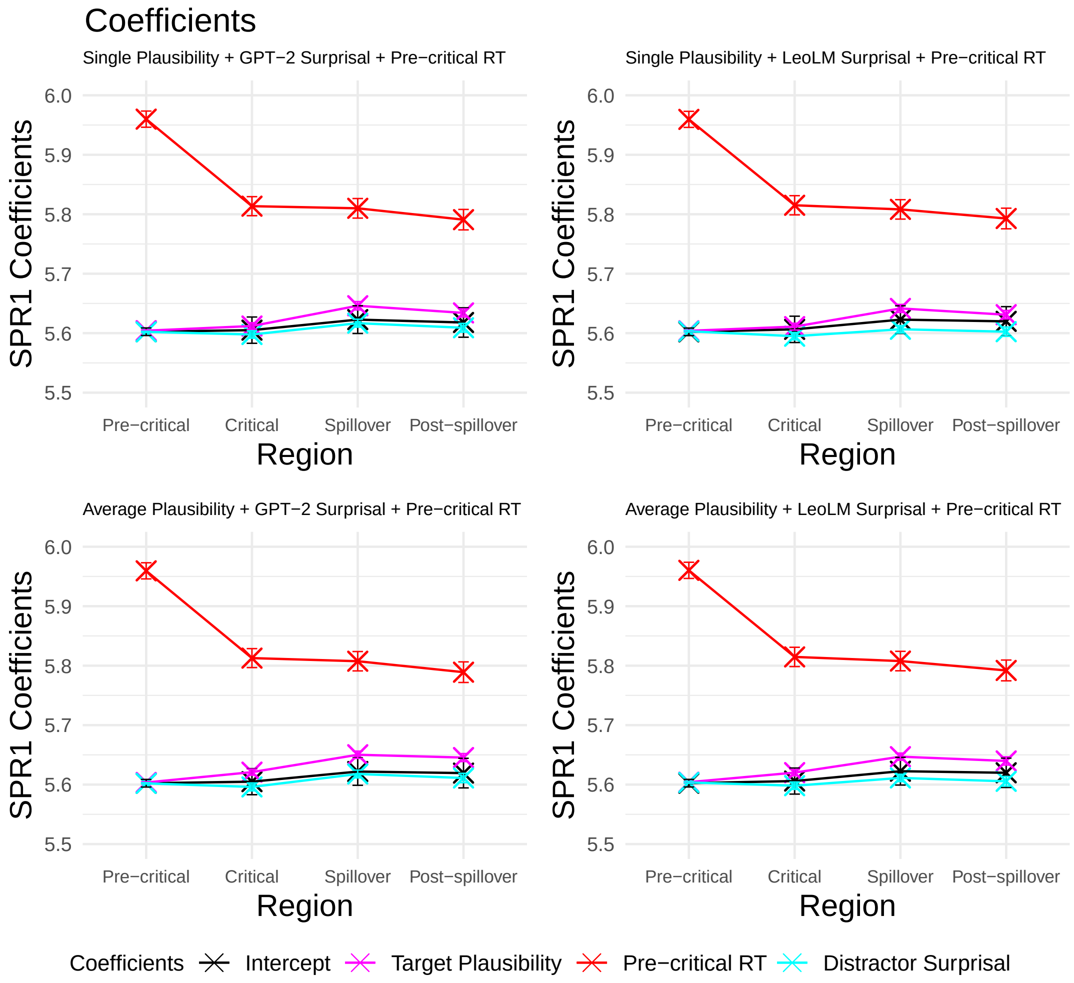

FIGURE 6.7: Coefficients, added to their intercept, from the models including Pre-critical reading time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

6.4.4 Model Comparison

Since offline target word plausibility proved to be a better predictor of RTs than online target word plausibility, as evidenced by smaller residuals (see Figure 6.4), the question arises whether adding single-trial target word plausibility as an additional predictor, i.e., including both averaged and single-trial target word plausibility, significantly improves model fit. A likelihood ratio test (LRT) was performed to compare the goodness of fit between a complex model, in this case including both offline and single-trial target word plausibility, and a simple model, including only offline target word plausibility. The LRT determines whether the complex model (alternative hypothesis) significantly improves the fit to the data compared to the simple model (null hypothesis) by comparing the observed test statistic to a critical value from the chi-squared distribution, where the degrees of freedom are equal to the difference in the number of parameters between the complex and the simple models. If the test statistic exceeds the critical value from the chi-squared distribution

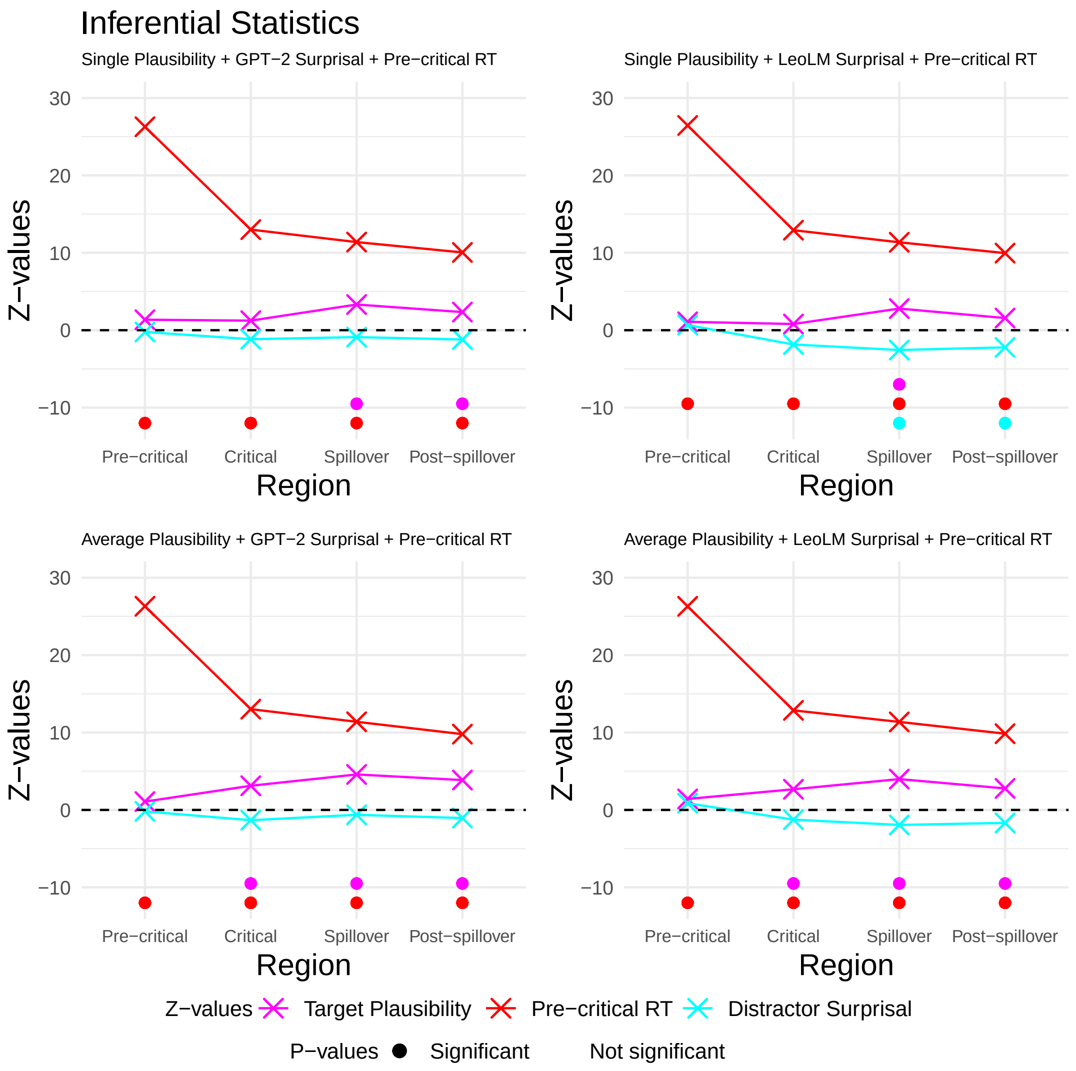

FIGURE 6.8: Effect sizes (z-values) and p-values from the models including including Pre-critical reading time as a predictor.

for the given significance level and degrees of freedom, the null hypothesis is rejected, indicating that the more complex model significantly improves the model fit. The results of the LRT are reported in Table 6.4.

Although only offline and single-trial target word plausibility are reported in Table 6.4, GPT-2 distractor word surprisal was also used as a predictor in the simple and complex models. The LRT was also performed with LeoLM surprisal instead of GPT-2 surprisal; however, as the results were not significantly different, only the results obtained with GPT-2 are reported. The relatively high chi-squared values and low p-values (< 0.05) indicate that the complex model provides a significantly better fit to the RT data than the simple model, which is also evidenced by the lower AIC values for the complex model in all regions except for the Spillover region. These results suggest that while single-trial plausibility ratings alone yield a less accurate estimate of RTs than averaged plausibility ratings, they capture additional variability that improves the overall model fit when combined with the offline plausibility ratings. The reason why the BIC value contrasts with other metrics – shown by the lower BIC value in all regions for the simple model, suggesting a better fit to the RT data than achieved by the complex model – is not entirely clear. It may be related the property of the BIC to penalise model complexity more severely, as the penalty term grows faster with the number of parameters and sample size.

		AIC	BIC	Chi-Sq.	P-value
Pre-critical	Avg. Plausibility	396.00	489.27		
	Avg. + Single Plausibility	388.71	543.44	25.29	0.00267
Critical	Avg. Plausibility	774.41	867.67		
	Avg. + Single Plausibility	760.70	906.43	31.70	<0.001
Spillover	Avg. Plausibility	796.16	889.43		
	Avg. + Single Plausibility	800.41	946.14	13.75	0.132
Post-spillover	Avg. Plausibility	1169.0	1262.2		
	Avg. + Single Plausibility	1147.9	1293.6	39.10	<0.001

TABLE 6.4: Results from the likelihood ratio test via ANOVA for model comparison.

Since the LRT indicated that including both single-trial and averaged target word plausibility as predictors significantly improves the model fit, a separate model including both plausibility predictors and GPT-2 surprisal was fitted for each critical region. The resulting estimated RTs and residuals are shown in Figure 6.9. The same models were also fitted using LeoLM surprisal instead of GPT-2 surprisal. However, the figures of the estimated RTs and residuals are not included, as visual inspection did not reveal any differences to the results obtained when GPT-2 surprisal was used. Furthermore, the estimated RTs and residuals obtained from the complex model show no observable differences compared to the estimates and residuals obtained when RTs were predicted based on averaged target word plausibility and GPT-2 or LeoLM distractor word surprisal alone (see Figure 6.3; bottom left and Figure 6.4; bottom right).

Figure 6.10 shows the coefficients added to their intercept (left) and z- and p-values (right) for the models including single-trial and averaged target word plausibility as well as GPT-2 (top) or LeoLM (bottom) distractor word surprisal, as the significance of the predictors differs across regions depending on whether the models were fitted with GPT-2 or LeoLM distractor word surprisal as a third predictor. The coefficients for both GPT-2 and LeoLM distractor word surprisal are still negative in all regions, indicating that lower surprisal (higher predictability) predicts slower reading. Similarly, the model coefficients for averaged target word plausibility are still positive in all regions, suggesting that lower plausibility predicts slower RTs, except in the Pre-critical region of the model fitted with LeoLM distractor word surprisal, where the coefficient for target word plausibility is now negative. The coefficient for single-trial target word plausibility changes sign depending on the region: It is positive in the Pre-critical and the Spillover regions and negative in the Critical and Post-spillover regions, suggesting that it predicts slower or faster RTs depending on the region of interest. The fact that single-trial plausibility does

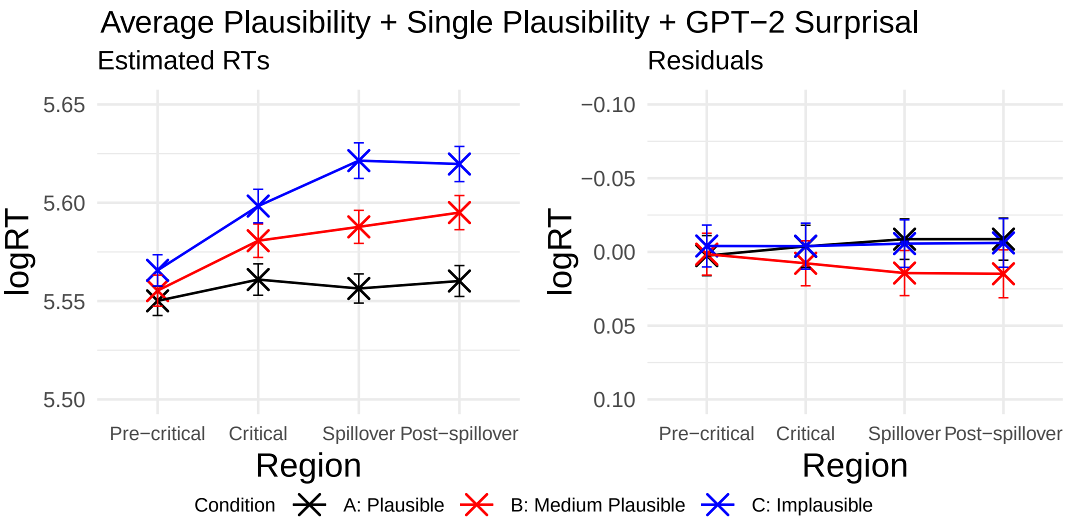

FIGURE 6.9: Estimated log reading times (left) and residuals (right) obtained from the models including averaged (offline) plausibility, single-trial (online) plausibility and GPT-2 surprisal as predictors per condition on the four critical regions.

not predict the same direction of the effect across regions suggests that it does not account for the plausibility effect, which is roughly the same across regions.

The z- and p-values in Figure 6.10 show that when the models are fitted with single-trial target word plausibility, averaged target word plausibility and GPT-2 distractor word surprisal, averaged target word plausibility significantly predicts RTs only in the Spillover and Post-spillover regions, while single-trial plausibility is not a significant predictor of RTs in any of the four critical regions. At the same time, GPT-2 Surprisal significantly influences RTs in the Spillover region, although it was previously not significant in combination with either averaged or single-trial target word plausibility. In contrast, in the models fitted with single-trial plausibility, averaged plausibility and LeoLM surprisal, neither single-trial nor averaged plausibility is significant in any region. The only significant predictor in this model is LeoLM distractor word surprisal, which, similarly to the model fitted with averaged plausibility and LeoLM surprisal alone, is only significant in the Spillover region. The corresponding p-values are shown in Table B.1.

6.5 Discussion

The results of the pre-tests (see Chapters 5.1.2 and 5.2.2) showed that the expectancy of the distractor word was successfully lowered in Condition B, resulting in a lower expectancy of the distractor word compared to the target word in all conditions[2], while the three plausibility levels were maintained $(A > B > C)$. The results of the self-paced reading study demonstrated that RTs gradually increase as plausibility decreases, such that items were read slower on average the less plausible they

[2] Except for the surprisal values from LeoLM in Condition C, which are on average slightly higher for the target word than for the distractor word.

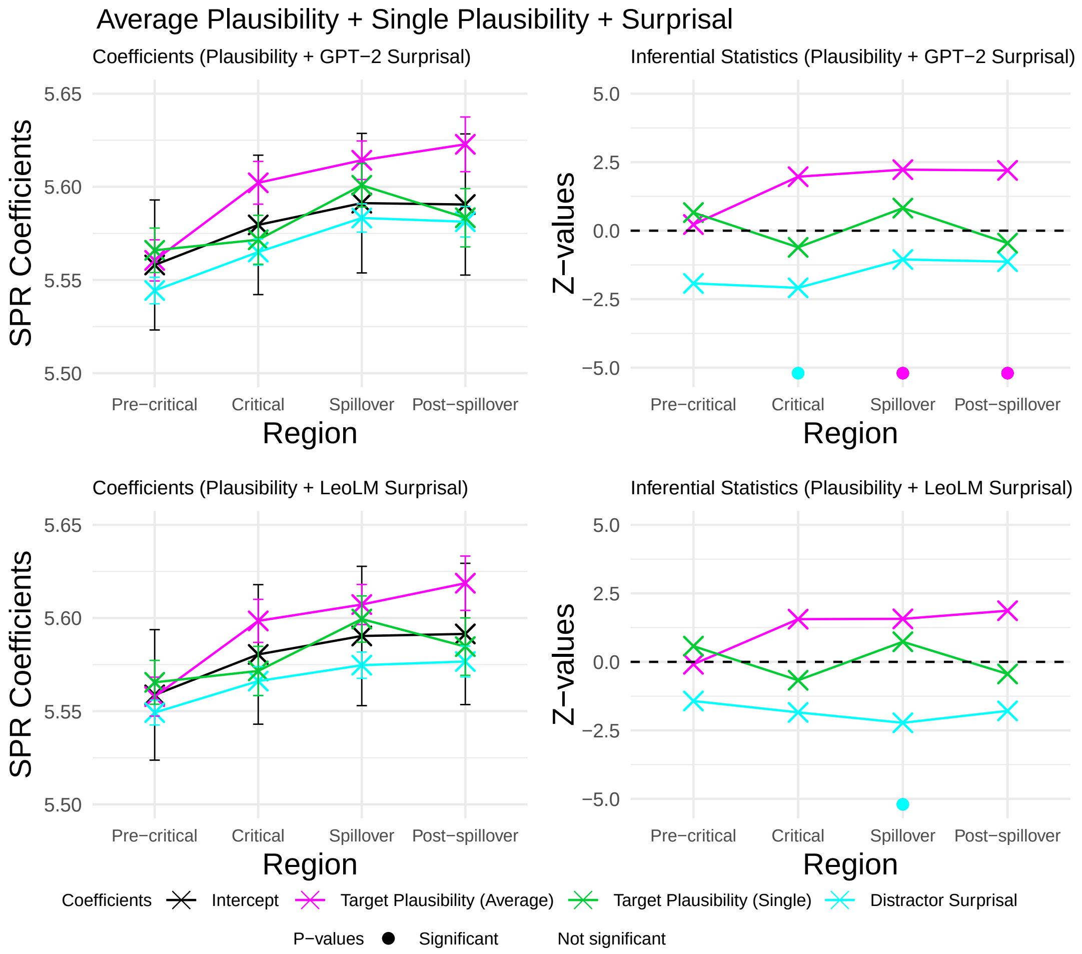

FIGURE 6.10: Coefficients, added to their intercept, (left) and z- and p-values (right) obtained from the models including averaged (offline) plausibility, single-trial (online) plausibility and GPT-2 surprisal (top) or LeoLM surprisal (bottom) as predictors per condition on the four critical regions.

were (see Figure 6.2). However, in contrast to the RTs observed in the study by Aurnhammer et al. (2023), which formed a consistent pattern with Conditions A, B and C, the RTs observed in the current self-paced reading study did not increase as strongly after the Critical region, i.e., after reading the target word, and there was only a small difference in RTs between Conditions B and C. A possible explanation for the similarity in RTs between Conditions B and C could be that the modification of Condition B led to lower plausibility, resulting in relatively high RTs. However, the results of the plausibility pre-test and the online plausibility rating task indicated that plausibility is evenly graded in Conditions A, B and C, which is inconsistent with this assumption. The alternative interpretation that the items in Condition C were too plausible and led to relatively low RTs is also unlikely, given that Condition C was not modified for the current study and in the original study by Aurnhammer et al. (2023), RTs in Condition C were notably higher than those in Condition B. Therefore, it may have been the combination of the self-paced reading study and the online plausibility rating task that influenced the RTs, as participants read the word-by-word presented final sentence in anticipation of the upcoming rating task. To investigate whether this was the case, or to identify other potential factors that might have contributed to the observed RT pattern in the self-paced reading study, post-hoc analyses were carried out, which are described in Appendix B.3 - B.6.

The regression-based analysis showed that both single-trial plausibility ratings and averaged plausibility ratings from the pre-test continuously predict RTs. However, single-trial plausibility was a significant predictor of RTs either in the Spillover region only or in both the Spillover and Post-spillover regions, depending on whether GPT-2 or LeoLM surprisal was used as the second predictor, whereas averaged plausibility was a significant predictor in the Critical, Spillover and Post-spillover regions in both combinations with surprisal and even after accounting for pre-critical RT differences (see Figure 6.8). Interestingly, the averaged plausibility ratings from the pre-test predicted RTs better than the single-trial plausibility ratings collected online, as evidenced by more accurate estimates and smaller residuals (see Figures 6.3 and 6.4). This may be surprising, as it would be reasonable to assume that single-trial plausibility ratings would be a better predictor of RTs due to their ability to capture individual differences and variations, allowing for a more granular analysis. A possible reason may be that, precisely because of this granularity, single-trial ratings are disproportionately affected by outliers, which may affect the overall analysis. Even after removing outliers based on extremely low or high reading or reaction times, the single-trial plausibility ratings can fluctuate from trial to trial within each participant due to individual differences in perception, i.e., some participants may be more lenient and others stricter in their judgments, as well as factors such as decreasing attention or inconsistent use of the rating scale. In contrast, the averaged plausibility ratings are based on the responses of multiple participants, i.e., they reflect the overall perceived plausibility of each item and therefore smooth out individual variation. This robustness to variability may be what ultimately makes them more effective predictors of RTs.

Further analyses (see Appendix B.1) showed that the averaged single-trial plausibility ratings predicted the RTs better than the single-trial plausibility ratings but worse than the averaged plausibility ratings collected in the pre-test. Surprisingly, this shows that the averaged plausibility ratings from the pre-test provide a better fit to the RT data than the plausibility ratings collected online, beyond the higher predictive power of the average, despite the fact that the former were collected from a different set of participants than the RTs. The reason for this may be that the combination of self-paced reading and the plausibility rating task may have not only affected the RTs, but also the single-trial plausibility ratings, and therefore the average of these ratings, albeit to a lesser extent due to their relative stability. An important difference between the online and offline plausibility rating tasks was that for the online plausibility rating task, the plausibility scale could not be presented along with the context paragraph and the final sentence due to the design of the self-paced reading study, which prevented participants from revisiting previous sections before assigning a rating. This may have influenced the online plausibility ratings. Therefore, the reason why the averaged plausibility ratings per condition from the online study were more similar to each other (see Table 6.2) than those from the plausibility pre-test, may be that participants could only provide ratings based on the content they remembered, which may have made them more hesitant to choose extreme ratings (very plausible or very implausible) in the online rating task.

In terms of surprisal, using LeoLM distractor word surprisal instead of GPT-2 surprisal in the models with single-trial plausibility reduced the residual error across conditions, but did not lead to an observable difference when combined with the more effective averaged target word plausibility predictor. Furthermore, the results showed no significant effects of GPT-2 distractor word surprisal on RTs when combined with either single-trial or averaged target word plausibility. This was expected, as GPT-2 surprisal was on average higher, i.e., expectancy was lower, for the distractor word than for the target word across all conditions (see section 5.2.2). However, even if distractor word surprisal were lower (i.e., distractor word expectancy were higher) than target word surprisal, this would not necessarily result in a significant RT modulation of distractor word surprisal. The results of Aurnhammer et al. (2023) showed no significant effect of distractor word cloze probability on RTs, even though in their design distractor word expectancy was higher than target word expectancy in Condition B, suggesting that RTs may not be affected by unfulfilled expectations.

Interestingly, LeoLM distractor word surprisal significantly predicted RTs in the Spillover region when used together with averaged target word plausibility and was additionally significant in the Critical and Post-spillover regions when used with single-trial target word plausibility. After including Pre-critical RT as a predictor to account for pre-critical RT differences, LeoLM distractor word surprisal was no longer significant when used together with averaged target word plausibility, but was still significant in the Spillover and Post-spillover regions in the models using single-trial target word plausibility. The reason for this is not entirely clear but it may be related to the pattern observed for the LeoLM distractor word surprisal (see section 5.2.2), which is the same as the pattern observed for averaged and single-trial target word plausibility ($A > B > C$). However, the reason could also be that p-values were not corrected for multiple comparisons, which increases the likelihood of a false positive result (Type I error), i.e., the likelihood of observing a significant effect or difference when, in reality, there is none.

Finally, the results of a LRT showed that the complex model, including both single-trial and averaged target word plausibility (along with distractor word surprisal), provided a better fit to the RT data than the simple model, including only averaged target word plausibility (and distractor word surprisal), in all regions except the Spillover region (see Table 6.4). Surprisingly, including averaged and single-trial target word plausibility along with distractor word surprisal did not improve the RT predictions, as indicated by slightly larger residual errors (see Figure 6.9). The difference in prediction accuracy is small and cannot be detected by visual inspection, but examination of the residual error values revealed that in ten out of twelve models (one model per region for each condition) the residuals of the complex model were slightly larger than those of the simple model. This suggests that although the single-trial plausibility ratings account for variation in RTs beyond the variation explained by the averaged pre-test plausibility ratings, this improvement does not explain the effects observed between Conditions A, B and C better. In other words, the LRT only indicates that the single-trial plausibility ratings improve the model fit in general, which does not imply that the complex model also better captures the RTs grouped by Conditions A, B, and C better.

Chapter 7

Self-Paced Reading Study II

Additional analyses (see Appendix B.5) examining the RT data from the first self-paced reading study revealed that when RT were grouped into three equal groups based on the assigned plausibility ratings rather than based on Conditions A, B and C, items rated as medium plausible (i.e., assigned a plausibility rating of 3, 4 or 5 on a seven-point Likert scale) had higher RTs on average than items rated as plausible (i.e., assigned a plausibility rating of 5, 6 or 7) or items rated as implausible (i.e., assigned a plausibility rating of 1, 2 or 3). A possible explanation is that items rated as medium plausible are inherently more difficult to rate than highly plausible or implausible items, and as participants anticipate the upcoming rating task, the increased difficulty of judging medium plausible items may be reflected in increased RTs. Since the majority of items rated as medium plausible belong to Condition B, and the observed RTs from the first self-paced reading study (see Figure 6.2) were similarly high for Conditions B and C, this suggests that the combination of the online plausibility rating task with the self-paced reading study may have distorted the RTs. More specifically, the RTs observed for items in Condition B may have been higher and the RT pattern $A < B < C$ less pronounced than expected because most items in Condition B are perceived as medium plausible and are therefore more difficult to rate. To confirm whether the similarity in RTs between Conditions B and C in the first self-paced reading study was due to the online rating task affecting the RTs, a second self-paced reading study was conducted in which the plausibility rating task was removed. Otherwise, the materials were identical to those used in the first self-paced reading study. Thus, the procedure of the second self-paced reading study is identical to the procedure in the study by Aurnhammer et al. (2023). The only difference is the expectancy of the distractor word, which is low in all conditions in the current study and high only in Condition B in the study by Aurnhammer et al. (2023).

7.1 Participants

Participants were again recruited via Prolific Academic Ltd., to take part in the second self-paced reading study, which was conducted on the PCIbex platform. A total of forty-five participants were recruited, but data from three participants were excluded from all statistical analyses due to inattentive reading, as indicated by low response accuracy (less than 70% correct) on the comprehension questions. This exclusion threshold was slightly lower than the threshold used in the first self-paced reading study (less than 80% correct). Consequently, data from five participants, who answered less than 80% (but more than 70%) of the questions correctly, were not excluded in the current self-paced reading study. The remaining forty-two participants (mean age 24.83; SD 2.9; age range 19-31; 16 male, 26 female) were all native German speakers (including seven early bilinguals) who did not report any language-related disorders or literacy difficulties. To ensure that participants had no previous exposure to the study materials, participants who had taken part in the first self-paced reading study, the plausibility pre-test or any of the equivalent studies in Aurnhammer et al. (2023) were excluded from this self-paced reading study.

7.2 Procedure

The materials and procedure were the same as in the first self-paced reading study (see Chapter 6.2). However, unlike the first self-paced reading study, participants were not asked to judge the plausibility of the word-by-word presented final sentence given the preceding context paragraph. Instead, they moved directly to the comprehension question, if there was one, or to the next item if there wasn't one.

7.3 Analysis

The data was analysed using the same linear mixed effects regression re-estimation technique (see also Aurnhammer et al., 2021, 2023) that was used in the first self-paced reading study (see Chapter 6). The procedure and all data pre-processing steps are described in detail in Chapter 4.3. Trials were excluded if the reading time on any of the four critical regions was lower than 50 ms or higher than 2500 ms and if the reaction time on the task, i.e., on the comprehension question, if there was one, was lower than 50 ms or higher than 10,000 ms. Based on these criteria, 58 out of 2520 trials (2.3%) were excluded.

7.4 Results

The results of the comprehension questions, as well as the observed RTs and their statistical analyses are described in the following subsections.

7.4.1 Comprehension Questions

Participants answered comprehension questions for almost half of the experimental items (46% of all trials) and on two-fifths of the filler items. Descriptive statistics for response accuracy and reaction time on the comprehension questions were calculated across subjects. The mean accuracy was 91.4% (SD = 8.0, range = 70.0% - 100.0%) and mean reaction time was 3196 ms (SD = 737, range = 1768 ms - 5362 ms). The mean response accuracies and response times per condition are shown in Table 7.1. Response accuracy was highest in Condition A (94.7%), followed by Condition B (91.0%), and Condition C (89.1%), while mean reaction times were highest in Condition C (3296 ms), followed by Condition A (3143 ms) and Condition B (3142). The mean response accuracies across conditions in the current study were slightly lower than in the first study, probably due to the lower exclusion threshold in the second study (less than 70% correct) compared to the first study (less than 80%

correct). At the same time, the average reaction times in the second study were higher than those in the first study. One reason for this could be that the first self-paced reading study included a rating task, which participants may have spent more time on.

	Accuracy			Reaction Time		
Condition	Mean	SD	Range	Mean	SD	Range
A	94.7%	9.0	60.0% - 100.0%	3143 ms	851	1881 ms - 6112 ms
B	91.0%	11.3	60.0% - 100.0%	3142 ms	783	1652 ms - 4620 ms
C	89.1%	10.6	60.0% - 100.0%	3296 ms	823	1665 ms - 5503 ms

TABLE 7.1: Task performance on the comprehension questions in the second self-paced reading study. Accuracy and reaction times were computed across subjects.

7.4.2 Reading Times

The observed log-transformed RTs per condition on the four critical regions are shown in Figure 7.1. Already in the Pre-critical region, RTs differ between conditions, with Condition C being read slightly slower on average than Condition B and especially Condition A. In the Critical region, i.e., on the target word, RTs increase in all three conditions, although to different extents: RTs in Condition C rise more sharply than RTs in Conditions B and A, while Condition B shows a slightly faster increase in RTs than Condition A. In the Post-spillover region, RTs diverge further: RTs in Condition C continue to increase strongly, while RTs in Condition B increase only slighty and RTs in Condition A decrease again. Thus, RTs pattern with the three plausibility levels of Conditions A, B and C, reflecting high, medium and low levels of plausibility, in the Spillover and even more so in the Post-spillover region. The RTs observed in the current study show a more pronounced gradation of RTs for plausibility compared to the RTs observed in the first self-paced reading study (see Figure 6.2), which also patterned with the three levels of Conditions A, B and C, but to a lesser extent as the RTs in Conditions B and C were almost identical. Conversely, in the current study, the RTs in Conditions B and C show notable differences in all regions, except for the Pre-critical region, where RTs are similarly low in all three conditions. Furthermore, the RTs observed in this study, appear to be lower than those observed in the first self-paced reading study. They are also lower than the RTs observed in the self-paced reading study conducted by Aurnhammer et al. (2023), but otherwise show a similar pattern.

Figure 7.2 shows the estimated RTs using averaged target word plausibility and either GPT-2 distractor word surprisal (left) or LeoLM distractor word surprisal (right). Figure 7.3 shows the corresponding residuals, representing the differences between the observed RTs and those predicted by the linear mixed effects models. Visual inspection reveals no differences between the estimates and residuals of the models including GPT-2 distractor word surprisal and those including LeoLM distractor word surprisal. In both cases, the residual error is small in all regions and conditions, especially in Conditions A and C, indicating that the models accurately capture the effects structure in the observed RT data. Similar to the first self-paced reading study, the model estimates in Condition B are slightly less accurate, as

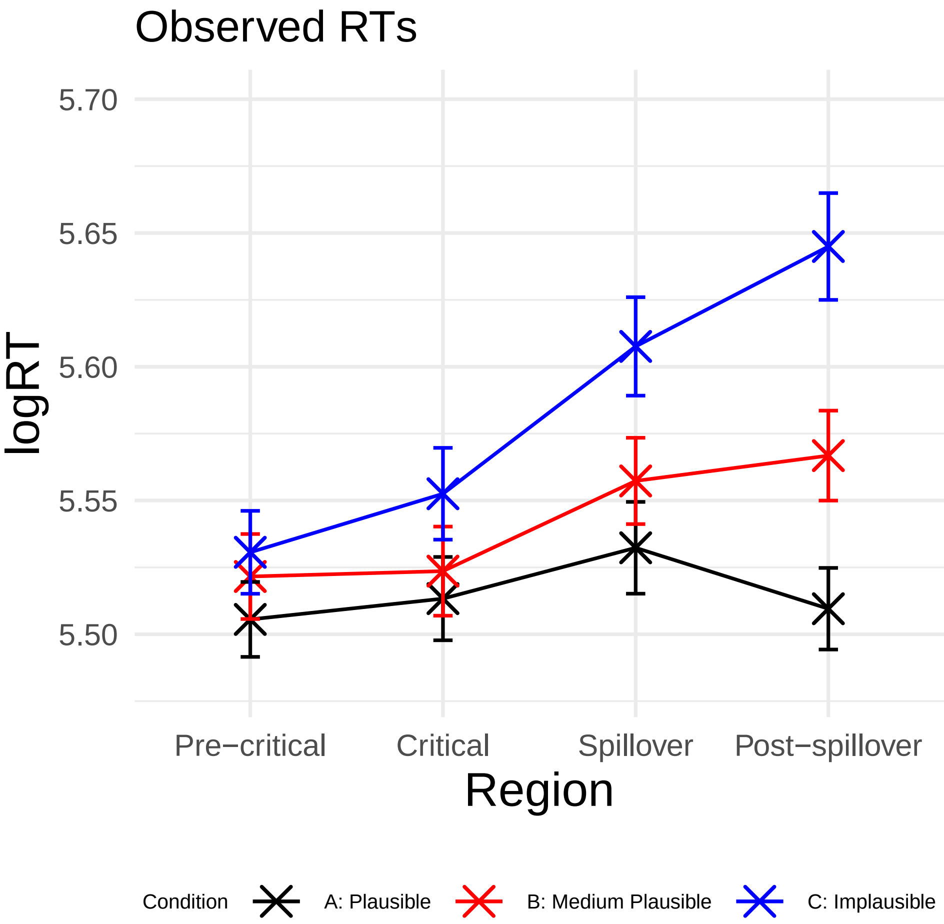

FIGURE 7.1: Log reading times per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions in the second self-paced reading experiment. The error bars show the standard error calculated from the per-subject per-condition averages.

evidenced by a higher residual error. However, the first study showed a positive residual error (indicating that RTs were overestimated), whereas the residual error in the second self-paced reading study in Condition B is negative (indicating that RTs were underestimated).

Figure 7.4 shows the model coefficients, added to their intercept, for averaged target word plausibility and GPT-2 (left) or LeoLM (right) distractor word surprisal, while the corresponding effect sizes (z-values) and significance levels are shown in Figure 7.5. The p-values are reported in Table B.4. The coefficients for averaged target word plausibility obtained from the model including GPT-2 distractor word surprisal and the model including LeoLM distractor word surprisal are positive in all regions. This suggests that, similar to the findings from the first self-paced reading study, lower plausibility predicts slower reading. In contrast, the coefficients for GPT-2 surprisal and LeoLM surprisal are negative in all regions, indicating that lower surprisal (higher predictability) predicts slower RTs.

The relatively large z-values and small p-values (< 0.05) associated with target

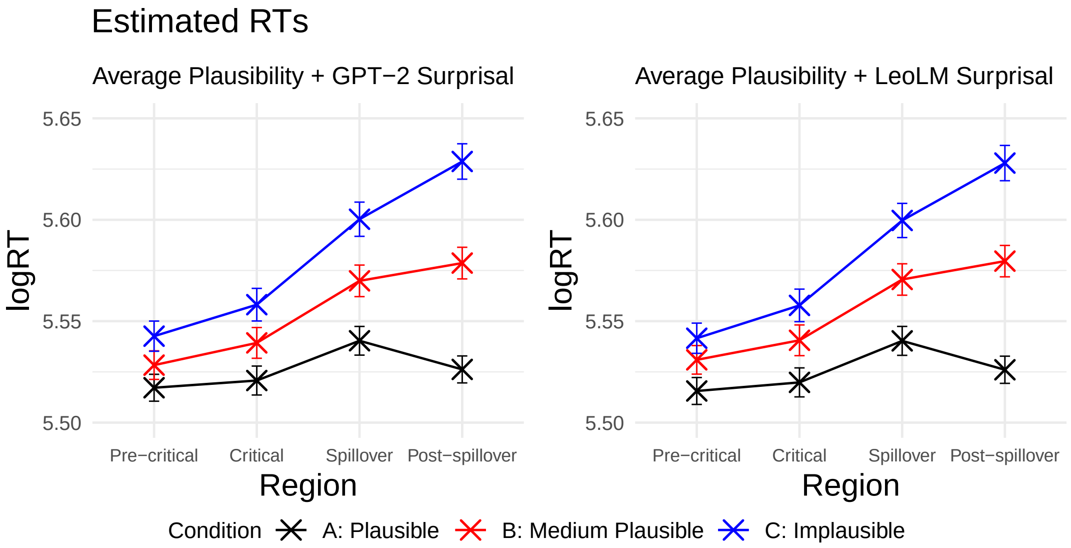

FIGURE 7.2: Estimated log reading times using the predictors averaged plausibility and GPT-2 surprisal (left) and averaged plausibility and LeoLM surprisal (right) per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

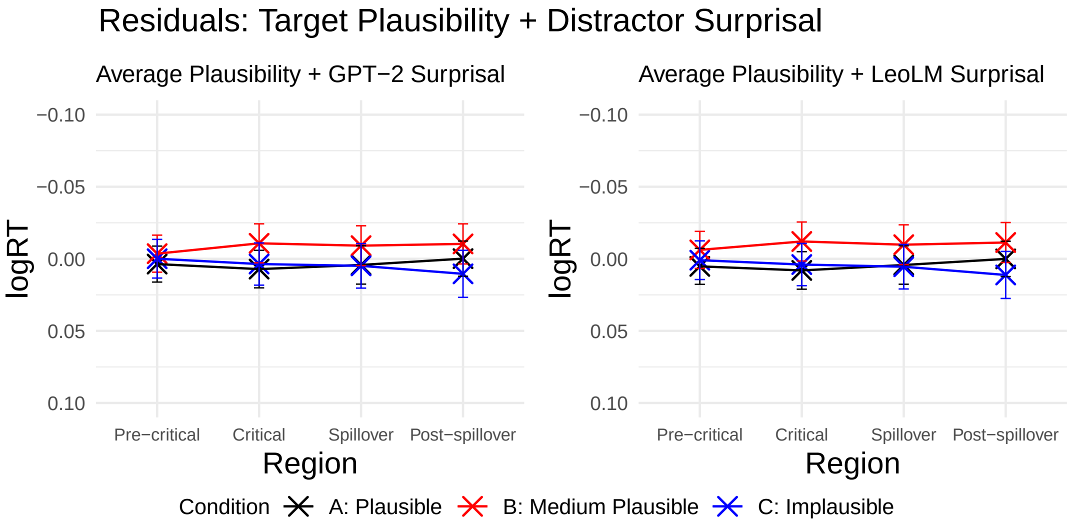

FIGURE 7.3: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

word plausibility indicate that plausibility is significant in both cases, the model including GPT-2 distractor word surprisal as a second predictor and the model including LeoLM distractor word surprisal instead. In the first case, target word plausibility is statistically significant in all four critical regions, whereas in the second case, target word plausibility significantly modulates RTs in all regions except the Pre-critical region. In contrast, neither GPT-2 surprisal nor LeoLM surprisal is significant in any of the four regions, as indicated by small z-values and large p-values (> 0.05). The z-values obtained for averaged target word plausibility in the second self-paced reading study are larger than those obtained in the first study in

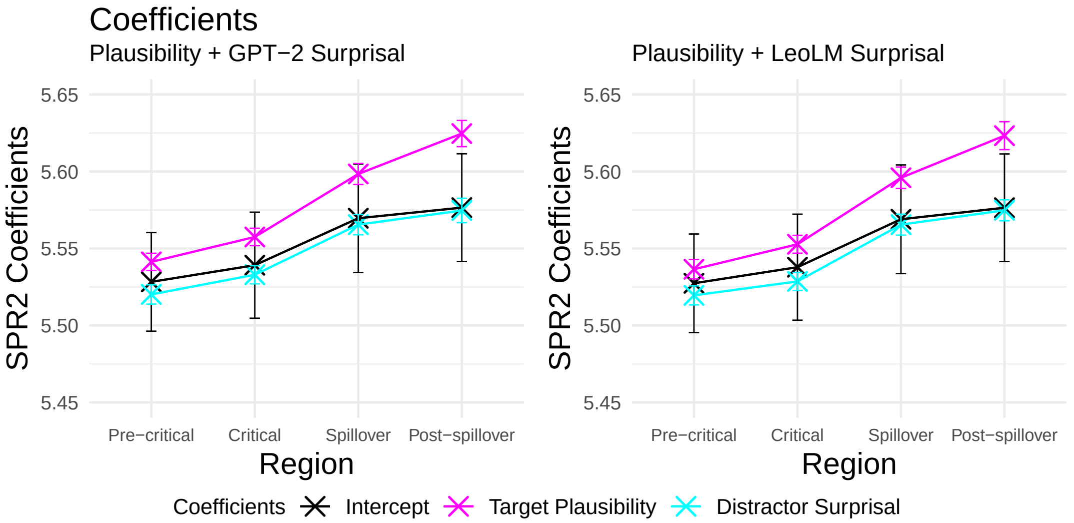

FIGURE 7.4: Coefficients, added to their intercept, for averaged plausibility and GPT-2 surprisal (left) and averaged plausibility and LeoLM surprisal (right). Error bars indicate the standard error of the coefficients in the fitted statistical models.

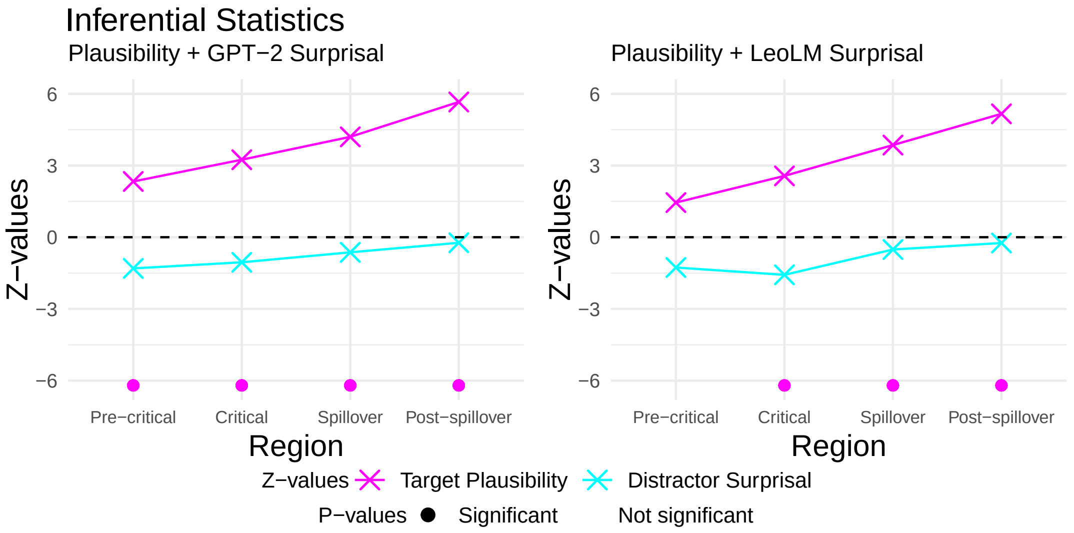

FIGURE 7.5: Effect sizes (z-values) and p-values.

all regions (see Figure 6.6), suggesting that plausibility is likely to have a stronger effect on RTs in the second study than in the first study. In contrast, the z-values for GPT-2 and LeoLM distractor word surprisal in the second self-paced reading study are smaller than those in the first study across all regions, especially in the Spillover and Post-spillover regions, suggesting that surprisal has a smaller (or rather no) effect in the second self-paced reading study than in the first study. A comparison of the p-values (see Tables B.1 and B.4) moreover shows that plausibility has a significant effect on RTs in all regions, whereas surprisal has no significant effect on the RTs in any ragion in the second self-paced reading study. While averaged target word plausibility is a significant predictor of RTs in the Critical, Spillover and Post-spillover regions in the first self-paced reading study, averaged target word plausibility is significant in all four critical regions when combined with GPT-2 distractor word surprisal in the second study. In contrast, GPT-2 distractor word surprisal has no significant effect on RTs in either the first or in the second self-paced reading study. LeoLM distractor word surprisal significantly predicts RTs in the first self-paced reading study in the Critical, Spillover, and Post-spillover regions when combined with single-trial target word plausibility, and in the Spillover region when combined with averaged target word plausibility, but is not significant in any region in the second self-paced reading study.

In order to account for Pre-critical reading time differences, Pre-critical RT was again used as a predictor to determine whether the observed RT differences were due to the plausibility of the target word itself or to the different contexts created by the different main verbs in each condition. After standardising the Pre-critical RT predictor, it was included in the models along with averaged target word plausibility and either GPT-2 or LeoLM surprisal. The coefficients of the corresponding predictors are shown in Figure 7.6 and the corresponding z-values and significance levels are shown in Figure 7.7. The corresponding p-values are reported in Table B.4. As shown by the positive coefficient for Pre-critical RT across all regions, words that were read slower in the Pre-critical region were also read slower in all subsequent regions. The resulting z- and p-values indicate that RTs are significantly predicted by Pre-critical RT across all regions when target word plausibility and GPT-2 distractor surprisal are used, as well as when LeoLM distractor surprisal is used. When plausibility and GPT-2 surprisal are used as predictors, plausibility remains a significant predictor of RTs in all regions except in the Pre-critical region. Conversely, when plausibility and LeoLM surprisal are included, plausibility is still significant in the Spillover and Post-spillover regions but no longer in the Critical region. GPT-2 and LeoLM surprisal still have no significant effect on RTs in any region.

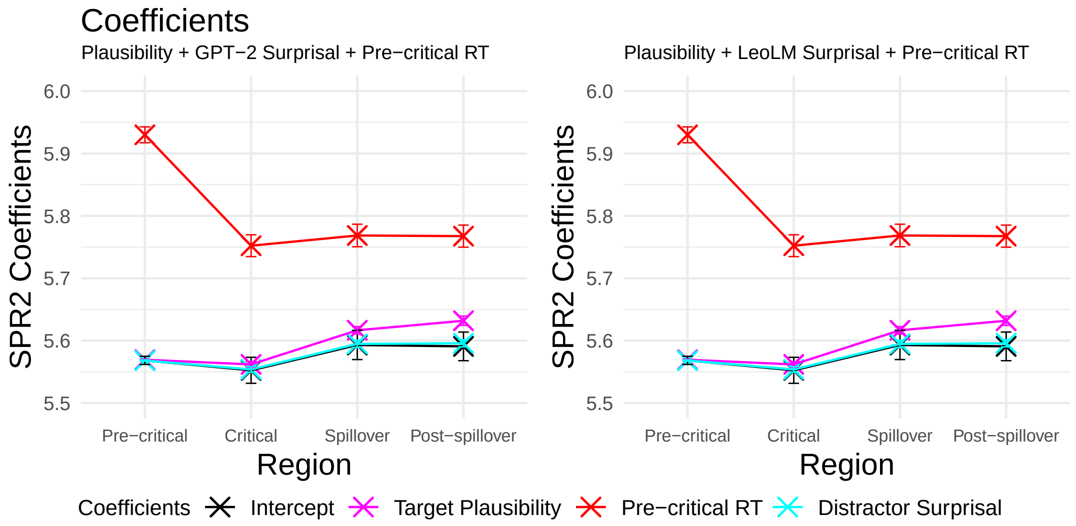

FIGURE 7.6: Coefficients, added to their intercept, from the models including Pre-critical reading time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

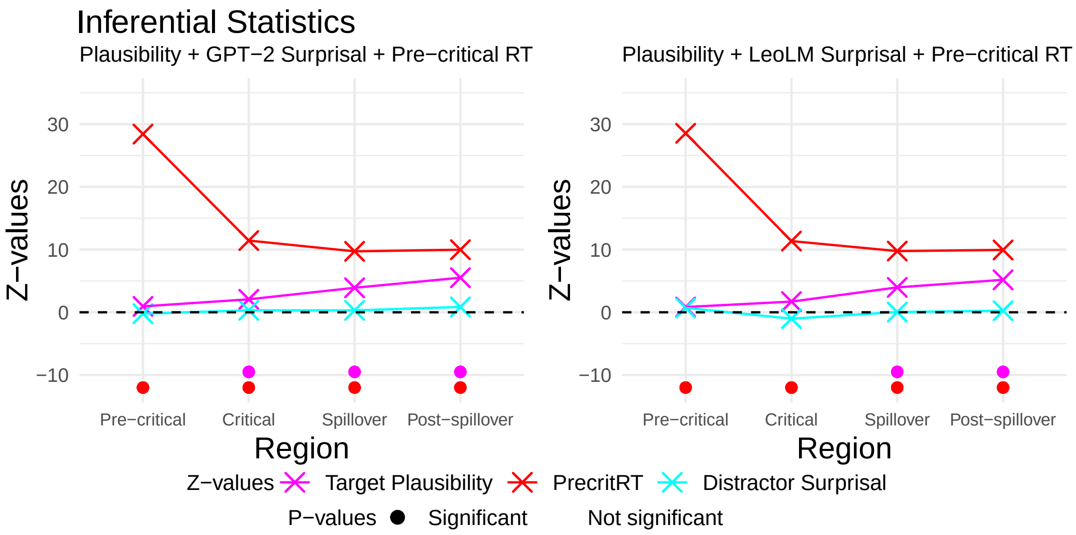

FIGURE 7.7: Effect sizes (z-values) and p-values from the models including Pre-critical reading time as a predictor.

7.5 Discussion

The results of the second self-paced reading study show that RTs pattern with the three plausibility levels of Conditions A, B and C (see Figure 7.1). However, the RT pattern $A < B < C$ is more pronounced, i.e., the difference in RTs between Condition B and C is greater, compared to the RT pattern observed in the first self-paced reading study (see Figure 6.2). Thus, the results of the first self-paced reading study, and especially those of the second self-paced reading study, are consistent with the findings of Aurnhammer et al. (2023), which indicate that RTs gradually increase as plausibility decreases, reflecting the increased effort to integrate less plausible content.

Furthermore, the observed RT pattern in the second self-paced reading study showed that it was not the plausibility manipulation of Condition B, but rather the combination of the plausibility rating task and the self-paced reading study that led to the observed RT pattern in the first self-paced reading study, which was only minimally graded due to the similarity of RTs in Conditions B and C. The reason for this seems to be that the inclusion of an online rating task in the self-paced reading study causes participants to anticipate the upcoming rating during self-paced reading, leading to higher RTs for items that are less straightforward to rate. In contrast, highly plausible (Condition A) or implausible (Condition C) items may result in faster RTs because participants have less ambiguity to resolve in order to rate them. As shown in Figure B.10, items rated as medium plausible, regardless of their condition, have higher RTs in the critical regions due to the increased difficulty of judging a medium level of plausibility compared to very high or low plausibility levels. Moreover, Figure B.11 illustrates the effect of the rating task during self-paced reading on the pre-critical regions. RTs increased strongly in the third pre-critical region in Condition C due to the implausibility of the main verb, which may have attenuated the plausibility effect when encountering the target word during the first self-paced reading experiment. Conversely, analysis of the pre-critical

RTs from the second study indicates that the strong increase in RTs observed in Condition C disappeared after the exclusion of the plausibility rating task. Finally, the generally lower RTs in the second self-paced reading study suggest that participants read the items faster on average after the removal of the rating task, regardless of their condition. It would have been interesting to analyse the reaction times on the plausibility rating task to determine whether the increased difficulty of rating medium plausible items is reflected not only in the RTs of the final sentence but also in the reaction times on the rating task itself. Unfortunately, reaction times on the plausibility rating task were not measured as a task effect was not expected. An examination of the reaction times on the subsequent task, i.e., on the comprehension questions, did not show that the reaction times for Condition B, when grouped by conditions, or for Group 2, when grouped by plausibility ratings, were higher than for the other conditions or groups.

Since the per-item averages of the single-trial ratings were shown to be a more accurate predictor of the RT data than the single-trial ratings, but not when compared to the averaged plausibility ratings from the pre-test (see Figure B.2), this suggests (1) that averaged plausibility ratings are generally more accurate predictors of the RT data due to the greater stability and robustness of the average, and (2) that the combination of the plausibility rating task and the self-paced reading experiment affected RTs and possibly also plausibility ratings. This could be due to the fact that during the online rating task, participants could not revisit the context paragraph or the final sentence when they reached the rating task, and thus could only provide their ratings based on their memory.

Finally, the linear mixed effects regression analysis showed that plausibility continuously predicted RTs. The estimated RTs (see Figure 7.2) and residual errors (see Figure 7.3) of the models including GPT-2 or LeoLM distractor word surprisal together with averaged target word plausibility from the pre-test show no observable differences. However, the z- and p-values indicate that after accounting for pre-critical RT differences, plausibility is significant in the Spillover and Post-spillover regions in the model including LeoLM distractor word surprisal and additionally in the Critical region when GPT-2 distractor word surprisal is used (see Figure 7.7). In contrast, neither GPT-2 nor LeoLM distractor word surprisal is significant in any of the critical regions in the second self-paced reading study, whereas it was significant in the first self-paced reading study for unknown reasons.

Chapter 8

General Discussion

Usually, plausibility ratings are collected in offline pre-tests by averaging the plausibility ratings obtained from a group of participants in order to examine how specific plausibility manipulations are reflected in measures of processing effort, such as RTs. Consequently, these plausibility ratings represent an aggregate measure of the plausibility of each item and do not explicitly capture individual perceptions of the plausibility. However, the perception of the plausibility of an item is highly subjective and depends on the individual's experiences and knowledge of the world, which may differ across cultures and among individuals based on other factors. One approach to address this issue is to collect plausibility ratings on a trial-by-trial basis during an online experiment, for example, during self-paced reading. In this way, each plausibility rating is obtained directly from the participant whose RTs are being measured. Thus, single-trial plausibility ratings not only improve the understanding of individuals' perceptions of plausibility, but may also lead to more accurate predictions of the observed RT data. Furthermore, online plausibility ratings are collected in closer temporal proximity to the RT data. However, while the immediate context and direct correspondence between plausibility ratings and RTs may mitigate biases due to individual differences, it doesn't eliminate potential influences of reflection and strategic effects, as participants still have the same amount of (unlimited) time to provide a plausibility rating as in offline plausibility rating studies.

Given these considerations, the main goal of this thesis was to determine whether single-trial plausibility ratings collected during a self-paced reading experiment capture the effects structure in the observed RTs better than offline plausibility ratings averaged over a group of participants. Thus, the first research question outlined in Chapter 3 was:

(1) Are online plausibility ratings collected on each trial during a self-paced reading experiment a better predictor of reading times than offline plausibility ratings?

On the one hand, single-trial plausibility ratings were expected to better capture the effects structure in the observed RT data, as they are collected from the same participants in the same context. On the other hand, this very granularity makes single-trial ratings more susceptible to variability, whether random or systematic, and therefore less robust, which may negatively affect RT predictions in terms of error minimisation. Thus, there was no strong prediction as to whether operationalising plausibility at the single-trial level or through group averages would predict the RT data more accurately. However, it was assumed that single-trial plausibility ratings may better capture the effects in the observed RT data, as they can directly be linked to the corresponding individuals' RTs and therefore may account for more variability in their RT data.

The RT predictions of the single-trial plausibility ratings collected during the self-paced reading study and the averaged plausibility ratings collected in a pre-test were assessed using the stimuli from Aurnhammer et al. (2023). Their study employed a context manipulation design in which a context paragraph was followed by a final sentence that varied in the plausibility of the target word $(A > B > C)$ and the low (A, C) or high (B) expectancy of the distractor word (see Figure 2.1). Since the objective of the current study was to compare RT predictions only for different levels of plausibility (rather than contrasting theories), the main verb in Condition B was changed to lower the expectancy of the distractor word. This modification should ensure that the expectancy of the distractor word was lower than the expectancy of the target word in all conditions, leading to the second research question:

(2) Are reading times (still) graded for plausibility after modifying the main verb in Condition B to achieve a lower distractor word expectancy?

A plausibility rating study was conducted to assess the plausibility of the final sentence given the preceding context, and surprisal values were computed to assess the expectancy of the target and distractor words. In the case of a successful plausibility and expectancy manipulation of Condition B, RTs were predicted to be graded for target word plausibility, similar to the results obtained by Aurnhammer et al. (2023), and distractor word surprisal was predicted to be higher for the distractor words than for the target words in all conditions.

Although it was not tied to a specific research question, two different LMs, GPT-2 and LeoLM, were used to compute surprisal values to assess the expectancy of the target and distractor words. Distractor word surprisal (together with target word plausibility) was used to predict the RT data to assess its influence on RTs after removing the ambiguity in Condition B. If the distractor word expectancy was successfully lowered in Condition B, distractor word surprisal was predicted not to modulate RTs, if RTs are at all sensitive to unfulfilled expectations, which was not the case in the study by Aurnhammer et al. (2023). However, depending on whether the GPT-2 or LeoLM distractor word surprisal was used, it was assumed that the predicted RTs might differ. Specifically, it was suggested that the models including LeoLM distractor word surprisal could provide more accurate RT predictions due to the greater number of parameters and the larger amount of data on which it was trained (see Chapter 4.2).

With regard to the first research question, the results indicated that the single-trial plausibility ratings obtained during the self-paced reading experiment captured the effects structure in the observed RT data less accurately, as evidenced by larger residuals (see Figure 6.4), than the offline plausibility ratings. This suggests that offline plausibility ratings better capture the effects structure in the observed RT data, as they are more robust to variability - whether meaningful or noisy - and provide a more stable measure of plausibility. Regarding the second research question, RTs were graded for plausibility $(A < B < C)$, reflecting graded integration difficulty (see Figure 6.2). However, the pattern was not as pronounced as it was the case in the study by Aurnhammer et al. (2023). As the contrast between the RTs in Conditions B and C was minimal, additional analyses were carried out (see Appendix B.3 - B.6). The analyses revealed that the RTs were affected by the plausibility rating task (see Figure B.10), probably because participants anticipated the upcoming task and evaluated the plausibility of the final sentence already while reading, which affected their RTs. Specifically, RTs were highest on average for items rated as medium plausible (i.e., assigned a rating of 3, 4, or 5), regardless of their conditions, suggesting that judging items perceived as medium plausible requires more cognitive effort than judging items perceived as clearly plausible or implausible. This cognitive load was on average even higher than the processing effort introduced by items perceived as implausible and explains why the RTs in Condition B, which contained the majority of items perceived as medium plausible, were so similar to the RTs in Condition C. Unfortunately, as no task effects were expected, the reaction times were not measured on the plausibility rating task. Otherwise, the reaction times might have revealed that participants took longer to rate items perceived as medium plausible than items perceived as highly plausible or implausible, providing further evidence that rating medium plausible items requires higher processing effort. However, a second self-paced reading study in which the rating task was removed confirmed this hypothesis. The results of this experiment show that RTs are again graded for plausibility (see Figure 7.1), but that the pattern is much more pronounced, i.e., the contrast in mean RTs between Condition B and Condition C is greater than in the first self-paced reading study.

Returning to the first research question, one might ask whether the offline plausibility ratings more accurately captured the effects in the observed RT data due to the relative stability of the average itself, or whether they better predicted the RTs because of other properties of the collected data (e.g., due to the task effect). Specifically, the combination of the self-paced reading and plausibility rating task may have influenced not only the RTs but also the online plausibility ratings, potentially contributing to the poorer RT predictions of the single-trial plausibility ratings. Since the seven-point Likert scale was presented only after participants had read the context paragraph and final sentence – preventing participants from rereading the context paragraph or the final sentence – they had to rely on their memory to provide a rating, while also concentrating on the self-paced reading task, which could have affected both the RTs and the online plausibility ratings. To investigate this, the single-trial plausibility ratings were averaged per item, similar to the offline plausibility ratings, and used to predict the RT data. The resulting estimates and residuals (see Figures B.1 and B.2) show that the models including the averaged online plausibility ratings capture the effects structure in the observed RT data better than the models including the single-trial plausibility ratings, but worse than the models using the averaged plausibility ratings collected in the pre-test. This suggests that (1) the average is generally a better predictor of RTs due to its relative stability and robustness to variability, and that (2) the offline plausibility ratings provide even more accurate RT predictions than the averaged online plausibility ratings, despite being collected at a greater temporal distance and from a different group of participants, which again suggests that the RTs, and possibly the online plausibility ratings, were influenced by the online rating task.

Finally, a likelihood ratio test showed that including single-trial target word plausibility (in addition to averaged pre-test plausibility and GPT-2/LeoLM distractor word surprisal) significantly improved the model fit in all regions except the Spillover region (see Table 6.4). This suggests that although the single-trial plausibility ratings provide less accurate RT estimates than the offline plausibility ratings, they contain useful information that explains additional variance in the RT data. However, using single-trial target word plausibility as a predictor alongside offline target word plausibility (and GPT-2/LeoLM distractor word surprisal) does not improve prediction accuracy (see Figure 6.9) in almost any region or condition. This suggests that although the single-trial plausibility ratings contain useful information that explains additional variability, they do not specifically improve the model's ability to predict RTs for Conditions A, B, and C, i.e., they only lead to an overall improvement in RTs that is not reflected in the RTs grouped by the three conditions.

The following sections of this chapter discuss challenges and opportunities associated with the manipulations of plausibility and expectancy and outline predictions for future research based on the current stimuli and design.

8.1 The Role of Offline and Online Plausibility Ratings

The current study investigated the effects of averaged (offline) and single-trial (online) plausibility on RTs. It was hypothesised that single-trial plausibility ratings might be more predictive of the RT data than averaged plausibility ratings, as taking into account variation in individuals' perceived plausibility might be more predictive of their RTs. For example, a participant who rated an item as highly plausible (e.g., with a "7") might have read the item faster than another participant who rated the same item only as medium plausible (e.g., with a "4"). However, precisely because single-trial plausibility ratings take into account individual variability, they are more susceptible to random noise, whereas group averages smooth out both random errors and individual idiosyncrasies, leading to a more stable estimate of plausibility. This robustness appears to be the main reason why the averaged plausibility ratings captured the effects structure in the observed RT data more effectively than the single-trial plausibility ratings (see Figure B.2). The averaging process increases the signal-to-noise ratio, since the signal becomes clearer when random noise is minimised, which makes it easier to detect statistically significant differences. It is therefore not surprising that the offline plausibility ratings were found to be a significant predictor of RTs in the Critical, Spillover, and Post-spillover regions when combined with GPT-2 and LeoLM distractor word surprisal, whereas single-trial plausibility ratings were significant only in the Spillover and Post-spillover regions when combined with GPT-2 distractor word surprisal and only in the Spillover region when combined with LeoLM distractor word surprisal (see Figure 6.8).

These findings may be considered to reflect the assumptions of Featherston (2007), who argued that variability between and within individuals' judgments is, at least to some extent, random rather than systematic, and in both cases is smoothed out when considering the mean, which can be taken as "the 'underlying' value, free of the noise factor" (Featherston, 2007, p. 284). He advises to average plausibility ratings since comparing individuals' ratings amplifies the error variance because the variance in their judgments may be in opposite directions, whereas the averaging process cancels out these errors.

While the results have shown that averaged plausibility ratings are a better predictor of RTs in terms of reducing the unexplained variance in the data, this does not imply that they provide more information about the variability between and within participants. For example, if the research goal is to identify and understand systematic processing differences between and/or within individuals, single-trial plausibility ratings are certainly more informative due to their granularity compared to averaged judgments. Verhagen et al. (2019) found that individual variation in actual ratings is rarely just noise; rather, it represents meaningful information that reflects characteristics of language use and linguistic representations. They point out that what might be considered noise, such as unnoticed typos or instances where participants get bored and assign random scores to finish more quickly, are not real judgments. This means that differences in true individual ratings always reflect systematic biases, and a participant may rate certain stimuli as more or less plausible based on personal experiences with language and the world. Similarly, within-individual variation may arise from systematic differences in cognitive processes such as working memory, such that individuals with a higher attention span may be less likely to be distracted, or some individuals may apply the rating scale more consistently across items. While systematic differences in plausibility ratings, similar to random noise, are obscured when ratings are averaged over a group of participants, these differences can be uncovered and examined through single-trial analyses. A different problem from this perspective is the distinction between ratings that are real and those that are not. This involves determining when the variability is meaningful and when it is random, as well as identifying the factors that contribute to these differences and controlling for them in the analyses. In addition, in the current study, individuals' plausibility ratings were also used to predict their corresponding RTs. In this context, random noise could also arise from the self-paced reading task (e.g., accidental key presses or delayed key presses due to distraction), affecting the extent to which the plausibility ratings are predictive of the RTs.

Troyer et al. (2020) also performed a single-trial analysis using participants' reports (*yes*/*no*) of whether they knew a Harry Potter (HP) fact as a predictor of their ERPs. Given the similarity to the findings of Troyer and Kutas (2018), in which participants were not asked to indicate whether they knew a fact, Troyer et al. (2020) reported that the N400 effects were not influenced by task effects based on participants' reports. However, in the study by Troyer et al. (2020), participants had to read single sentences while their ERPs were recorded, whereas in the current study, participants first had to read a context paragraph and then a final sentence, during which they had to press the *Space* bar after each word to move through the sentence. Moreover, assigning a plausibility rating on a 1-7 Likert scale is likely to be more demanding, especially for medium plausible items, than simply indicating whether a fact is known or unknown. Thus, in the current study, it appears that the cognitive load imposed by the plausibility rating task affected participants' behaviour, such that RTs were influenced based on the perceived plausibility of the items.

Task effects may be avoided by following the approach of Troyer and Kutas (2020), who collected participants' trial-level reports offline. This could be done by presenting the participants with the same items again after the self-paced reading study and asking them to rate their plausibility. While this would extend the duration of the study, it offers the possibility of using trial-level plausibility ratings as predictors of RTs (or ERPs), while avoiding confounding RTs by task effects. In addition, Verhagen et al. (2019, p. 284) point out that, unlike "time-pressured performance data" (such as RTs), offline judgement tasks provide participants with ample time for reflection and their ratings are less influenced by factors such as sneezing, lapses in attention, or other unintentional distractions. In the current study, participants had the same amount of time to provide their plausibility ratings online as during the offline plausibility rating task. However, once they saw the rating scale, they could not reread the context paragraph or final sentence, so they had to provide their ratings more quickly while still remembering the content.

As discussed, the current study showed that the variability in plausibility ratings for medium plausible items (Condition B) is significantly greater than for plausible (Condition A) and implausible (Condition C) items (see Table 6.2 and Figure 6.1) and that, when combined with a rating task, the average RTs for items perceived as medium plausible are higher than the RTs for items perceived as plausible or implausible (see Figure B.10). The reason for this may not only be the ambiguity introduced by the medium level of plausibility, but also be related to the presentation of the plausibility levels on a seven-point Likert scale. In the current study, the numbers "7" (plausible) and "1" (implausible) are probably easier to assign because they sit at the extremes, indicating that an item is perfectly plausible or implausible. All values in between suggest that an item is more or less medium plausible, but the distinctions between these intermediate values are rather vague and not well defined (see for example Knapp, 1990). Although "4" is exactly in the middle of the scale and represents a perfectly medium plausible item, "2", "3", "5" and "6" could also be interpreted as reflecting varying degrees of medium plausibility. This makes it more difficult to categorise medium plausible items, leading to increased variability and making the results more difficult to interpret. Therefore, an alternative could be to use a three-point scale with the response options "plausible," "medium plausible," and "implausible". This would reduce the variability in responses (which, however, may be desired especially when investigating individual differences), make the results more interpretable and potentially allow the models to better capture the RTs in the medium plausible condition.

Finally, the additional analyses showed that the plausibility manipulation of some items may not be consistent with the intended plausibility levels of Conditions A, B or C (see Appendix B.4 and B.5). A qualitative analysis of the plausibility ratings obtained for each item is unfortunately beyond the scope of this thesis. However, future studies using the current stimuli for their research purposes might consider revising the stimuli. Specifically, they could examine and modify items that were intended to be plausible (Condition A) that received relatively low plausibility ratings on average, or items that were intended to be implausible (Condition C) that received relatively high plausibility ratings on average, as well as items intended to be medium plausible (Condition B) that received either very low or very high plausibility ratings on average. For example, in the case of Condition A, this could include items that, when grouped by the averaged plausibility ratings, do not fall into Group 3, which contains only the third of items with the highest averaged plausibility ratings. An item in Condition A that falls into Group 2 or 1 did not receive a high plausibility rating on average, indicating that it is less plausible than desired.

8.2 LM Surprisal as a Measure of Expectancy

Although the primary goal of this thesis was not to compare how well the surprisal values of different LMs predict the observed RTs, the operationalisation of distractor word expectancy using GPT-2 and LeoLM surprisal provides an opportunity to discuss their impact on the RT predictions.

The coefficients obtained from all models for GPT-2 and LeoLM distractor word surprisal were negative across all regions (see Figures 6.5 and 7.5), suggesting that lower surprisal (higher expectancy) predicts slower reading. This seems counterintuitive as lower surprisal would typically be expected to lead to faster reading. If the expectancy of the distractor word was high but the distractor word was not presented in the target position, as was the case in Condition B of the study by Aurnhammer et al. (2023), this could indeed lead to slower reading due to the processing cost of disconfirmed expectations. However, as the surprisal pre-test showed, the expectancy of the distractor word was low in all conditions (see Table 5.1).

Furthermore, Aurnhammer et al. (2023) also found no significant modulation of RTs due to distractor word cloze probability, even though their item manipulation created a prediction disconfirmation in Condition B, suggesting that RTs are not sensitive to unfulfilled expectations. Given that the expectancy of the distractor word was low in all conditions in the current study, distractor word surprisal was not expected to have a significant effect on RTs anyway. Indeed, neither GPT-2 nor LeoLM distractor word surprisal was significant in any of the four critical regions after accounting for pre-critical RT differences (see Figures 6.8 and 7.7) in the models including offline target word plausibility. Surprisingly, when single-trial target word plausibility and LeoLM distractor word surprisal were included as predictors, distractor word surprisal significantly predicted RTs in the Spillover and Post-spillover regions, even after accounting for pre-critical RT differences (see Table B.1). However, it should be noted that p-values were not corrected for multiple comparisons, which increases the likelihood of a false positive result (type I error), i.e., the likelihood of observing a significant effect when there is none. After applying the Bonferroni correction, where the significance level (α) is divided by the number of comparisons (4 in this case), the significance threshold for each test is lowered to 0.0125. As a result, in the models including LeoLM distractor word surprisal and single-trial target word plausibility, LeoLM distractor word surprisal is no longer significant in any region (see Table B.2). Regardless of the established significance threshold, it is noteworthy that the p-values for LeoLM distractor word surprisal are generally smaller than those for GPT-2, providing stronger evidence that the observed effect is inconsistent with the null hypothesis, i.e., is not due to random variation.

One reason for this observation could be the higher correlation between single-trial target word plausibility and LeoLM distractor word surprisal ($r = 0.16$; see Table 6.3). While this is not a high correlation per se, the variables cannot be said to be independent of each other, as was the case for target word plausibility and GPT-2 distractor word surprisal ($r = 0.01$) or in the study by Aurnhammer et al. (2023) for target word plausibility and distractor word cloze probability ($r = 0.01$). However, LeoLM distractor word surprisal was not significant in combination with offline target word plausibility, although the correlation between these predictors was even higher (r = 0.28). Since only LeoLM distractor word surprisal was significant, and only in combination with single-trial target word plausibility, it can be assumed that a combination of factors related to the properties of LeoLM distractor word surprisal and single-trial target word plausibility contributed to the observed statistical significance (e.g., differences in surprisal estimates due to differences in training and tokenisation, and higher variability in single-trial plausibility ratings, leading to a lower signal-to-noise ratio, which increases the difficulty of detecting a consistent relationship between plausibility and RTs).

As mentioned above, it was assumed that including LeoLM distractor word surprisal along with online or offline target word plausibility could lead to more accurate RT predictions given that LeoLM is significantly larger in terms of parameters and training resources compared to GPT-2. Indeed, LeoLM distractor word surprisal explained more variance in the RT data than GPT-2 distractor word surprisal in combination with online target word plausibility (see Figure 6.4). These findings can be seen as consistent with studies that found larger Transformer-based LMs to be more predictive of RTs (Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Wilcox et al., 2020), and in contrast to the findings of Oh and Schuler (2022, 2023); Oh et al. (2024), who observed that surprisal estimates derived from more sophisticated LMs were less predictive of RTs. However, LM predictions depend on the nature of the task (e.g., carefully constructed items vs. self-paced reading of naturalistic texts), the scientific context, the goals of the study, and many other factors. Moreover, this study does not offer the optimal context to assess the predictive power of LMs, given that distractor word expectancy was low and therefore predicted to not predict RTs.

In addition, the surprisal estimates of LeoLM were generally more consistent with human plausibility judgments than the surprisal estimates of GPT-2. Specifically, the correlations between LeoLM target word surprisal and offline target word plausibility (r = - 0.51) or online target word plausibility (r = - 0.35) were clearly higher than the correlations between GPT-2 target word surprisal and offline target word plausibility (r = - 0.36) or online target word plausibility (r = - 0.25; see Table 6.3). This suggests that LeoLM models language in a way that is more consistent with human expectations and experiences.

The obtained surprisal values and RT predictions revealed that results can vary significantly even among LMs with the same architecture. Thus, results can vary especially when the same variables are measured using different expectancy metrics. Since Aurnhammer et al. (2023) operationalised expectancy by cloze probability rather than LM surprisal, this likely affected the RT predictions differently than the use of LM surprisal in the current study. However, it is difficult to determine the extent to which this is the case, as there are also other differences between the two studies: While the first self-paced reading study differs from the study by Aurnhammer et al. (2023) in terms of the online plausibility rating task, the low expectancy of the distractor word in Condition B, and the operationalisation of expectancy using LM surprisal, the second self-paced reading study differs from the study by Aurnhammer et al. (2023) only in the latter two aspects. These discrepancies ultimately resulted in distractor word cloze probability not significantly predicting RTs despite the high expectancy of the distractor word in Condition B in the study by Aurnhammer et al. (2023), and in LeoLM distractor word surprisal significantly predicting RTs despite the low expectancy of the distractor word in all conditions in the current study. Moreover, in the study by Aurnhammer et al. (2023), the residuals were notably smaller than in the current study, even when compared to the models that included offline target word plausibility as a predictor. As the offline plausibility ratings were collected in the same way as in Aurnhammer et al. (2023), and the RTs observed in the second self-paced reading study showed a similar pattern to the RTs observed in Aurnhammer et al. (2023), this suggests that the different operationalisations of expectancy are likely to have influenced this outcome.

Furthermore, in the study by Aurnhammer et al. (2023), cloze probability was high when plausibility was high (i.e., for the target word in Condition A and for the distractor word in Condition B), but in the current study, the surprisal values computed by GPT-2, and especially those computed by LeoLM for the distractor word, were on average lower (i.e., the expectancy was higher) in less plausible conditions (see Table 6.3), although lower plausibility is usually associated with lower expectancy. This suggests that LM surprisal may not fully capture more subtle differences in expectancy between generally less expected words. As there are conflicting findings as to whether cloze probability (e.g., Brothers and Kuperberg, 2021) or LM surprisal (e.g., Hofmann et al., 2022) predicts RTs better, it is reasonable to assume that each method of measuring expectancy has its strengths and weaknesses, and that the choice of method may therefore depend on the processing index under consideration, as well as on the specific research context and the theoretical motivation.

8.3 Predicting ERP Components Based on Graded RTs

The results of both self-paced reading studies, especially the second one, showed that RTs scaled gradually with plausibility (see Figures 6.2 and 7.1) – in line with the findings of Aurnhammer et al. (2023) – reflecting graded integration difficulty.

Even though it was not possible to conduct an EEG study within the timeframe of this thesis, the observed RTs provide an opportunity to discuss whether the current design would result in N400 or P600 effects. Given the similarities to the study by Aurnhammer et al. (2023) in terms of the experimental design and the graded nature of the observed RTs, a graded P600 and no N400 effect are also predicted based on the current design. Furthermore, the offline plausibility ratings would likely be a continuous and significant predictor of P600 amplitude, similar to the findings of Aurnhammer et al. (2023). The same is true for the single-trial plausibility ratings collected during the first self-paced reading study, although their predictive power may be somewhat lower, as observed in the RT analysis (see Figure 6.4), due to their variability.

Although the current study adopted the context manipulation design of Aurnhammer et al. (2023), the aim was to compare offline and online plausibility as predictors of RTs, rather than to contrast the predictions of different theories. Therefore, the expectancy of the distractor word in Condition B was lowered, while maintaining graded target word plausibility. Based on this modification, the predictions of multi-stream models change. By reducing the expectancy of the distractor word in Condition B, it no longer serves as a semantically attractive alternative that, when encountering the less plausible target word, could repair the semantic anomaly. Under this account, no P600 effect is predicted for Condition B since, after the semantically attractive alternative is removed, the semantic anomaly is rendered irreparable. Therefore, similar to Condition C, multi-stream models now predict an N400 effect for Condition B, reflecting the challenge of forming a semantically plausible interpretation. In contrast, RI theory still predicts a graded P600 effect based on the current materials, due to the increased integration difficulty of less plausible and especially implausible words with the unfolding utterance interpretation. Thus, the first question that an EEG study based on the materials from the current study and building on the findings of Aurnhammer et al. (2023) could address is:

(1) Is the P600 amplitude still graded for plausibility after lowering the expectancy of the distractor word in Condition B, so that it no longer represents a semantically attractive alternative?

The predictions of both multi-stream models and RI theory based on the current stimuli and those used by Aurnhammer et al. (2023) are presented in Table 8.1. Whereas in the study by Aurnhammer et al. (2023) the predictions of multi-stream models and RI theory only differed as to whether an N400 or a P600 effect would be observed in Condition C, their predictions differ for Conditions B and C based on the current stimuli. Note that, unlike RI theory, multi-stream models can typically only make binary predictions about N400 and P600 effects, as they often lack the computational specification for quantitative predictions. Therefore, an N400 effect of similar size and no P600 effect relative to baseline would be predicted for Conditions B and Condition C based on the current stimuli. An exception is the *sentence gestalt* (SG) model proposed by Rabovsky et al. (2018), which follows most of the assumptions of multi-stream models while being computationally specified. The SG model posits that the amplitude of the N400 changes based on the cues provided by each incoming word, constraining the formation of a probabilistic representation that implicitly captures the meaning of the event described by the sentence. Assuming the logic of the SG model, a graded N400 effect is predicted based on the materials of the current study. However, this is usually not the case for other multi-stream models. In contrast, RI theory predicts a graded P600 effect based on the current materials, similar to the graded effect observed in the study by Aurnhammer et al. (2023), since its predictions are independent of the availability of a semantically attractive alternative. Instead, the P600 is predicted to be graded for plausibility, with increasing amplitude for Conditions $A < B < C$, due to the increased difficulty of integrating less plausible and implausible words into the unfolding utterance interpretation. Furthermore, RI theory does not predict an N400 effect, as both target and distractor words are equally primed by lexical repetition in the context paragraph of each item.

In addition, Aurnhammer et al. (2023) observed an early negativity (~250 - 400 ms post-stimulus onset) for Condition B. Importantly, the early negativity does not correspond to the N400 ERP component, as it disappeared by 400 ms, whereas the N400 peaks around 400 ms post-stimulus onset. Aurnhammer et al. (2023) suggested that the early negativity may have emerged as a result of the prediction disconfirmation in Condition B, since participants expected the distractor word but were instead presented with the less expected target word (e.g., "The lady *weighed* the **suitcase**" rather than "The lady *weighed* the tourist"). They also suggested that

	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 8.1: Predictions of multi-stream models and Retrieval-Integration theory for the N400 and the P600 ERP component based on the materials developed Aurnhammer et al. (2023) (a) and the modified materials used in the current study (b).

the early negativity may have been observed due to the absence of an N400 effect, whereas in previous studies investigating prediction disconfirmations it may have been obscured by overlap with the N400. In addition to the P600 with a parietal peak predicted by plausibility, Aurnhammer et al. (2023) observed a left-frontally peaking positivity predicted by cloze probability. Although the left-frontal positivity was not statistically significant in the analyses, they suggested that it occurred because the encountered target words disconfirmed the expected distractor words, which is in line with previous studies reporting left-frontal positivities due to a prediction mismatch (DeLong et al., 2014; Quante et al., 2018). Therefore, the second question that an EEG study based on the materials from the current study could answer is:

(2) Do the early negativity (~250 - 400 ms) and the left-frontal positivity (~600 - 1000 ms) disappear after lowering the expectancy of the distractor word in Condition B?

By lowering the expectancy of the distractor word in Condition B, the lexical mismatch between the target and distractor words is eliminated. As a result, the early negativity and the left-frontal positivity observed in Aurnhammer et al. (2023) should disappear, if they were previously observed due to a prediction mismatch. If the observed early negativity and left-frontal positivity do not disappear, this could be taken as evidence that factors beyond the lexical mismatch between the target and distractor words contributed to their presence in the study by Aurnhammer et al. (2023).

Chapter 9

Conclusion

This thesis assessed the effectiveness of offline plausibility ratings, collected in a pre-test, and online plausibility ratings, collected on each trial during a self-paced reading experiment, in predicting RTs. Given that offline plausibility ratings represent an aggregate measure of plausibility, whereas single-trial plausibility ratings capture individual differences in perceived plausibility, it was hypothesised that single-trial plausibility ratings might be more predictive of the processing effort reflected in individuals' RTs than offline plausibility ratings. Since the context manipulation design developed by Aurnhammer et al. (2023), in which plausibility was varied across three levels, was adapted for this work, it was also expected that RTs would be graded for plausibility due to the increased cognitive effort required to integrate less plausible words. In addition, distractor word expectancy, which was high in the medium plausible condition in the study by Aurnhammer et al. (2023), was assessed using the surprisal estimates computed by two LMs, GPT-2 and LeoLM. However, as distractor word expectancy was lowered in the current work, distractor word surprisal was not expected to significantly predict RTs.

The results showed that, although RTs were graded for plausibility with increasing integration effort for Conditions $A < B < C$, the RTs were almost identical in the medium plausible and implausible conditions. Grouping RTs by plausibility ratings, rather than by conditions, revealed that items perceived as medium plausible (i.e., assigned ratings of 3, 4, or 5) had the slowest RTs. This suggested that the rating task itself influenced the RTs. Apparently, it was more difficult to classify items perceived as medium plausible than items perceived as highly plausible or implausible, and this increased cognitive effort was reflected in slower RTs as participants anticipated the rating task while reading. A second self-paced reading study, in which the rating task was removed, confirmed this hypothesis by showing more pronounced differences in RTs, especially between the medium plausible and implausible conditions. The regression-based analyses from the first self-paced reading study showed that the offline plausibility ratings captured the effects structure in the observed RT data better than the single-trial plausibility ratings. Furthermore, averaging the single-trial plausibility ratings also yielded more accurate RT predictions than the single-trial plausibility ratings themselves, suggesting that averaging the plausibility ratings enhances their predictive power, independent of the effect of the rating task on the RTs. Moreover, the models that included the single-trial plausibility ratings showed slightly more accurate RT predictions when combined with the surprisal values of the more powerful LM, LeoLM, compared to GPT-2, whereas there was no difference in prediction accuracy when combined with the offline plausibility ratings. Surprisingly, LeoLM distractor word surprisal was significant in combination with the single-trial plausibility ratings, whereas GPT-2 distractor word surprisal was not significant in combination with either offline or single-trial plausibility ratings.

The results show that offline plausibility ratings yield more accurate RT predictions precisely because the individual variability in perceived plausibility is smoothed out, resulting in a more stable measure of plausibility. Whether this variability is systematic or random is another question that was not investigated in this thesis. Although offline plausibility ratings are preferable in terms of reducing the unexplained variance in the data, single-trial plausibility ratings provide more detailed insights into individual perceptions of plausibility, which may be valuable if the research goal is to identify and understand systematic differences between and within individuals.

Future studies using single-trial plausibility ratings might consider collecting these ratings offline, after the online experiment has been completed, to avoid potential effects of the rating task on the measure of processing effort. In addition, the stimuli adapted for this work provide an opportunity to address two research questions based on the findings of Aurnhammer et al. (2023) in an EEG study. First, the materials used in the current study allow for further testing of the hypotheses of multi-stream models and RI theory. Second, they allow for examination of whether the early negativity observed by Aurnhammer et al. (2023) in the medium plausible condition disappears.

Acknowledgements

I would like to thank my supervisors Prof. Dr. Matthew Crocker, Dr. Francesca Delogu and Dr. Christoph Aurnhammer for their support and guidance throughout all stages of this Master's thesis. Their feedback and expertise have shaped the direction of this thesis and greatly improved its quality.

In particular, I would like to thank Dr. Christoph Aurnhammer for sharing his stimuli and resources with me. His support and his patience in answering all my (trivial) questions were essential for the completion of this thesis.

Furthermore, I want to thank Benedict Krieger for his efforts in calculating the surprisal values used in this thesis and for his availability to answer any other questions.

I would also like to thank the members of the Psycholinguistics Group for the opportunity to work as a student assistant and make small contributions to different projects over the past year. I am particularly grateful to Dr. Francesca Delogu for her kindness and support.

I am grateful to the Department of *Language Science and Technology* for the opportunity to participate in the LST Master's programme, where I was able to gain knowledge in different fields and had the chance to meet exceptional people from all over the world.

Finally, I would like to thank Pavel Smirnov and my brother Jens Richter who have always supported me during this Master's programme.

Bibliography

Amouyal, S. J., Meltzer-Asscher, A., and Berant, J. (2024). Large Language Models for Psycholinguistic Plausibility Pretesting. In Graham, Y. and Purver, M., editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181.

Aurnhammer, C., Delogu, F., Brouwer, H., and Crocker, M. W. (2023). The P600 as a Continuous Index of Integration Effort. *Psychophysiology*, 60(9).

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.

Aurnhammer, C. and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Bañón, M., Pinzhen, C., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Rojas, S. O., Sempere, L. P., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2008). An alternative perspective on “semantic P600" effects in language comprehension. *Brain Research Reviews*, 59(1):55–73.

Boudewyn, M. A. (2015). Individual Differences in Language Processing: Electrophysiological Approaches. *Language and Linguistics Compass*, 9(10):406–419.

Boudewyn, M. A., Long, D. L., and Swaab, T. Y. (2012). Cognitive control influences the use of meaning relations during spoken sentence comprehension. *Neuropsychologia*, 50(11):2659–2668.

Brothers, T. and Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116.

Brothers, T., Wlotko, E. W., Warnke, L., and Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, 1(1):135–160.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language. *Cognitive Science*, 41(6):1318–1352.

Brouwer, H., Delogu, F., Venhuizen, N. J., and Crocker, M. W. (2021). Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. *Frontiers in Psychology*, 12.

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.

Connell, L. and Keane, M. T. (2004). What plausibly affects plausibility? Concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory and Cognition*, 32(2):185–197.

De Varda, A. G., Marelli, M., and Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, pages 1–24.

Delogu, F., Brouwer, H., and Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, 1766.

DeLong, K. A., Quante, L., and Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61:150–162.

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics*, 33:269–318.

Federmeier, K. D., Kutas, M., and Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, 115(3):149–161.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In Reitter, D. and Levy, R., editors, *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–69.

Foulkes, P. (2006). *The Handbook of English Linguistics*, chapter Phonological Variation: A Global Perspective, pages 625–669. Blackwell.

Frank, S. L., 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.

Goodall, G. (2021). *The Cambridge Handbook of Experimental Syntax*, chapter Sentence Acceptability Experiments: What, How, and Why, pages 7–38. Cambridge University Press.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In Sayeed, A., Jacobs, C., Linzen, T., and van Schijndel, M., editors, *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

Haeuser, K. I. and Kray, J. (2022). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, 43(5):1193–1220.

Hagoort, P., Hald, L., Bastiaansen, M. C. M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669):438–441.

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. (2018). Finding syntax in human encephalography with beam search. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, page 2727–2736.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, volume 2, pages 1–8.

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.

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., and Kuchinke, L. (2022). Language Models Explain Word Reading Times Better Than Empirical Predictability. *Frontiers in Artificial Intelligence*, 4:1–20.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

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

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, 39(2):121–123.

Krieger, B., Brouwer, H., Aurnhammer, C., and Crocker, M. W. (2024). On the limits of LLM surprisal as functional Explanation of ERPs. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, pages 488–495.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–72.

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

Kuperberg, G. R., Brothers, T., and Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, 32(1):12–35.

Kuperberg, G. R. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59.

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(2):1126–1177.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929. European Language Resources Association (ELRA).

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., and McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology*, 37(4):913–934.

McLaughlin, J., Osterhout, L., and Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience*, 7:703–704.

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentin, G., and Osterhout, L. (2010). Brain Potentials Reveal Discrete Stages of L2 Grammatical Learning. *Language Learning*, 60(2):123–150.

Merkx, D. and Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Michaelov, J. A., Bardolph, M. D., Coulson, S., and Bergen, B. K. (2021). Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude? In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pages 300–306.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., and Coulson, S. (2023). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, pages 1–29.

Michaelov, J. A. and Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In Fernández, R. and Linzen, T., editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, page 652–663.

Michaelov, J. A., Coulson, S., and Bergen, B. K. (2022). So Cloze yet so far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. In *IEEE Transactions on Cognitive and Developmental Systems*.

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Nair, S. and Resnik, P. (2023). Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship? *Findings of the Association for Computational Linguistics: EMNLP 2023,*, page 11251–11260.

Nakano, H., Saron, C., and Swaab, T. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience*, 22(12):2886–2898.

Ngo, H., Araújo, J. G. M., Hui, J., and Frosst, N. (2021). No News is Good News: A Critique of the One Billion Word Benchmark. *arXiv preprint arXiv:2110.12609*.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., Rueschemeyer, S. A., Segaert, K., Tuomainen, J., and Von Grebmer Zu Wolfsthum, S. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, 375(1791):20180522.

Nieuwland, M. S., Otten, M., and van Berkum, J. J. A. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience*, 19(2):228–236.

Nieuwland, M. S. and van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701.

Oh, B. and Schuler, W. (2022). Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 9324–9334.

Oh, B. and Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Oh, B., Yue, S., and Schuler, W. (2024). Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to

Reading Times. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, page 2644–2663.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., and Iliadi, C., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, pages 9–16. Leibniz-Institut f"ur Deutsche Sprache.

Payne, B. R. and Federmeier, K. D. (2017). Pace yourself: Intraindividual variability in context use revealed by self-paced event-related brain potentials. *Journal of Cognitive Neuroscience*, 29(5):837–854.

Payne, T. W. and Lynn, R. (2011). Sex differences in second language comprehension. *Personality and Individual Differences*, 50(3):434–436.

Pernet, C. R., Sajda, P., and Rousselet, G. A. (2011). Single-Trial Analyses: Why Bother? *Frontiers in Psychology*, 2(322).

Plüster, B. (2023). LeoLM: Igniting German-Language LLM Research.

Quante, L., Bölte, J., and Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: evidence from late-positivity ERPs. *PeerJ*, 6(4):e5717.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report*.

Rayner, K., Warren, T., Juhasz, B. J., and Liversedge, S. P. (2004). The Effect of Plausibility on Eye Movements in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290–1301.

Rich, S. and Harris, J. (2021). Unexpected guests: When disconfirmed predictions linger. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, page 2246–2252.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving Lexical and Syntactic Expectation-Based Measures for Psycholinguistic Modeling via Incremental Top-down Parsing. In Koehn, P. and Mihalcea, R., editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333.

Schwanenflugel, P. J. and Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, 24(2):232–252.

Schweter, S. (2020). German GPT-2 Model (Version 1.0.0). *Zenodo*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, page 1715–1725.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 8(107307).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855. European Language Resources Association (ELRA).

Smith, N. J. and Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1637–1642.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Stanovich, K. E. and West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, 112(1):1–36.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, 30(4):415–433.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., and Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models.

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*, 113:104111.

Troyer, M., Urbach, T. P., and Kutas, M. (2020). Lumos!: Electrophysiological tracking of (wizarding) world knowledge use during reading. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 46(3):476–486.

Van Petten, C. and Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *Psychophysiology*, 83(2):176–190.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In et al., I. G., editor, *Advances in Neural Information Processing Systems*, volume 30, page 5998–6008.

Venhuizen, N., Crocker, M. W., and Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56(3):229–255.

Verhagen, V., Mos, M., Schilperoord, J., and Backus, A. (2019). Variation is information: Analyses of variation across items, participants, time, and methods in metalinguistic judgment data. *Linguistics*, 58(1):37–81.

Warren, T., McConell, K., and Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Memory and Cognition*, 34(4):1001–1010.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Zehr, J. and Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX).

Appendix A

Stimuli

1. 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.
   Dann *[verabschiedete / begrüßte / unterschrieb]* die Dame den Touristen und danach ging er zum Gate.

2. Ein engagierter Lehrer sah eine alte Weltkarte in der Vitrine eines Antiquitätengeschäfts. Ein solch authentisches Artefakt schien dem Lehrer sehr geeignet für sein Klassenzimmer zu sein und er sprach die Verkäuferin an. Aufgeregt fragte der Lehrer die sympathische Verkäuferin, wie viel die Weltkarte kosten sollte. Obwohl er für eine zusätzliche Weltkarte selbst bezahlen musste, sagte der Lehrer der Verkäuferin, dass er dies gerne tun würde. Die Verkäuferin sagte daraufhin, wie beschämend es sei, dass die Schule nicht einmal für eine Weltkarte bezahlen würde.
   Dann *[kaufte / unterschrieb / füllte]* der Lehrer die Weltkarte und danach verließ er das Geschäft.

3. Eine Redakteurin hatte von ihrer Firma eine Streifenkarte erhalten. Mit dieser Streifenkarte konnte die Redakteurin günstig mit dem Bus zur Arbeit fahren und musste nicht jedes Mal eine Karte bei dem Busfahrer kaufen. Leider hatte die Tochter der Redakteurin eines Tages eine Zeichnung auf die Streifenkarte gemalt. Deswegen hatte die Redakteurin etwas Angst, als sie bemerkte, dass der Busfahrer heute nicht gut gelaunt war, als sie ihm die Streifenkarte überreichte.
   Dann *[stempelte / zeigte / aß]* der Busfahrer die Streifenkarte und sofort fuhr er viel zu schnell weiter.

4. Während er einen Tisch baute, brach ein Schreiner seinen schönen Hammer in zwei Teile. Der Schreiner hatte den Hammer immer gemocht. Deswegen schien es ihm eine Schande, ihn einfach wegzuwerfen. Es erschien dem Schreiner eine viel bessere Idee, den Hammer von seinem Lehrling reparieren zu lassen.
   Dann *[nahm / bemalte / aß]* der Lehrling den Hammer und sofort machte er sich an die Arbeit.

5. Ein Opa wollte einen Apfelkuchen bei einem Konditor kaufen. Der Konditor versicherte dem Opa, dass der Apfelkuchen heute besonders gelungen sei.

Der Opa schaute auf den Apfelkuchen in der Vitrine und sah glücklich den Konditor an.
Daraufhin *[verpackte / backte / spülte]* der Konditor den Apfelkuchen und dann wandte er sich an den nächsten Kunden.

6. Eine Lieferbotin brachte einem nervigen Kunden eine Frühlingsrolle. Der Kunde forderte jedoch von der Lieferbotin eine neue Frühlingsrolle, da diese kalt war. Nach einer Stunde kehrte die Lieferbotin einfach mit derselben kalten Frühlingsrolle zum Kunden zurück.
Nichtsahnend *[nahm / wusch / reparierte]* der Kunde die Frühlingsrolle und sogleich schloss er hinter sich die Tür.

7. In einem Restaurant unterhielt sich eine Vegetarierin mit einem befreundeten Metzger über eine Fleischwurst auf seinem Teller. Der Metzger sah die Vegetarierin an und erklärte, diese Fleischwurst zu essen, wäre ein reines Vergnügen. Er verglich es sogar damit, eine schöne Oper zu hören. Die Vegetarierin hielt dies jedoch für einen schlechten Vergleich und wies den Metzger darauf hin, dass ein Tier für diese Fleischwurst getötet worden war.
Dann *[durchschnitt / bestellte / mietete]* der Metzger die Fleischwurst und sofort begann er zu essen.

8. Ein gemeiner Kutscher schlug seinen Gaul immer sehr heftig mit einer Peitsche. Eines Tages wurde der Kutscher dabei von einem Tierliebhaber beobachtet, der Mitleid mit dem Gaul hatte. Sofort lief der Tierliebhaber zum Kutscher und seinem Gaul und nahm ihm die Peitsche weg.
Dann *[bedrohte / bezahlte / füllte]* der Tierliebhaber den Kutscher und darüber hinaus forderte er ihn auf, den Gaul in Ruhe zu lassen.

9. Mitten im Meer sah ein Kapitän ein Pärchen auf einem kleinen Segelboot. Schon aus großer Entfernung konnte der Kapitän sehen, dass das Segelboot kaputt und das Pärchen in großer Not war. Schnell änderte der Kapitän seinen Kurs und steuerte zum Segelboot, um dem Pärchen zu helfen.
Dann *[bestieg / sichtete / verschloss]* der Kapitän das Segelboot und sofort half er dem Pärchen.

10. Da der Wasserhahn einer älteren Hausfrau nicht mehr aufhörte zu tropfen, rief die Hausfrau schließlich einen Handwerker. Zuerst betrachtete der Handwerker den Wasserhahn ausführlich und versuchte dann, ihn zu reparieren. Geduldig wartete die Hausfrau daneben. Nach einer Weile sagte der Handwerker, dass der Wasserhahn schon zu kaputt sei und er einen neuen installieren müsse.
Daraufhin *[lobte / verständigte / knickte]* die Hausfrau den Handwerker und noch lange ärgerte sie sich über die Mängel moderner Geräte.

11. In einer fremden Stadt buchte ein Urlauber eine Stadtführung. Der Guide freute sich über das Interesse des Urlaubers und schenkte ihm noch einen Flyer. Der Guide erklärte dem verwunderten Urlauber, dass der Flyer zusätzliche Informationen enthalte, auf die er selbst während der Führung nicht eingehen werde. Der Urlauber freute sich über den Flyer und dankte dem Guide.

Nach der Führung *[faltete / besorgte / kochte]* der Urlauber den Flyer und dann machte er sich auf den Weg zu seinem Hotel.

12. Ein Paparazzi stellte seine große Kamera auf und wartete auf eine berühmte Schauspielerin. Es war eine sehr gute Kamera und er wollte unbedingt tolle Bilder schießen. Als die Schauspielerin den Paparazzi entdeckte, wurde sie sehr wütend, da sie nicht fotografiert werden wollte. Deshalb warf die Schauspielerin die Kamera um.
    Daraufhin *[bedrohte / erkannte / färbte]* der Paparazzi die Schauspielerin und ferner sagte er, dass er sich so nicht behandeln lasse.

13. Ein Schneider und seine Assistentin suchten für eine neue Schaufensterpuppe, die der Schneider auf einer Messe ersteigert hatte, einen Platz in dem Laden. Zuerst stellte die Assistentin sie in den hinteren Teil des Ladens. Doch dann überzeugte sie den Schneider, die Schaufensterpuppe in die Nähe des Eingangs zu stellen, da das Licht dort besser war. Tatsächlich befand die Assistentin, dass die Schaufensterpuppe dort durch das viele Licht sehr gut zur Geltung komme.
    Daraufhin *[lobte/entdeckte/schnitt]* der Schneider die Assistentin und dann sagte er, dass der Platz am Eingang eine gute Idee war.

14. Ein Schwimmer übte einen besonders schwierigen Sprung vom Sprungbrett, als er am Beckenrand ein Mädchen entdeckte. Seit einiger Zeit schon bewunderte er das Mädchen aus der Ferne, hatte sich aber nie getraut, es anzusprechen. Doch heute wollte der Schwimmer dies nachholen und ihm kam die Idee, dass er es mit dem anspruchsvollen Sprung vom Brett beeindrucken könnte. So wartete er einen Moment ab, in dem das Mädchen zum Brett blickte und sprang dann ins Wasser. Nach dem geglückten Sprung ging der Schwimmer sofort zu dem Mädchen und sprach es an.
    Danach *[musterte / besuchte / salzte]* das Mädchen den Schwimmer und nach einer Weile verriet es ihm seine Handynummer.

15. Erfreut zeigte eine Sekretärin ihrem Chefarzt die neue Diktiermaschine. Damit konnte der Chefarzt seine Arztberichte nun selbst aufzeichnen und war nicht mehr auf die Hilfe seiner Sekretärin angewiesen. Bisher hatte sie nämlich seine Berichte selbst aufschreiben müssen. Deswegen freute sie sich besonders über die neue Diktiermaschine. Da der Chefarzt heute besonders viele Patienten gehabt hatte, schlug die Sekretärin ihm vor, die neue Diktiermaschine direkt auszuprobieren.
    Dann *[verabschiedete / fand / leerte]* der Chefarzt die Sekretärin und dann machte er Feierabend.

16. Eine Reporterin wollte einen Bericht über eine Farm schreiben. Dafür hatte sie sich ein paar Fragen überlegt, die sie dem Bauern stellen wollte. Am Hof angekommen begrüßte ein Mitarbeiter die Reporterin freundlich und brachte sie zum Bauern. Auf dem Weg erzählte der Mitarbeiter, dass er schon seit zwanzig Jahren auf der Farm arbeite. Beim Farmhaus angekommen, stellte der Mitarbeiter die Reporterin dem Bauern vor und wünschte ihnen ein erfolgreiches Interview.
    Daraufhin *[verabschiedete / suchte / ordnete]* die Reporterin den Mitarbeiter und anschließend machte sie ein paar Fotos vom Bauernhof.

17. Ein Gärtner war sehr stolz auf seinen schönen neuen Rasenmäher, denn der Rasenmäher war so groß, dass man auf diesem sitzen und wie mit einem Auto herumfahren konnte. Das erzählte der Gärtner auch der kleinen Tochter seines Chefs. Begeistert fragte die Tochter des Chefs, ob sie auch mal fahren dürfe. Die Tochter kletterte neben den Gärtner auf den Sitz des Rasenmähers und sie drehten eine große Runde über die Wiese.
    Danach *[parkte / bemalte / halbierte]* die Tochter den Rasenmäher und dann sagte sie begeistert, dass sie morgen wiederkommen würde.

18. Eine junge Dame wollte einen Edelstein von einem Juwelier beurteilen lassen. Stolz erzählte sie ihm, dass sie ihn von ihrer Großtante geerbt habe. Nun wollte die Dame von dem Juwelier wissen, um welche Art Edelstein es sich handelte. Der Juwelier betrachtete den Edelstein sehr lange und sagte dann zu der jungen Dame, dass er sehr selten und wunderschön sei.
    Entzückt *[entlohnte / empfing / würzte]* die Dame den Juwelier und danach bedankte sie sich für sein Fachwissen.

19. Ein Mechaniker machte einige Zaubertricks mit einem Schraubenzieher für seine kleine Nichte. Zu ihrer Überraschung war das Werkzeug plötzlich aus der Hand des Mechanikers verschwunden, doch kurz darauf zog er den Schraubenzieher hinter dem Ohr der Nichte hervor und lachte über ihren erstaunten Gesichtsausdruck. Geheimnisvoll erzählte der Mechaniker der Nichte, dass er gerade Magie benutzt habe, um den Schraubenzieher verschwinden zu lassen.
    Verblüfft *[nahm / sah / kochte]* die Nichte den Schraubenzieher und dann sagte sie, dass sie noch mehr Zaubertricks sehen wolle.

20. Ein Mopedfahrer war versehentlich gegen die Stoßstange eines Autos gefahren. Der Autofahrer verlangte nun, dass der Mopedfahrer für den Schaden aufkomme, doch dieser weigerte sich und sagte, dass die Stange ja überhaupt nicht beschädigt sei. Daraufhin rief der Autofahrer einen Polizisten zur Hilfe. Der Polizist eilte sofort herbei und begutachtete das Fahrzeug. Dann sagte der Polizist zum Mopedfahrer, dass dieser für die Reparaturkosten des Autofahrers aufkommen müsse.
    Daraufhin *[bestach / bemerkte / sortierte]* der Mopedfahrer den Polizisten und außerdem entschuldigte er sich für den Unfall.

21. Ein Segler und seine Freundin hatten einen Bootsausflug gemacht. Nun wollten sie das Boot wieder am Steg festbinden. Die Freundin griff nach dem Strick und wollte dem Segler helfen, doch dieser sagte der Freundin, dass er keine Hilfe benötige. Daraufhin packte er den Strick und wollte einen Knoten binden. Plötzlich glitt dem Segler der Strick aus den Händen und fiel ins Wasser.
    Daraufhin *[ermahnte / informierte / verschraubte]* die Freundin den Segler und dann sagte sie, er solle etwas aufmerksamer sein.

22. Als Piraten von riesigen Goldschätzen auf einer kleinen Insel mitten im Meer gehört hatten, machten sie sich sofort auf den Weg, um sie zu suchen. Zu ihrer Überraschung entdeckten sie Einheimische auf der Insel, die die Goldschätze bewachten. Die Piraten versteckten sich vor den Einheimischen, um in Ruhe ihren Überfall vorbereiten zu können. Die Piraten warteten ab, bis die

Einheimischen schliefen, um unbemerkt an die Goldschätze zu kommen.
Dann *[versklavten / begleiteten / wechselten]* die Piraten die Einheimischen und danach segelten sie Richtung Heimat.

23. Ein Junge verspürte Lust, einen Apfel zu essen. Erst gestern hatte er bei der Ernte geholfen und anschließend den vollen Korb nach Hause getragen. Bei dem Gedanken, wie schwer der Korb gewesen war und daran, wie frisch und saftig der Apfel sein musste, lief dem Jungen glatt das Wasser im Mund zusammen. Der Junge wusste, dass die Mutter den Korb mit seinem ersehnten Apfel im Keller versteckte.
Sofort *[suchte / füllte / schlug]* der Junge den Korb und dann entschied er sich für einen großen roten Apfel.

24. Schon seit einiger Zeit bereitete sich ein Sportler auf einen großen Wettkampf im Ringen vor. Der Vater des Sportlers half ihm täglich beim Training, denn gemeinsam wollten sie den Juror mit einer guten Technik überzeugen. Der Vater kannte den Juror schon seit langer Zeit und wusste, dass der Juror sehr auf die richtige Technik achtete. Am Tag des Wettkampfes war der Vater sehr aufgeregt, doch der Sportler beeindruckte alle mit seiner hervorragenden Technik und gewann den Wettbewerb.
Danach *[beglückwünschte / bewertete / öffnete]* der Juror den Sportler und außerdem lobte er dessen Sohn in höchsten Tönen.

25. Eine Geschäftsfrau hatte bei einer Auktion eine süße, alte Scheune ersteigert, die sie zu einer Bar herrichten ließ. Ihr Mann war nämlich Kellner und wollte sich schon lange selbstständig machen. Da der Mann in Bezug auf Ästhetik nicht sehr viel verstand, überließ er es ihr, die Renovierungsarbeiten anzuleiten. Diese hatten einige Zeit beansprucht, doch der Geschäftsfrau war das egal, denn sie war mit dem Resultat äußerst zufrieden. Die Bar war wunderschön geworden und hatte ihr altes Flair nicht verloren. Begeistert zeigte die Geschäftsfrau ihrem Mann die fertige Bar.
Daraufhin *[umarmte / rief / sortierte]* der Mann die Geschäftsfrau und dann lobte er sie für ihren guten Geschmack.

26. Ein Rentner wollte auf einem Trödelmarkt sein altes Zelt verkaufen, mit dem er schon viele schöne Urlaube verbracht hatte. Deshalb wollte er nun einen neuen Besitzer finden, der genauso viel Freude daran haben würde, wie er selbst sie gehabt hatte. Plötzlich tauchte ein kleines Kind neben ihm auf und starrte begeistert auf das Zelt. Das Kind stellte dem Rentner viele Fragen und erzählte ihm auch von seinen eigenen Campingausflügen mit der Familie. Schließlich fragte das Kind nach dem Preis für das Zelt.
Lachend *[holte / sah / aß]* der Rentner das Zelt und dann schenkte er es dem Kind.

27. Als ein Lehrer seine Unterlagen holen wollte, bemerkte er, dass er seine Tasche nicht bei sich hatte. Erschrocken überlegte er, wo er die Tasche hatte stehen lassen. Ihm fiel ein, dass er sich eine Limonade hatte kaufen wollen, aber nicht genügend Kleingeld gehabt hatte. Deswegen war der Lehrer nochmal zurück ins Lehrerzimmer gegangen, um mehr Geld zu holen. Dort war er von einem Kollegen in ein wichtiges Gespräch verwickelt worden, sodass er die Limonade

total vergessen hatte. In aller Aufregung über die Limonade hatte er bestimmt auch die Tasche in der Kantine stehen lassen.
Zurück in der Kantine *[kaufte / behielt / unterrichtete]* der Lehrer die Limonade und dann suchte er seine Tasche.

28. Eine junge Bergsteigerin hatte eine neue Spitzhacke geschenkt bekommen und war nun erpicht darauf, diese sogleich an einer sehr steilen Bergwand auszuprobieren. Ihre Mutter hatte ihr die Spitzhacke erst am Tag zuvor gekauft, nachdem der Verkäufer der Mutter versichert hatte, dass es ein sehr gutes Modell sei. Am Morgen hatte die Mutter ihr viel Erfolg gewünscht und danach war die Bergsteigerin voller Tatendrang aufgebrochen, den Berg zu erklimmen. Doch leider brach die Spitzhacke durch, nachdem die Bergsteigerin schon eine Weile geklettert war und sie musste von der Bergwacht gerettet werden.
Im Krankenhaus *[tröstete / verließ / stapelte]* die Mutter die Bergsteigerin und hinterher betrachtete sie die kaputte Spitzhacke.

29. Ein Förster und eine Praktikantin gingen in den Wald, um Wild zu sehen. Der Förster schlug vor, auf einen Hochsitz zu klettern, da sie dort einen besseren Überblick haben würden. Nach einer Weile entdeckte die Praktikantin einen Hirsch. Der Hirsch war groß und hatte ein mächtiges Geweih. Doch er war sehr weit entfernt, weshalb die Praktikantin enttäuscht sagte, dass sie kaum etwas erkennen könne. Daraufhin holte der Förster ein Fernglas aus seiner Tasche und gab es ihr, damit sie den Hirsch sehen konnte.
Dann *[umarmte / wechselte / sammelte]* die Praktikantin den Förster und anschließend bedankte sie sich für das Fernglas.

30. Ein Angeklagter wurde zum Gerichtssaal gebracht, wo der Richter und der Staatsanwalt schon auf ihn warteten. Der Mann wurde eines Raubüberfalls beschuldigt und heute war der erste Anhörungstag. Nachdem der Richter die Sitzung eröffnet hatte, trug der Staatsanwalt alle Punkte vor, die dem Angeklagten vorgeworfen wurden. Danach dankte der Richter dem Staatsanwalt und begann mit der Anhörung des Angeklagten.
Am Ende *[konsultierte / ersetzte / kopierte]* der Richter den Staatsanwalt und danach ließ er den ersten Zeugen herein.

31. Aufgeregt standen die Gäste in der Kirche und lauschten der rührenden Predigt des Pfarrers. Die Braut konnte den Moment kaum erwarten, in dem sie dem Bräutigam ihr Jawort geben und den Ring erhalten würde. Sie wusste, dass der Ring ein sehr besonderes Erbstück aus der Familie des Bräutigams war, das schon lange von Generation zu Generation weitergegeben worden war, und fühlte sich sehr geehrt, dieses zu erhalten. Als der Pfarrer die Predigt beendete und dem Brautpaar die Frage stellte, gaben sich der Bräutigam und die Braut das Jawort, während der Trauzeuge den schönen Ring hervorholte.
Glücklich *[küsste / verließ / vereinfachte]* die Braut den Bräutigam und dann übergab der Trauzeuge den Ring.

32. Während ein Ritter seinen Umhang anprobierte, besprach er das bevorstehende Turnier mit dem Burgfräulein. Das Burgfräulein fand den Umhang viel zu groß und schlug vor, ihn etwas zu kürzen. Aber der Ritter wollte nicht, dass das Burgfräulein irgendetwas veränderte. Er hatte den Umhang schon seit Jahren

und dieser hatte dem Ritter bisher immer Glück gebracht.
Daraufhin [faltete / trug / entleerte] das Burgfräulein den Umhang und dann wünschte es dem Ritter viel Erfolg.

33. In einem Museum konnte eine Besucherin einen bestimmten Raum nicht finden. Verzweifelt versuchte sie, sich an der Wegbeschreibung auf ihrer Eintrittskarte zu orientieren, aber ohne Erfolg. Dann entdeckte die Besucherin eine Aufsichtsperson am anderen Ende des Raumes und fragte sie nach Hilfe. Die Aufsichtsperson erzählte, dass einige Leute Probleme mit der Wegbeschreibung auf der Eintrittskarte hätten. Die Aufsichtsperson nahm die Eintrittskarte der Besucherin und versprach, ihr den Weg zu zeigen.
Daraufhin *[begleitete / grüßte / erfand]* die Aufsichtsperson die Besucherin und währenddessen erklärte sie ihr den Weg.

34. Ein Händler war auf dem Weg in den fernen Orient, um dort kostbare Gewürze einzukaufen. Dort angekommen begab er sich zum Marktplatz. Der Händler konnte schon von weitem die Rufe hören, mit denen die Sklaven zum Kauf angepriesen wurden. Der Händler fragte jemanden nach dem Stand mit den Gewürzen. Auf dem Weg zu den Gewürzen kam auch er an den Sklaven vorbei, welche seine fremdländischen Gewänder interessiert musterten.
Dann *[grüßte / verabschiedete / versiegelte]* der Händler die Sklaven und anschließend ging er weiter zu den Gewürzen.

35. Ein Kind entdeckte in einem Schaufenster einen Teddybären, den es unbedingt haben wollte. Der Ladenbesitzer bemerkte die bewundernden Blicke des Kindes und nahm ihn vom Regal. Das Kind sagte dem Ladenbesitzer, dass es den Teddybären gerne kaufen würde, worauf der Ladenbesitzer ihm den Teddybären überreichte.
Dann *[drückte / fand / bastelte]* das Kind den Teddybären und dann lachte es vor Freude.

36. Eine Hundeliebhaberin hatte ihren Nachbarn engagiert, um auf den Welpen aufzupassen, da sie über das Wochenende geschäftlich unterwegs war. Da die Hundeliebhaberin wusste, dass der Nachbar sich gut mit Tieren auskannte und den Welpen auch sehr gerne hatte, hatte sie keine Bedenken. Trotzdem war sie froh, als sie wieder zu Hause war. Als die Hundeliebhaberin die Haustüre aufschloss, rannte ihr der Welpe entgegen und der Nachbar begrüßte sie freundlich.
Daraufhin *[entlohnte / erkannte / sortierte]* die Hundeliebhaberin den Nachbarn und außerdem bedankte sie sich für seine Zeit.

37. Ein Schuhverkäufer hatte gerade einem Kunden ein Paar Schuhe verkauft, als er beobachtete, wie draußen vor seinem Laden ein Dieb dem Kunden seine Geldbörse entwendete. Auch sah der Schuhverkäufer, dass dieser nichts davon mitbekommen hatte und der Dieb sich geschickt aus dem Staub machte. Der Schuhverkäufer blickte dem Kunden hinterher und rannte schnell nach draußen, um den Dieb aufzuhalten.
Dann *[bemitleidete / schickte / hinterlegte]* der Schuhverkäufer den Kunden und sofort erzählte er ihm von dem beobachteten Diebstahl.

38. Ein Eskimo wollte auf die Jagd gehen, um eine Robbe zu jagen. Er nahm seine Freundin als Begleitung mit. Auf dem Weg sagte der Eskimo zu der Freundin, dass sie sich ganz still verhalten müsse und sich nicht mehr bewegen dürfe, sobald sie die Robbe erblickten. Nach einer Weile entdeckte der Eskimo die Robbe in geeigneter Entfernung und zeigte sie der Freundin.
    Dann *[ermahnte / versteckte / verpackte]* der Eskimo die Freundin und danach lud er sein Gewehr neu.

39. Nach einer Abendveranstaltung machte sich eine Tänzerin auf den Weg nach Hause. Sie beeilte sich, um schnell bei ihrer Tochter und der Babysitterin zu sein. Da die Babysitterin das erste Mal auf die Tochter aufgepasst hatte, wollte die Tänzerin schnell nach Hause, um nach dem Rechten zu schauen. Zuhause angekommen fand die Tänzerin eine glückliche Tochter und eine entspannte Babysitterin vor und war sehr erleichtert.
    Dann *[vergütete / testete / stapelte]* die Tänzerin die Babysitterin und anschließend schickte sie diese nach Hause.

40. Ein Minister und sein Berater waren erzürnt über den Präsidenten aus dem Nachbarland, da dieser sich nicht an ein Handelsabkommen hielt. Daraufhin riet der Berater dem Minister, mit Sanktionen gegen den Präsidenten vorzugehen. Der Berater organisierte ein Treffen, bei dem der Minister dem Präsidenten seine Forderungen überbringen konnte.
    Dann *[verhandelte / investierte / schminkte]* der Minister mit dem Präsidenten und dabei besprachen sie genauere Details.

41. Seit Monaten hatte sich der Athlet mit der Trainerin darauf vorbereitet, bei dem wichtigsten Wettkampf des Jahres den Pokal zu holen. Die Trainerin trieb ihn hart an, da sie sicher war, dass er gute Chancen hatte. Und tatsächlich hatte sich die harte Arbeit gelohnt, denn der Athlet gewann den Pokal und überglücklich bedankte er sich bei der Trainerin. Stolz hielt der Athlet den Pokal in den Händen.
    Im Hotel *[polierte / bezahlte / verspeiste]* die Trainerin den Pokal und anschließend stellte sie ihn auf den Tisch.

42. Ein Autor ging mit dem Hund spazieren, um an der frischen Luft neue Ideen für sein derzeitiges Buch zu bekommen. Der Autor hatte einen Ball dabei, da der Hund sehr verspielt war. Im Park angekommen, warf der Autor den Ball einige Meter weit. Sofort rannte der Hund dem Ball nach und brachte ihn brav zurück.
    Daraufhin *[nahm / tauschte / zitierte]* der Autor den Ball und wieder warf er ihn einige Meter weit.

43. Eine Oma und ein Kleinkind standen vor einem Hasenstall und streichelten das Kaninchen. Die Oma gab dem Kleinkind Löwenzahn, damit dieses das Kaninchen füttern konnte und dann ging sie noch mehr Löwenzahn holen. Doch plötzlich biss das Kaninchen das Kleinkind und dieses fing fürchterlich an zu weinen.
    Daraufhin *[fütterte / entdeckte / strickte]* die Oma das Kaninchen und nebenbei tröstete sie das Kleinkind.

44. Der Geschäftsführer und der Coach saßen nebeneinander und schauten einem bedeutenden Fußballspiel zu. Leider war die Mannschaft, die der Coach trainierte, deutlich unterlegen. Der Torwart hatte bisher fast keinen Ball gehalten. Der Geschäftsführer saß bekümmert auf der Bank und selbst die gute Leistung der anderen Spieler konnte die Ungeschicktheit des Torwarts nicht wieder gut machen. Auch der Coach wirkte verzweifelt, als der Torwart aus Versehen den Ball einem gegnerischen Spieler zuspielte, worauf dieser ein Tor schoss. Am Ende verlor die Mannschaft das Spiel. Entrüstet sagte der Geschäftsführer dem traurigen Coach, dass er mit dem Torwart reden wolle.
    Letztendlich *[suspendierte / lobte / reparierte]* der Geschäftsführer den Torwart und dann fuhr er immer noch wütend nach Hause.

45. Als ein Referendar den Weihnachtsmarkt seines Gymnasiums betrat, wurde er direkt von ein paar Schülern begrüßt. Die Schüler berichteten dem Referendar, dass sie eine Tombola organisiert hatten und nun versuchten, die Lose zu verkaufen. Die Schüler hatten schon sehr viele Lose verkauft und erzählten dem Referendar nun, was er alles Schönes mit den Losen gewinnen könne.
    Amüsiert *[kaufte / verglich / betrat]* der Referendar die Lose und tatsächlich gewann er einen Preis.

46. Eine Mutter ging mit ihren eineiigen Zwillingen zum Doktor, da diese geimpft werden sollten. Im Behandlungszimmer des Doktors machte dieser Witze darüber, wie ähnlich sich die Zwillinge sahen und zeigte ihnen die Spritzen, die er schon vorbereitet hatte. Der Doktor versicherte ihnen, dass sie keine Angst vor den Spritzen haben müssten. Da die Spritzen mit ihren langen, dünnen Nadeln tatsächlich angsteinflößend aussahen, bekamen die Zwillinge trotzdem Angst.
    Dann *[nahm / erhielt / bastelte]* der Doktor die Spritzen und anschließend begann er mit der Impfung.

47. In einem Kriegsgebiet wollte ein Soldat eine Zivilistin unbemerkt an den gegnerischen Truppen vorbei schmuggeln, da es für sie sehr gefährlich war, allein unterwegs zu sein. Da sie sich in einem Kriegsgebiet befanden, hielt der Soldat die Waffe bereit. So schlichen die Zivilistin und der Soldat mit seiner Waffe still die Häuser entlang. Die Zivilistin war sehr erleichtert über die Hilfe und fühlte sich durch die Waffe auch sicher, doch plötzlich tauchte vor ihnen ein Panzer des gegnerischen Lagers auf.
    Schnell *[zückte / sicherte / durchkämmte]* der Soldat die Waffe und sofort ging er in Deckung.

48. Ein Dirigent hatte ein neues Stück geschrieben und wollte es heute Abend zum ersten Mal dem Publikum zeigen. Lange hatte er mit dem Orchester geprobt und war gespannt auf die Reaktion des Publikums. Machte das Orchester heute Abend keinen Fehler, könnte das Stück die Karriere des Dirigenten voranbringen. Als der Abend gekommen war, betrat der Dirigent zusammen mit dem Orchester die Bühne, um dem Publikum das Stück zu präsentieren.
    An diesem Abend *[spielte / erwartete / engagierte]* das Orchester das Stück und das Publikum applaudierte.

49. Ein Doktorand hatte nach Jahren endlich seine Arbeit beendet und musste sie nun seiner Betreuerin und anderen Prüfern vorstellen. Obwohl der Doktorand eng mit der Betreuerin zusammengearbeitet hatte und wusste, dass die Arbeit sehr gut war, war er trotzdem sehr nervös. Vor der Prüfung ging der Doktorand noch einmal die wichtigsten Stichpunkte bezüglich der Arbeit durch, dann folgte er den anderen Prüfern ins Büro der Betreuerin.
    Dort *[begrüßte / wechselte / reparierte]* der Doktorand die Betreuerin und dann hielt er seinen Vortrag.

50. Eine Protestantin wollte nach Israel fliegen, um sich Jerusalem anzuschauen. Sie hatte die Reise geplant, seitdem der Pfarrer ihr Bilder von seinem Aufenthalt dort gezeigt hatte. Nun war die Reise fertig organisiert und der Abflug rückte immer näher. Doch die Protestantin machte sich Sorgen, da es in letzter Zeit vermehrt Unruhen gegeben hatte. So ging sie zu dem Pfarrer, um ihn um Rat zu fragen. Sie wollte, dass der Pfarrer ihr versicherte, dass sie sich keine Sorgen machen müsse. Dadurch würde sich die Protestantin bezüglich der Reise sicherer fühlen.
    Daraufhin *[segnete / kontaktierte / las]* der Pfarrer die Protestantin und dann wünschte er der Protestantin einen guten Flug.

51. Eine Erzieherin suchte einen Therapeuten auf. Dieser war ihr von einer Freundin empfohlen worden, nachdem sie über Symptome geklagt hatte. Die Symptome waren denen einer Depression ziemlich ähnlich und die Erzieherin hatte beschlossen, dass sie professionelle Hilfe von dem Therapeuten brauche. So war die Erzieherin sehr erleichtert gewesen, als sie endlich einen Termin bei dem Therapeuten bekommen hatte, da sich die Symptome in letzter Zeit noch verschlimmert hatten.
    In der Praxis *[erfragte / entwickelte / tauschte]* der Therapeut die Symptome und daraufhin verschrieb er ein Medikament.

52. Eine Designerin hatte den Auftrag bekommen, ein Buch grafisch zu gestalten. In dem Buch ging es um Geschichten über Eisbären. Die Geschichten waren für Kinder gedacht und der Verlag wollte, dass die Designerin die Eisbären bildlich darstellte. Nun hatte die Designerin die Geschichten über die Eisbären zu Ende gelesen und war bereit, mit der Arbeit zu beginnen.
    Dann *[malte / beobachtete / leerte]* die Designerin die Eisbären und bis spät in die Nacht arbeitete sie an der Geschichte.

53. Eines Abends wurde ein Architekt von dem Bürgermeister angerufen. Dieser sagte, dass die Stadt eine neue Turnhalle zu bauen beabsichtigte. Er beauftragte den Architekten, einen Plan der Turnhalle zu erstellen und diesen in einer Rede vor dem Gemeinderat näher auszuführen. In der Rede solle er auf die besonderen Merkmale seines Entwurfes eingehen. Der Architekt versicherte, dass er sofort mit der Konzeption der Turnhalle beginnen werde und bedankte sich für die Tipps bezüglich der Rede.
    Daraufhin *[schrieb / analysierte / rief]* der Architekt die Rede und dann goss er sich ein Glas Wein ein.

54. Ein Gitarrist wurde von einer Agentin engagiert, um zusammen mit einer Sängerin auf einer Party aufzutreten. Auf dem Weg zur Probe erzählte die

Agentin dem Gitarristen, dass sie lange nach einem guten Musiker gesucht habe und glaube, dass seine Art zu spielen ausgezeichnet mit der Stimme der Sängerin harmonieren würde. Dann holte der Gitarrist sein Instrument und die Agentin sagte, dass die Sängerin schon bereit sei und sie direkt mit der Probe beginnen könnten.
Dann *[verabschiedete / beschäftigte / kaufte]* der Gitarrist die Agentin und dann ging er schnell zur Bühne.

55. In einem Museum war ein Kurator dabei, eine neue Ausstellung zu gestalten. Da es um plastische Kunst ging, hatte sich der Kurator von einer befreundeten Galeristin eine Skulptur geliehen. Gerade war die Galeristin eingetroffen und sie überlegten nun gemeinsam, wo die Skulptur am Besten zur Geltung kommen würde. Lange suchten sie nach einem geeigneten Platz und fanden schließlich einen. Mühevoll installierte der Kurator die Skulptur, während die Galeristin Anleitungen gab.
Danach *[umarmte / buchte / sammelte]* der Kurator die Galeristin und dabei dankte er ihr für ihre Hilfe.

56. Eine Studentin war mit einer Kommilitonin in einer Kneipe. Da sie danach noch in einem Club feiern gehen wollten, beschlossen sie, sich auf der Toilette frisch zu machen. Dann fragte die Kommilitonin die Studentin, ob sie ihre Wimperntusche ausleihen dürfe, da sie ihre eigene vergessen hatte. Sofort gab die Studentin ihr die Wimperntusche. Die Kommilitonin fragte eine Bedienung nach der Toilette und machte sich mit der Wimperntusche in der Hand auf den Weg zur Toilette.
Dann *[betrat / beschrieb / las]* die Kommilitonin die Toilette und anschließend schminkte sie sich.

57. Ein Verbrecher war auf dem Weg zu einem Haus, wo ein Ermittler wohnte. Dieser untersuchte einen Fall, in den der Verbrecher verstrickt war. Deswegen wollte dieser den Ermittler aus dem Weg räumen. Am Haus angekommen verschaffte sich der Verbrecher Zutritt. Er wusste, dass es in dem Haus einen Schäferhund gab und bedacht achtete er darauf, dass der Schäferhund ihn nicht hörte. Bevor er den Ermittler suchte, gab er dem Schäferhund etwas zu Essen, um ihn abzulenken.
Dann *[streichelte / versorgte / faltete]* der Verbrecher den Schäferhund und danach machte er sich auf die Suche nach dem Ermittler.

58. Ein Beschuldigter und seine Anwältin betraten den Gerichtssaal, um bei der bevorstehenden Anhörung zu beweisen, dass der Beschuldigte die Tat nicht begangen hatte. Der Kläger saß schon an seinem Platz und warf den beiden böse Blicke zu. Die Anwältin ging noch ein paar ihrer Unterlagen durch, dann begann die Verhandlung. Der Kläger wurde nach vorne gebeten und von der Anwältin zur Tat befragt. Der Beschuldigte blickte nervös drein, als der Kläger ihn vor aller Augen der Tat bezichtigte.
Daraufhin *[verteidigte / prüfte / schwenkte]* die Anwältin den Beschuldigten und dann wandte sie sich an den Richter.

59. Ein Junge ging mit seinem Kumpel zum See, da er schwimmen wollte. Der Kumpel hatte seine Angel dabei und erzählte dem Jungen, dass er heute einen

Flussbarsch angeln wollte, von denen es viele im See gab. Er hatte einen besonderen Köder dabei, mit dem er den Flussbarsch anlocken wollte. Der Junge wünschte dem Kumpel viel Glück mit dem Flussbarsch und machte einen Salto ins Wasser.
Daraufhin *[angelte / reinigte / trocknete]* der Kumpel den Flussbarsch und danach ging er selbst ins Wasser.

60. Eine Schwangere betrat das Untersuchungszimmer einer Gynäkologin und wurde von der Gynäkologin freundlich begrüßt. Die Gynäkologin deutete auf die Liege im Zimmer und forderte die Schwangere auf, sich dort hinzulegen. Die Liege war etwas hoch eingestellt, doch die Schwangere schaffte es, hochzukommen und legte sich auf die Liege.
    Daraufhin *[verstellte / suchte / verordnete]* die Gynäkologin die Liege und dann wandte sie sich der Schwangeren zu.

Appendix B

Additional Analyses

B.1 Averaging Online Plausibility Ratings

To determine whether the averaged plausibility ratings collected in a pre-test better predict the RT data than the single-trial plausibility ratings collected during the self-paced reading study (Chapter 6.4.3) due to properties of the average itself or whether the pre-test plausibility ratings were simply more consistent with the RTs for some other reason, despite being collected from different participants than the RTs (e.g., because the combination of self-paced reading and rating task affected the RTs and/or online plausibility ratings), the single-trial plausibility ratings for each item were averaged per subject and used to predict the RT data.

Figure B.1 shows the estimated RTs based on single-trial target word plausibility (top), averaged target word plausibility collected online (middle) and averaged target word plausibility collected offline (bottom) combined with either GPT-2 (left) or LeoLM (right) distractor word surprisal. The corresponding residuals are presented in Figure B.2. The estimates and residuals indicate that the averaged online plausibility ratings (in combination with either GPT-2 or LeoLM surprisal) capture the patterns in the observed RT data more accurately than the single-trial plausibility ratings, but less accurately than the averaged offline plausibility ratings. This suggests that averaged plausibility ratings are generally a better predictor of RTs, because they are more stable and less susceptible to (systematic and random) variability than single-trial plausibility ratings. At the same time, the averaged offline plausibility ratings capture the effects structure in the observed RT data better than the averaged online plausibility ratings (even though the offline plausibility ratings were collected from different participants than the RT data), which suggests that the RTs (and potentially also the online plausibility ratings) were affected by the combination of the reading and rating tasks (which is less evident in the averaged online ratings due to the relative stability of the average). Moreover, the inclusion of LeoLM instead of GPT-2 surprisal appears to improve the predictions of the models fitted with single-trial plausibility and, to a lesser extent, the models fitted with averaged online plausibility. For the better-performing models fitted with averaged offline plausibility, there is no visible difference in prediction accuracy between the models including GPT-2 distractor word surprisal and those including LeoLM distractor word surprisal as a predictor.

The model coefficients, added to their intercept, for single-trial online, averaged online and averaged offline plausibility, GPT and LeoLM surprisal are presented in Figure B.3. The corresponding z-values and significance levels are displayed in Figure B.4, while the exact p-values are reported in Table B.1. The coefficients and z-values for averaged online plausibility are more similar to those for averaged offline plausibility than to those for single-trial plausibility in the models that include GPT-2 surprisal. Both averaged online and averaged offline plausibility are significant in the Critical, Spillover, and Post-spillover regions, while single-trial plausibility is only significant in the Spillover and Post-spillover regions. In contrast, GPT-2 surprisal is not significant in any region when combined with any type of plausibility. When LeoLM surprisal is included as a predictor instead of GPT-2 surprisal, the coefficients and z-values for averaged online plausibility fall between those for averaged offline plausibility and single-trial online plausibility. In this scenario, averaged online plausibility is significant in the Spillover and Post-spillover regions, while averaged offline plausibility is significant in the Critical, Spillover and Post-spillover regions and single-trial plausibility is significant only in the Spillover region. LeoLM surprisal is significant only in the Critical, Spillover, and Post-spillover regions in the models with single-trial or averaged online plausibility and only on the Spillover region in the model with averaged offline plausibility.

Finally, the model coefficients, added to their intercept, for plausibility, surprisal and Pre-critical RT are presented in Figure B.5 and the corresponding z-values and significance levels are shown in Figure B.6. The exact p-values are reported in Table B.1. The resulting coefficients and z-values of averaged online plausibility are similar to those of averaged offline plausibility in the models fitted with GPT-2 surprisal as a second predictor. In this case, Pre-critical RT is significant in all regions and both averaged online and offline plausibility are significant in the Critical, Spillover, and Post-spillover regions. In the models fitted with LeoLM surprisal instead, the coefficients and z-values of averaged online plausibility are similar to those of single-trial plausibility. In both cases, Pre-critical RT is significant across all regions, single-trial and averaged online plausibility are significant in the Spillover region, while LeoLM surprisal is significant in the Spillover and Post-spillover regions. In the model incorporating averaged offline plausibility, plausibility is additionally significant in the Critical region, whereas LeoLM surprisal does not significantly predict RTs in any region.

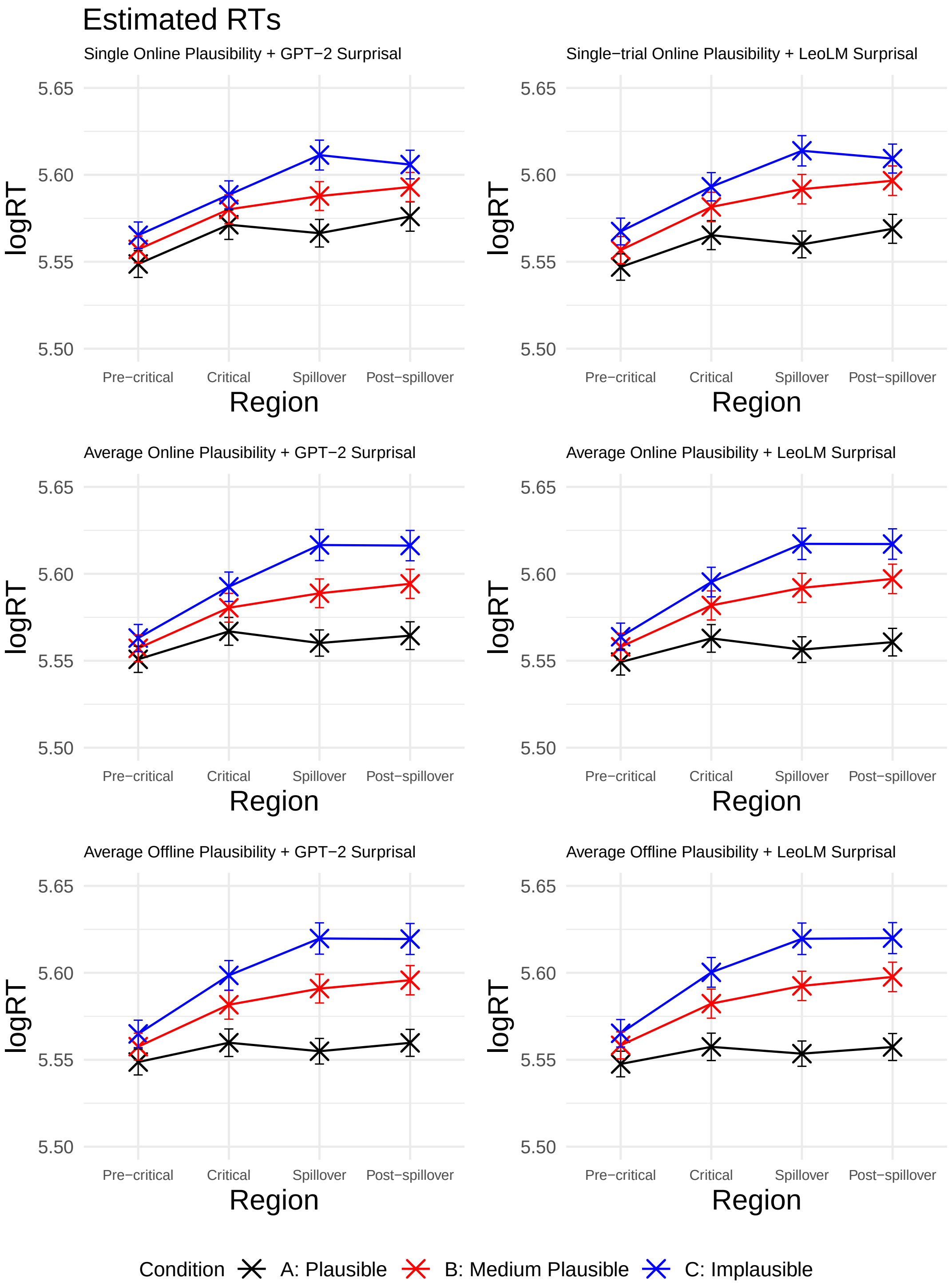

FIGURE B.1: Estimated log reading times using combinations of single-trial (online), averaged (online), averaged (offline) plausibility and GPT-2 (left) and LeoLM surprisal (right) as predictors per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

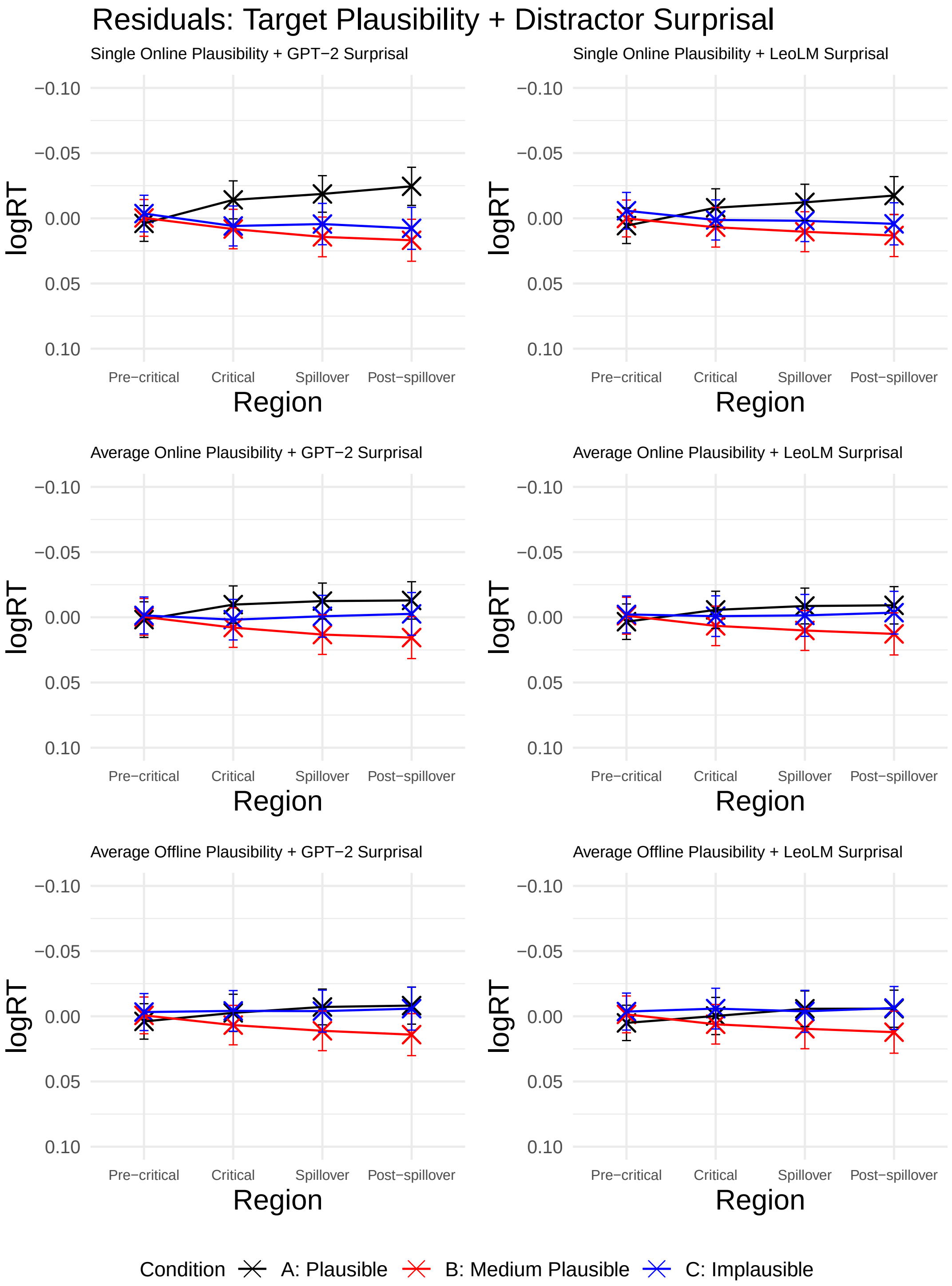

FIGURE B.2: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

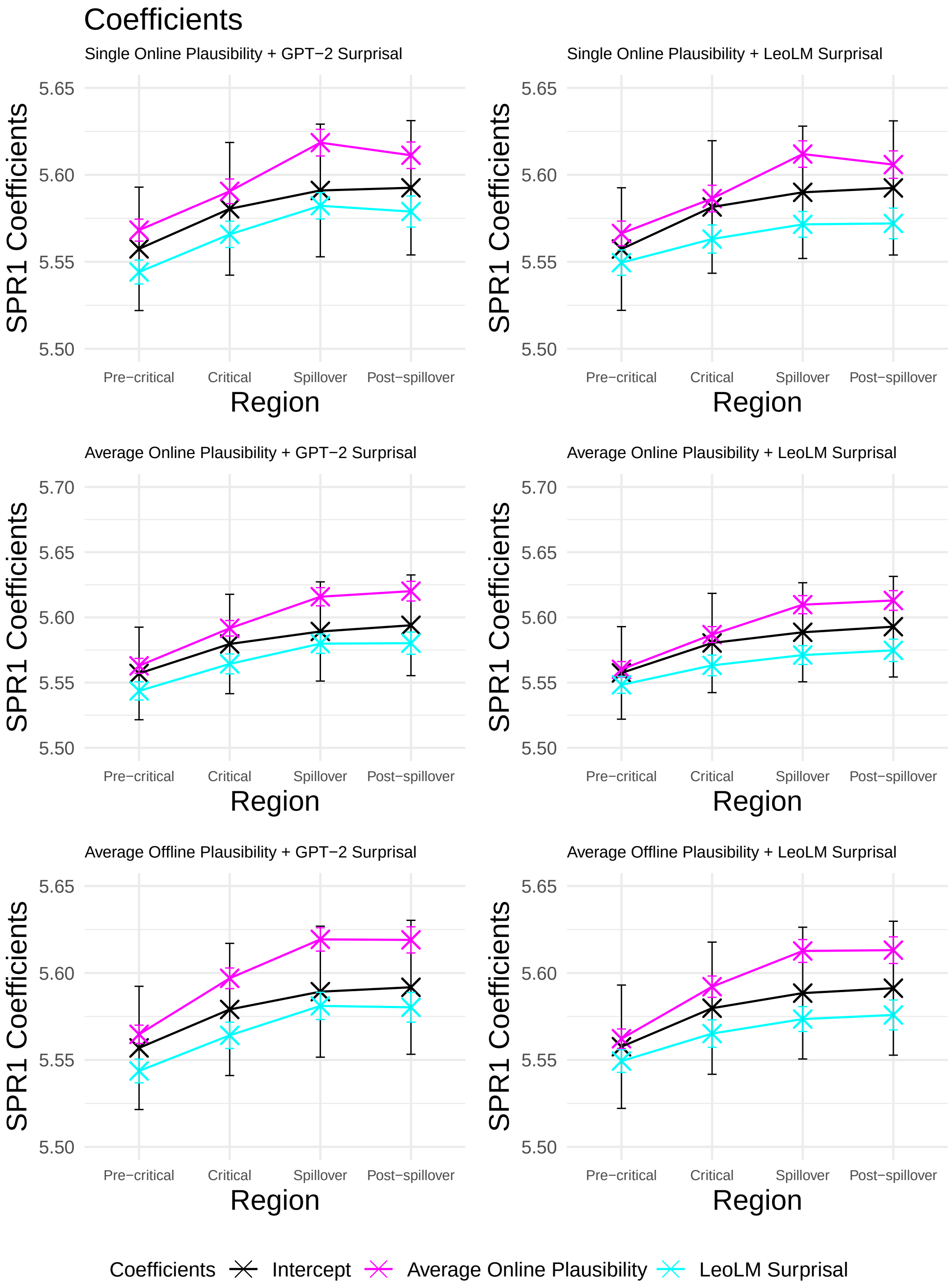

FIGURE B.3: Coefficients, added to their intercept, for the six predictor combinations used for fitting the models. Error bars indicate the standard error of the coefficients in the fitted statistical models.

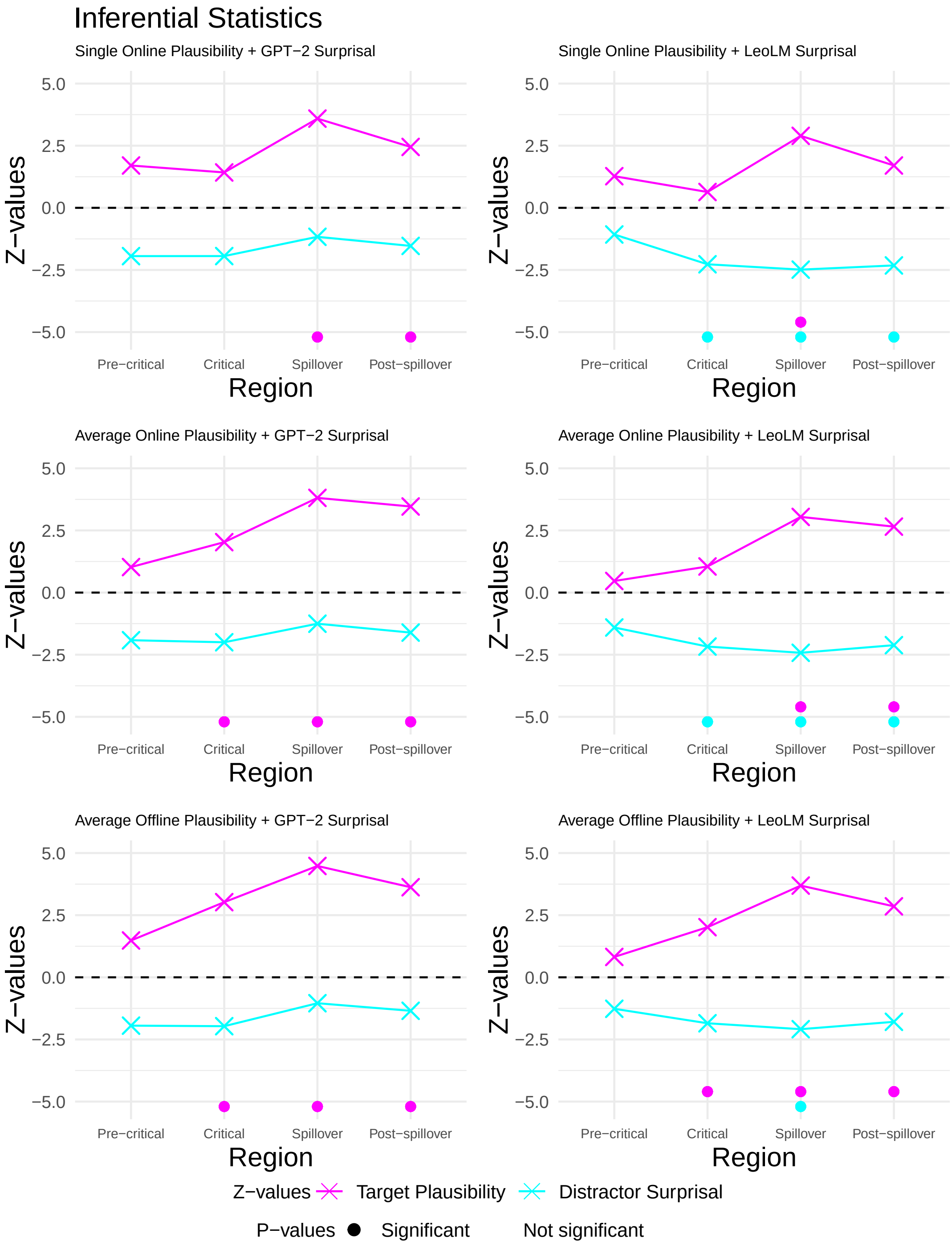

FIGURE B.4: Effect sizes (z-values) and p-values for the six predictor combinations used for fitting the models.

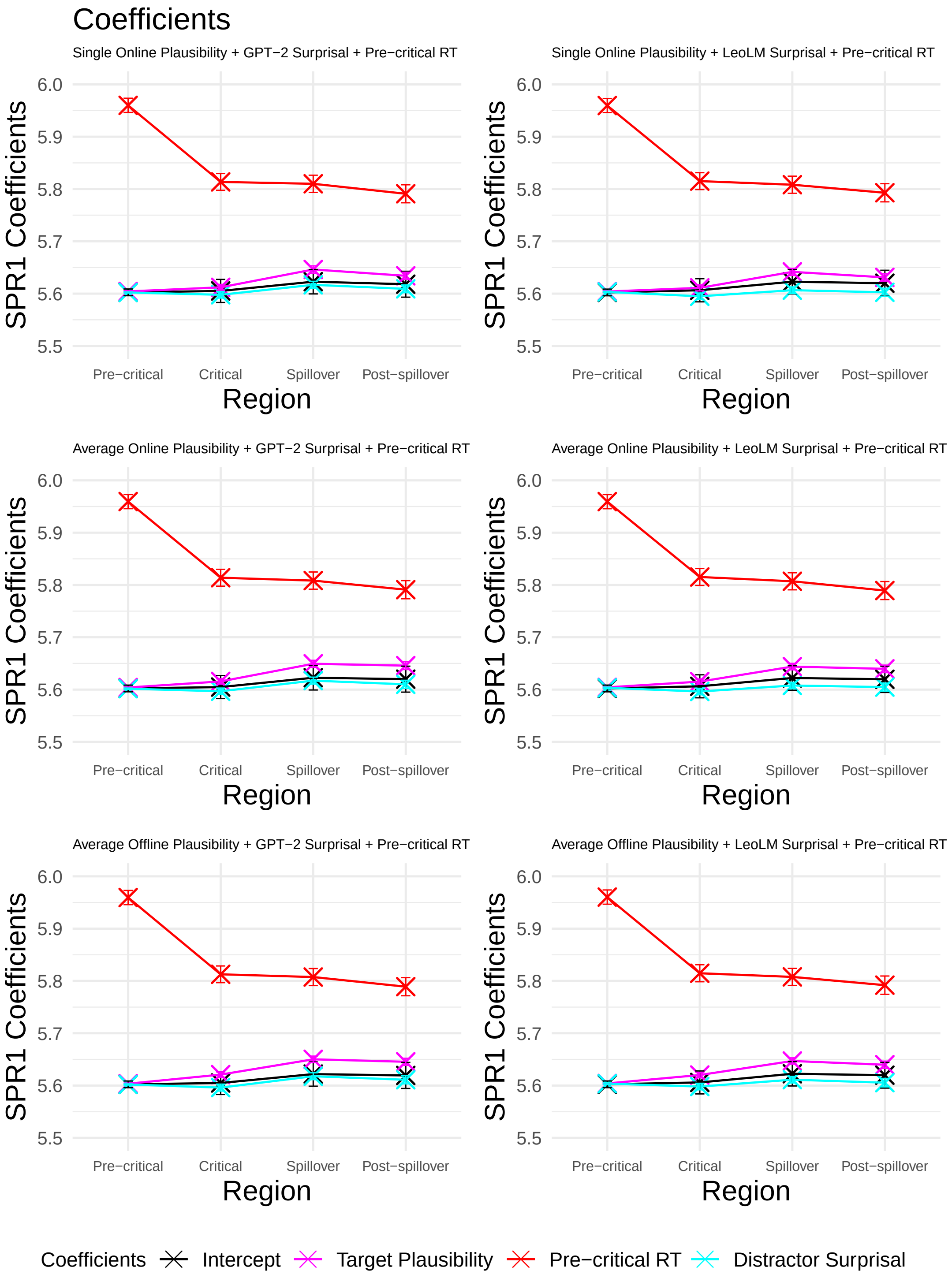

FIGURE B.5: Coefficients, added to their intercept, for the six predictor combinations used for fitting the models, including Pre-critical reading time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

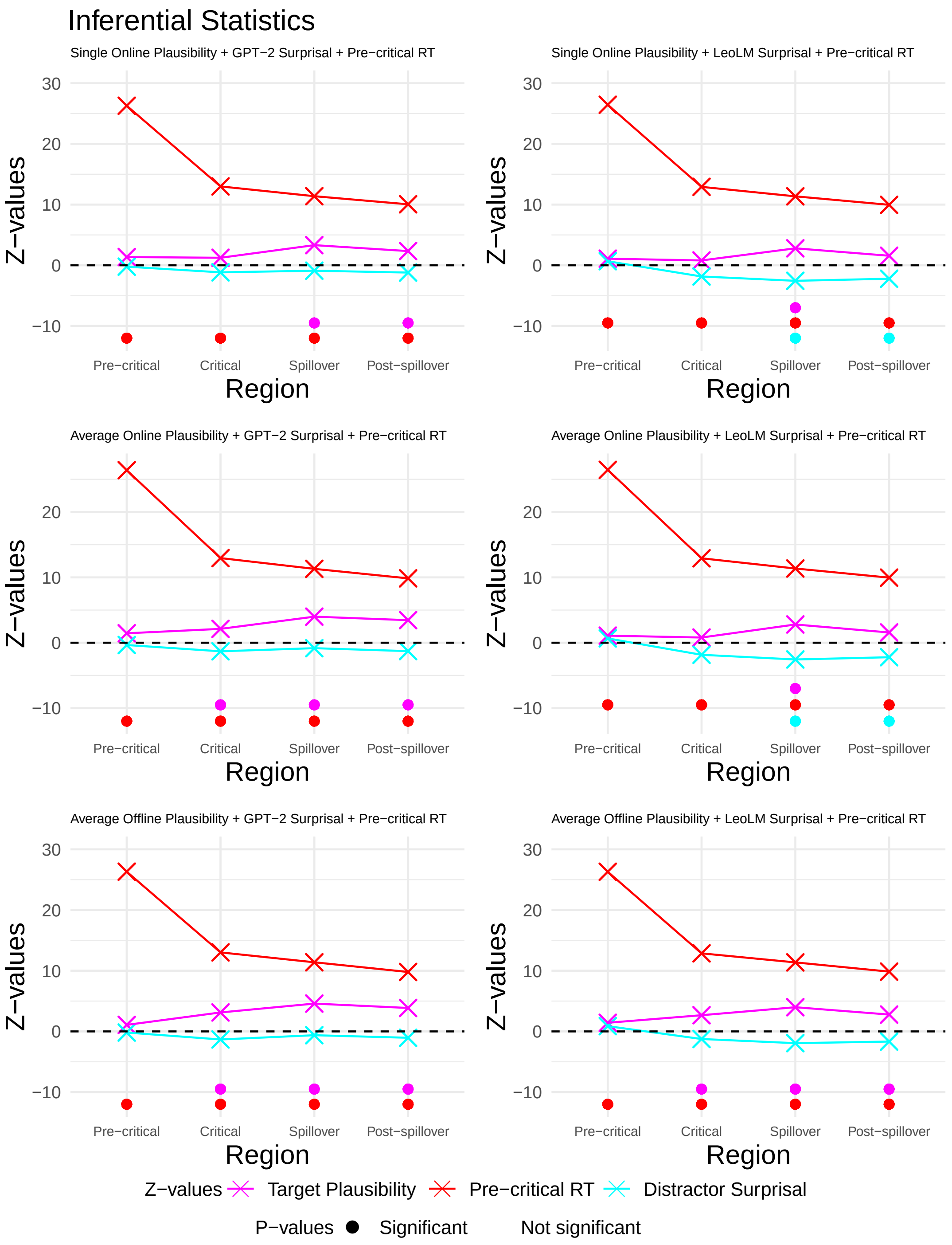

FIGURE B.6: Effect sizes (z-values) and p-values including Pre-critical reading time as a predictor.

B.2 P-Values

Table B.1 presents the p-values from the first self-paced reading study, which includes the predictors single-trial, averaged online and offline plausibility, GPT-2 and LeoLM surprisal in all critical regions. Furthermore, the p-values from the aforementioned models including Pre-critical RT are reported.

		Pre-critical	Critical	Spillover	Post-spillover
P-Value	Single Online Plaus.	0.0951	0.162	0.00617	0.0182
	GPT-2 Surprisal	0.0609	0.0624	0.250	0.134
P-Value	Single Online Plaus.	0.211	0.528	0.00617	0.0951
	LeoLM Surprisal	0.291	0.0287	0.0162	0.0243
P-Value	Single Online Plaus.	0.181	0.224	0.00193	0.0241
	GPT-2 Surprisal	0.820	0.255	0.381	0.238
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Single Online Plaus.	0.284	0.430	0.00795	0.123
	LeoLM Surprisal	0.513	0.0752	0.0131	0.0315
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Online Plaus.	0.309	0.0477	<0.001	<0.001
	GPT-2 Surprisal	0.0622	0.0543	0.220	0.116
P-Value	Avg. Online Plaus.	0.644	0.299	0.00388	0.00973
	LeoLM Surprisal	0.167	0.0359	0.0191	0.0390
P-Value	Avg. Online Plaus.	0.156	0.0367	<0.001	<0.001
	GPT-2 Surprisal	0.728	0.202	0.413	0.208
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Online Plaus.	0.284	0.430	0.00795	0.1234
	LeoLM Surprisal	0.513	0.0752	0.0131	0.0315
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Offline Plaus.	0.139	0.00309	<0.001	<0.001
	GPT-2 Surprisal	0.0605	0.0572	0.303	0.186
P-Value	Avg. Offline Plaus.	0.413	0.0449	<0.001	<0.001
	LeoLM Surprisal	0.214	0.0714	0.0415	0.0789
P-Value	Avg. Offline Plaus.	0.288	0.00224	<0.001	<0.001
	GPT-2 Surprisal	0.851	0.194	0.526	0.357
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Offline Plaus.	0.159	<0.001	<0.001	<0.001
	LeoLM Surprisal	0.393	0.225	0.0563	0.102
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001

TABLE B.1: P-values from the models fitted with single-trial, average online and average offline plausibility combined with GPT-2 or LeoLM surprisal and the same models fitted including Pre-critical RT as a predictor. Significant p-values (< 0.05) are highlighted in red.

Table B.2 contains the adjusted p-values using the Bonferroni correction, which divides the significance threshold by the number of comparisons, thereby reducing the likelihood of Type I errors in multiple comparisons. In this case, four different predictor combinations were used to predict the RTs, resulting in four tests on each region. Consequently, the new significance threshold is 0.0125 (0.05/4).

		Pre-critical	Critical	Spillover	Post-spillover
P-Value	Single Online Plaus.	0.0951	0.162	0.00617	0.0182
	GPT-2 Surprisal	0.0609	0.0624	0.250	0.134
P-Value	Single Online Plaus.	0.211	0.528	0.00617	0.0951
	LeoLM Surprisal	0.291	0.0287	0.0162	0.0243
P-Value	Single Online Plaus.	0.181	0.224	0.00193	0.0241
	GPT-2 Surprisal	0.820	0.255	0.381	0.238
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Single Online Plaus.	0.284	0.430	0.00795	0.123
	LeoLM Surprisal	0.513	0.0752	0.0131	0.0315
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Online Plaus.	0.309	0.0477	<0.001	<0.001
	GPT-2 Surprisal	0.0622	0.0543	0.220	0.116
P-Value	Avg. Online Plaus.	0.644	0.299	0.00388	0.00973
	LeoLM Surprisal	0.167	0.0359	0.0191	0.0390
P-Value	Avg. Online Plaus.	0.156	0.0367	<0.001	<0.001
	GPT-2 Surprisal	0.728	0.202	0.413	0.208
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Online Plaus.	0.284	0.430	0.00795	0.1234
	LeoLM Surprisal	0.513	0.0752	0.0131	0.0315
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Offline Plaus.	0.139	0.00309	<0.001	<0.001
	GPT-2 Surprisal	0.0605	0.0572	0.303	0.186
P-Value	Avg. Offline Plaus.	0.413	0.0449	<0.001	<0.001
	LeoLM Surprisal	0.214	0.0714	0.0415	0.0789
P-Value	Avg. Offline Plaus.	0.288	0.00224	<0.001	<0.001
	GPT-2 Surprisal	0.851	0.194	0.526	0.357
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Avg. Offline Plaus.	0.159	<0.001	<0.001	<0.001
	LeoLM Surprisal	0.393	0.225	0.0563	0.102
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001

TABLE B.2: Adjusted P-values from the first self-paced reading study using Bonferroni correction. Significant p-values (< 0.05) are highlighted in red.

Table B.3 shows the p-values of the predictors employed in the complex model, which includes single-trial and offline plausibility in combination with either GPT-2 or LeoLM surprisal. P-values for Pre-critical RT are not reported, as the models were not fitted with Pre-critical RT as an additional predictor. Table B.4 presents the p-values from the second self-paced reading study for the predictors offline plausibility and either GPT-2 or LeoLM surprisal, as well as the p-values of the models including Pre-critical RT. As plausibility ratings were not collected in the second self-paced reading study, only p-values for averaged pre-test plausibility are reported.

		Pre-critical	Critical	Spillover	Post-spillover
P-Value	Averaged Plausibility	0.824	0.0543	0.0297	0.0324
	Single Plausibility	0.513	0.545	0.415	0.652
	GPT-2 Surprisal	0.0608	0.0451	0.299	0.264
P-Value	Averaged Plausibility	0.931	0.125	0.122	0.0679
	Single Plausibility	0.569	0.505	0.467	0.661
	LeoLM Surprisal	0.161	0.0725	0.031	0.0795

TABLE B.3: P-values on each critical region for the complex model fitted with single-trial plausibility, averaged pre-test plausibility and GPT-2 or LeoLM surprisal. Significant p-values (< 0.05) are highlighted in red.

		Pre-critical	Critical	Spillover	Post-spillover
P-Value	Averaged Plausibility	0.0231	0.00165	<0.001	<0.001
	GPT-2 Surprisal	0.196	0.295	0.529	0.817
P-Value	Averaged Plausibility	0.153	0.0138	<0.001	<0.001
	LeoLM Surprisal	0.208	0.120	0.609	0.809
P-Value	Averaged Plausibility	0.350	0.0426	<0.001	<0.001
	GPT-2 Surprisal	0.832	0.753	0.762	0.394
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001
P-Value	Averaged Plausibility	0.406	0.0919	<0.001	<0.001
	LeoLM Surprisal	0.482	0.312	0.967	0.802
	Pre-critical RT	<0.001	<0.001	<0.001	<0.001

TABLE B.4: P-values on each critical region for the model fitted with averaged pre-test plausibility and GPT-2 surprisal or LeoLM surprisal and the same models fitted including Pre-critical reading time as predictor. Significant p-values (< 0.05) are highlighted in red.

B.3 Comparing Raw and Log-Reading Times

In order to gain further insights into the RT data and explore why the RTs do not clearly align with the three levels of Conditions A, B, and C – as observed in the study by Aurnhammer et al. (2023) – but instead show almost identical RTs for Conditions B and C, several analyses were conducted.

One potential explanation for the observed differences in RTs being less pronounced than anticipated is the log-transformation of the RTs (Figure 6.2), since the logarithmic function compresses the original scale, particularly for larger values, which may have resulted in differences in higher value ranges being less prominent. To determine whether the use of the logarithmic scale was the reason for the differences in RTs between conditions to appear smaller, particularly between Conditions B and C, the observed raw RTs were visualised. Figure B.7 shows the observed raw RTs along with the corresponding estimated RTs and residuals per condition and region. The raw RTs appear to increase more strongly than the log-transformed RTs, especially in the Spillover and Post-spillover regions, such that the difference between the RTs in Conditions B and C seems slightly more pronounced. However, given that the difference in the observed RT patterns between the log-transformed and raw RTs is minimal, it can be concluded that the use of the logarithmic scale is not a highly influential factor in the observed similarity of RTs in Conditions B and C.

B.4 Filtering Reading Times by Plausibility

In order to determine whether the observed RT pattern (see Figure 6.2) was due to the plausibility manipulation or rather related to the rating task itself, RTs were filtered based on two different thresholds (A or B) of the plausibility ratings assigned to the items of each condition. Specifically, RTs were excluded if items belonged to Condition A and were not rated as plausible (below 6; A or below 5; B), belonged to Condition B and were not rated as medium plausible (below 3 or above 5; A, B) or belonged to Condition C and were not rated as implausible (above 2; A or above 3; B). The filtered RTs based on the single-trial plausibility ratings as well as the corresponding estimated RTs and residuals are shown in Figure B.8. The majority of the removed items belong to Condition B but were rated as either plausible (above 5) or implausible (below 3), suggesting that the perceived plausibility of the items in Condition B varies the most, as previously also shown by the densities in Figure 6.1. The reason for this could be either the choice of the main verb, which could introduce more or less plausibility than intended, or simply a greater difficulty of classifying items of medium plausibility than items of rather high or low plausibility. After filtering the RTs based on the plausibility ratings, the RT pattern changed and items perceived as medium plausible that belong to Condition B are read even slower on average than items perceived as implausible that belong to Condition C, especially when considering range A (Figure B.8; left). The filtered RTs based on range A and B, as well as the unfiltered observed RTs, illustrate that the average RTs increase in Condition C and decrease in Condition B when the range is extended to include RTs of items rated as more implausible (and plausible) in Condition B and RTs of items rated as rather medium plausible (and plausible) in Condition C. This observation

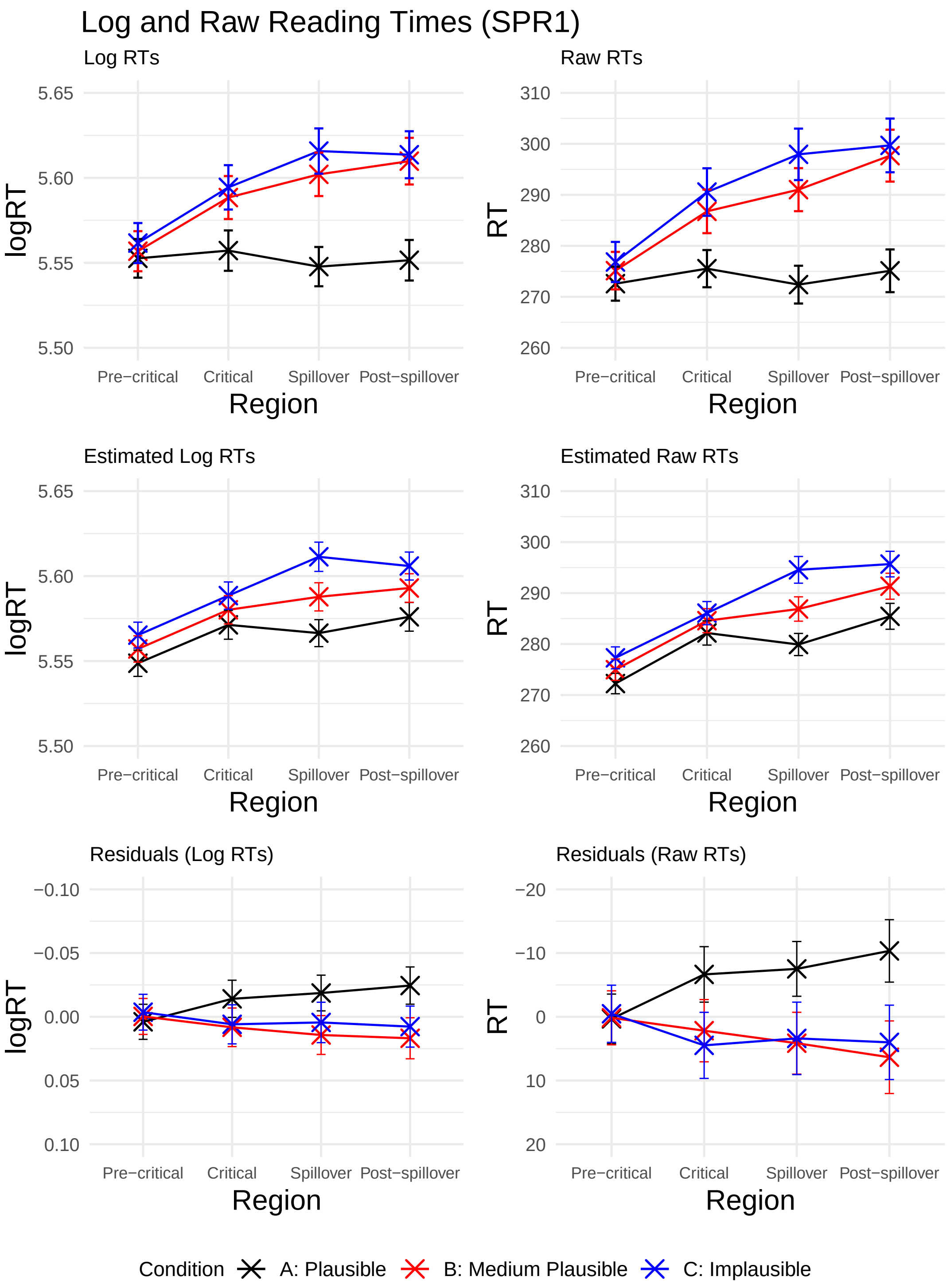

FIGURE B.7: Log reading times and raw reading times (top), estimated reading times using single-trial plausibility and GPT-2 surprisal (middle) and residuals (bottom) from the first SPR study per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate the standard error computed from the per-subject per-condition averages.

suggests that items rated as medium plausible are read more slowly, regardless of whether they belong to Condition B or C. Since the majority of, though not all, items rated as medium plausible correspond to Condition B, RTs are on average the slowest for Condition B when filtered by plausibility, indicating medium plausibility (3, 4, or 5). Consequently, when RTs are not filtered based on their respective plausibility ratings, RTs for Conditions B and C may be similar for different reasons: the generally higher degree of implausibility in Condition C and the increased difficulty of judging a medium level of plausibility, which is mostly present in Condition B, may lead to similar, increased RTs in both Conditions B and C compared to Condition A. Finally, the residual errors of the filtered RTs indicate that the models capture the patterns in the observed RT data relatively accurately for Conditions A and C, but strongly underpredict the RTs in Condition B.

RTs were also filtered based on the averaged plausibility ratings collected in the pre-test. Specifically, RTs were excluded if items belonged to Condition A and were not rated as plausible (below 5), belonged to Condition B and were not rated as medium plausible (below 3 or above 5) or belonged to Condition C and were not rated as implausible (above 3). The filtered observed RTs based on the averaged plausibility ratings, as well as the corresponding estimated RTs and residuals, are shown in Figure B.9. Similar to the RTs filtered by single-trial plausibility, the majority of the items excluded based on averaged plausibility (26 out of 33) belong to Condition B. Fewer RTs were excluded based on the average than based on the single-trial plausibility ratings, most likely because the averaged ratings per item are more robust and therefore correspond more reliably to the three plausibility levels of Conditions A, B, and C (which is also what makes them better predictors of the RTs). The pattern of the observed RTs filtered based on the averaged plausibility ratings $(A < B < C)$ is more pronounced, i.e., RTs in Condition B are notably lower than RTs in Condition C, compared to the originally observed RTs (see Figure 6.2) and the RTs filtered based on the single-trial plausibility ratings (see Figure B.8). As indicated by a smaller residual error, the models capture the observed RT pattern better when the variability in plausibility ratings is reduced, especially in Condition B, by filtering based on the averaged plausibility ratings than when RTs are not filtered or are filtered by single-trial plausibility ratings.

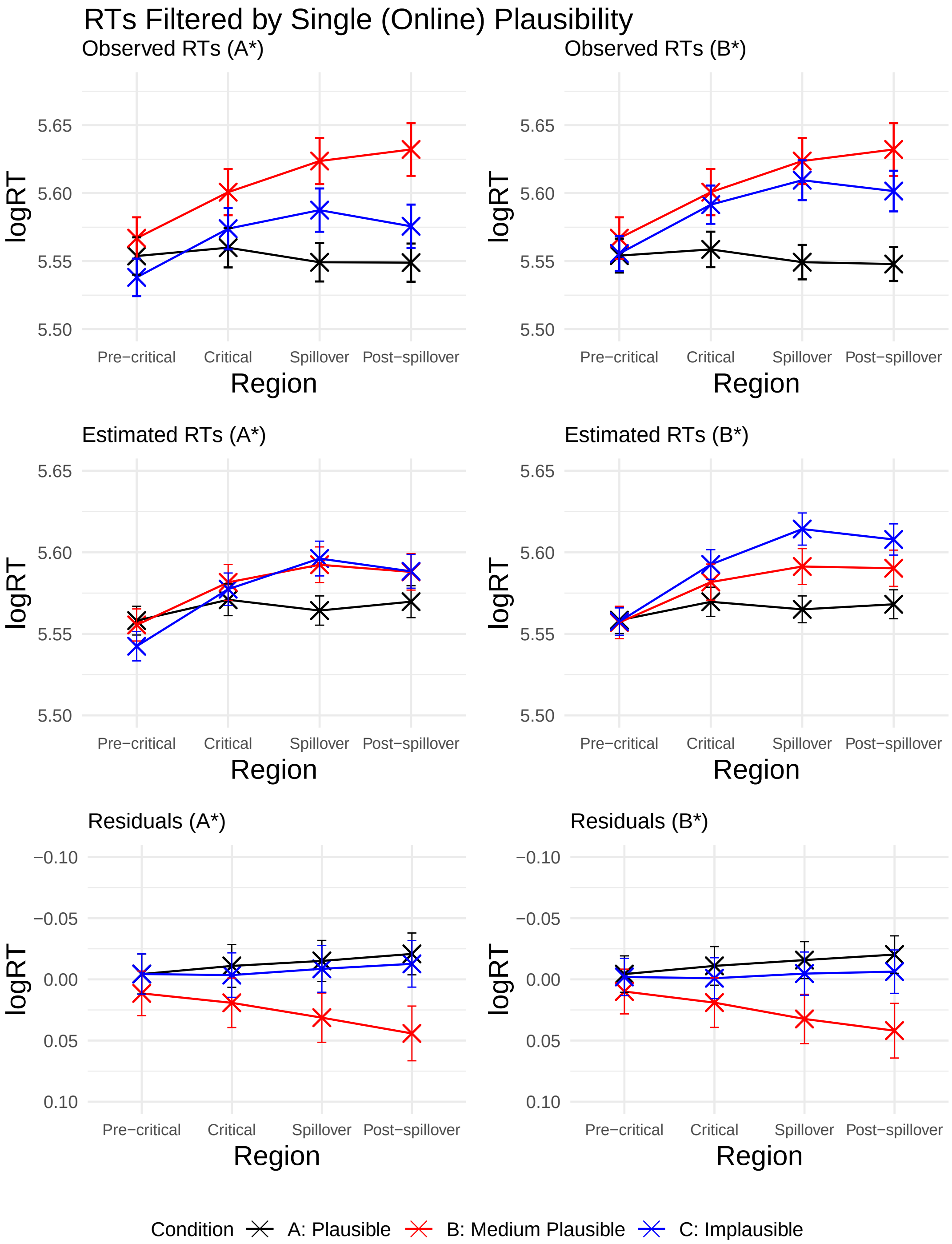

FIGURE B.8: Observed reading times (top), estimated reading times (middle) and residuals (bottom) per condition across all critical regions, filtered based on single-trial plausibility. This includes only reading times of items rated within range **A*** (**Condition A: 6, 7; Condition B: 3, 4, 5; Condition C: 1, 2)** or range **B*** (**Condition A: 5, 6, 7; Condition B: 3, 4, 5; Condition C: 1, 2, 3)**.

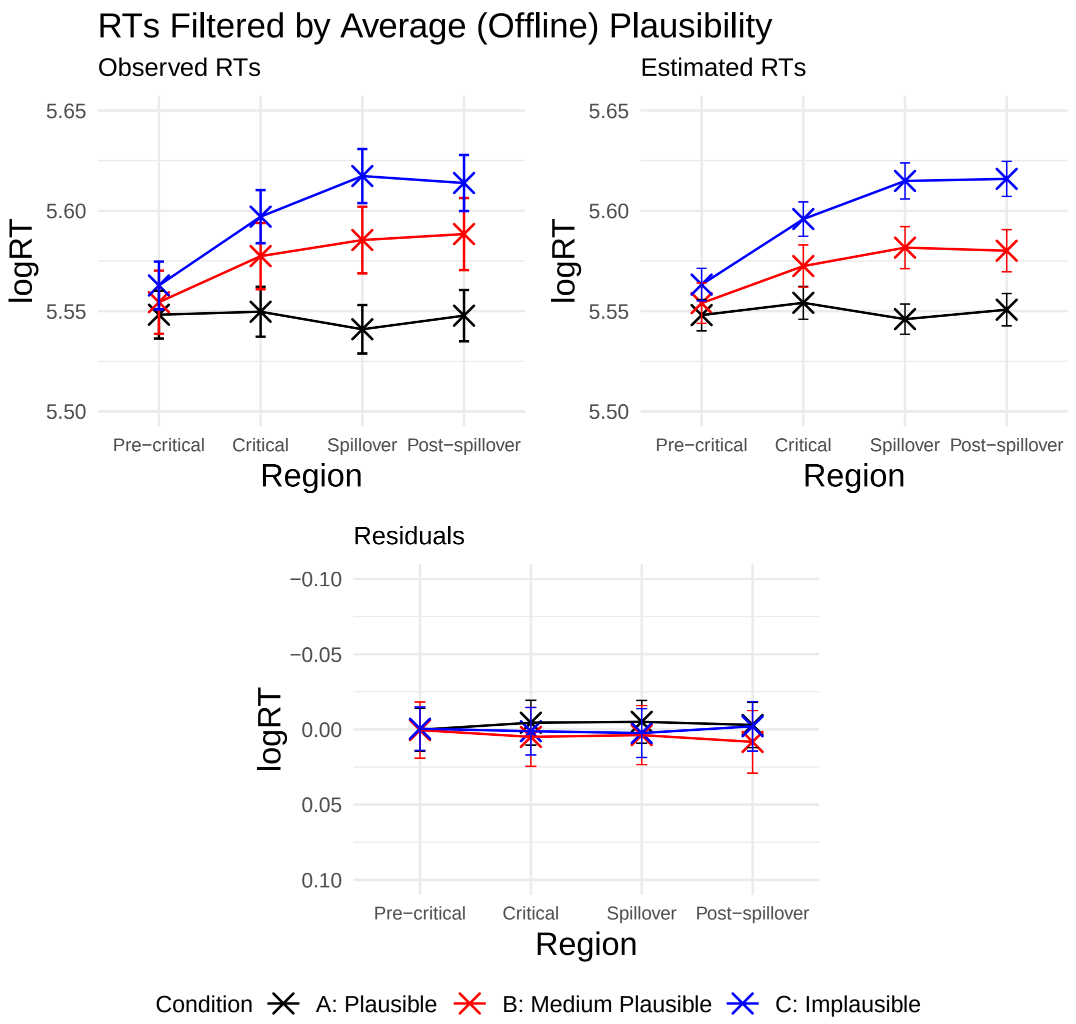

FIGURE B.9: Observed reading times (left), estimated reading times (middle) and residuals (right) per condition across all critical regions, filtered based on averaged pre-test plausibility. This includes only reading times of items rated within the range 5-7 (Condition A), 3-5 (Condition B), 1-3 (Condition C).

B.5 Grouping Reading Times by Plausibility

The previous analyses have highlighted the challenge of rating medium plausible items compared to highly plausible or implausible items. B.4 has shown that the averaged RTs in Condition B are higher when considering only items that were rated as medium plausible (assigned a plausibility rating of 3, 4 or 5), suggesting a potential influence of the online rating task on the RTs. To investigate whether the observed RT pattern (Figure 6.2) is related to the online rating task and whether increased RTs are specific to Condition B when focusing only on items rated as medium plausible, or whether items are affected regardless of their condition, the RT data were divided into three equal groups based on the single-trial plausibility ratings, rather than based on their respective conditions. Group 1 includes RTs from items assigned low plausibility ratings (1, 2, 3), Group 2 includes RTs from items assigned medium plausibility ratings (3, 4, 5), and Group 3 includes RTs from items assigned high plausibility ratings (5, 6, 7). This means that, for example, items in Group 1 were perceived as rather implausible, regardless of their condition, although the majority belonged to Condition C (571/838 items). Figure B.10 shows the RTs grouped by single-trial plausibility ratings, together with the estimated RTs and the corresponding residual errors. Items in Group 2, i.e., items associated with a medium level of plausibility, were clearly read slower on average than items in Group 3 or even items in Group 1, which contained only items perceived as implausible. This suggests that the inclusion of an online rating task in the self-paced reading study has a strong influence on the RTs. Specifically, the increased difficulty in assigning plausibility ratings to items rated as medium plausible may lead to slower RTs already on the word-by-word presented final sentence when anticipating the upcoming plausibility rating task.

RTs were also grouped based on the averaged plausibility ratings from the pre-test, although these ratings were not collected from the same participants as the RTs. 50/60 items in Condition C fall into Group 1, 41/60 items in Condition B fall into Group 2 and 51/60 items in Condition A fall into Group 3. This indicates that the items grouped by averaged plausibility mostly overlap with the items grouped by condition, especially in the case of Group 1 and Condition C and Group 3 and Condition A. However, Condition B and Group 2 exhibit higher variability in terms of the groups and conditions they include, even when considering the more robust averaged plausibility ratings. The observed RTs grouped by the averaged plausibility ratings, as well as the estimated RTs and residuals, are shown in Figure B.10. Interestingly, the RT pattern $A < B < C$ is more pronounced when the RTs are divided into three groups based on the averaged plausibility ratings from the pre-test than when they are grouped based on the three conditions. This suggests that the three groups formed based on the plausibility ratings assigned by the participants on average provide a clearer contrast between the RTs than the three plausibility levels created for Conditions A, B and C. In the future it may therefore be useful to review and possibly modify items that were rated above or below a certain threshold on average, indicating too high or too low plausibility for the respective condition.

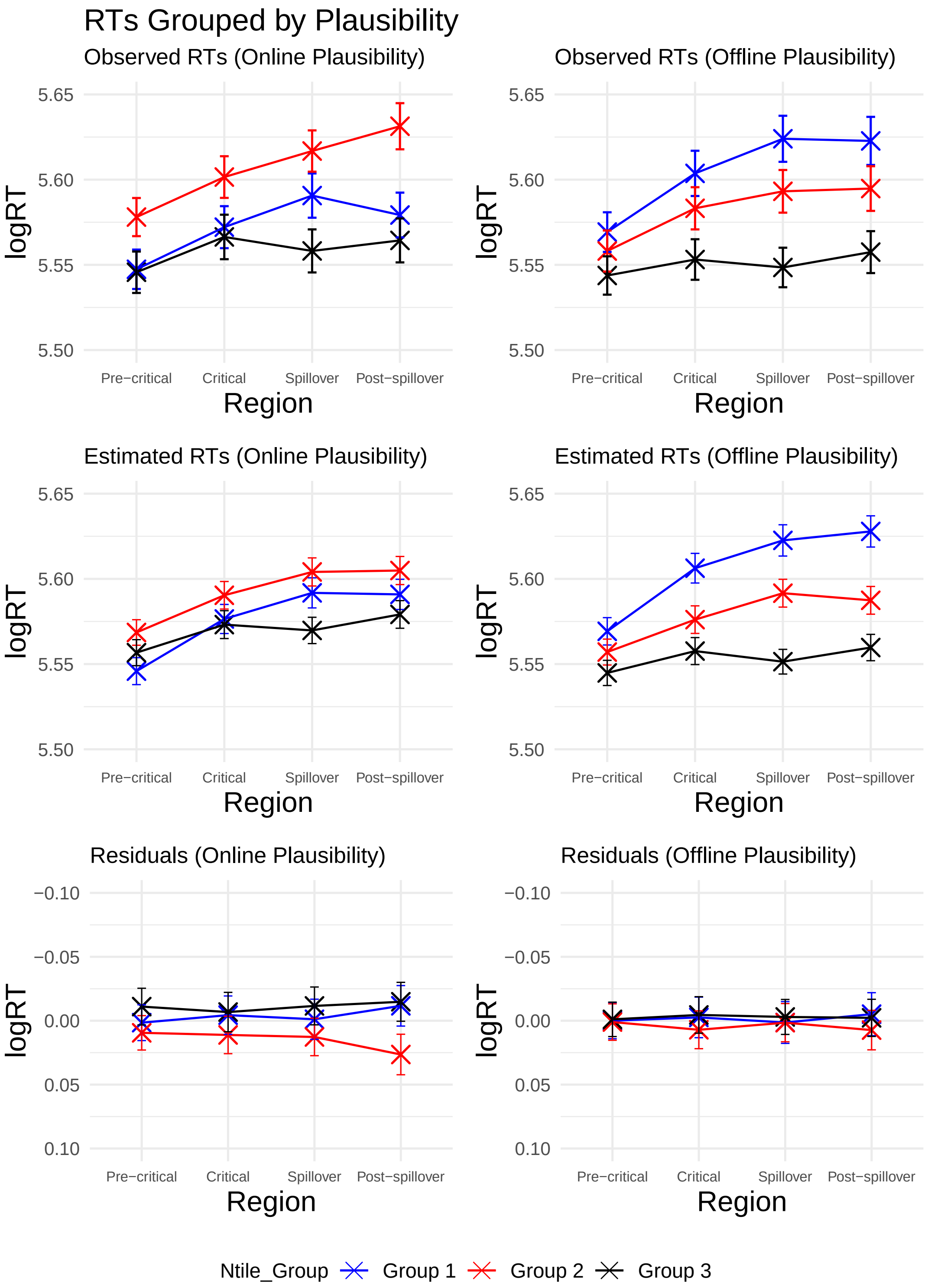

FIGURE B.10: Observed reading times (top), estimated reading times (middle) and residuals (bottom) grouped by single-trial plausibility (left) and by offline plausibility (right).

B.6 Measuring Reading Times on the Pre-Critical Regions

Figure B.11 shows the observed RTs on the four critical regions as well as on the main verb (e.g., "begrüßte" / "*welcomed*"), the third word preceding the target word (the determiner of the noun before the target word, e.g., "die" / "*the*") and the second word before the target word (a noun, e.g., "Dame" / "*woman*") from the first self-paced reading study (top), as well as from the second self-paced reading study (bottom).

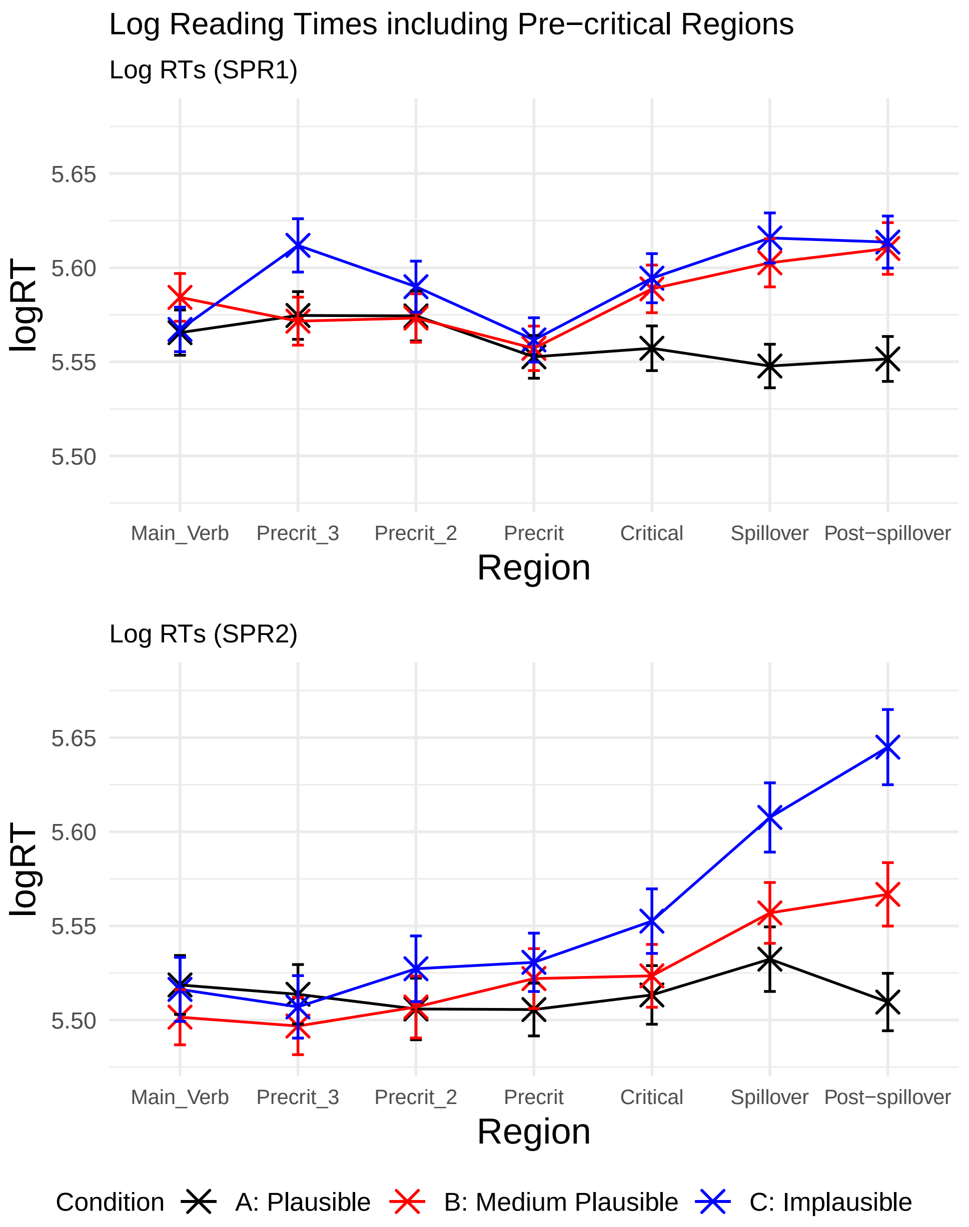

FIGURE B.11: Log reading times per condition on the main verb, Pre-critical3, Pre-critical2, Pre-critical, Critical, Spillover, and Post-spillover regions from the first (top) and the second (bottom) SPR studies. Error bars indicate the standard error computed from the per-subject per-condition averages.

In the first self-paced reading study, RTs on the main verb and the three pre-critical regions are similarly low in Conditions A and B, while RTs in Condition C sharply increase at the third pre-critical region (i.e., the region following the main verb) due to the implausibility introduced by the main verb. Subsequently, the RTs in Condition C decrease again and and become almost identical to those in Conditions A and B in the Pre-critical region. On the Critical region, that is, when reading the target word, RTs diverge again, particularly in Conditions B and C where they rise more sharply compared to Condition A. As discussed earlier, RTs in Conditions B and C increase to a similar extent, probably due to the online plausibility rating task, which leads to increased RTs especially for items rated as medium plausible. Simultaneously, the observed RT pattern in the pre-critical regions raises the question of whether RTs in Condition C are relatively low due to an attenuation of the plausibility effect on the target word by the implausibility of the main verb, which already led to increased RTs in the third pre-critical region.

In the second self-paced reading study, RTs are also already higher in Condition C than in Condition B from the main verb onwards and higher than in Condition A from the second pre-critical region onwards. However, there is no sharp increase in RTs in Condition C at the main verb or in any of the subsequent pre-critical regions. Only from the Critical region onwards, RTs significantly diverge, particularly in comparison to the RTs of the first self-paced reading study. The absence of a rapid increase in RTs in any of the pre-critical regions in Condition C, along with the more pronounced RT pattern in the regions following the target word in the second self-paced reading study, suggests that the increase in RTs on the third pre-critical region in the first study was likely due to the online plausibility rating task. While RTs in Condition C may be slightly higher in the pre-critical regions since the main verb introduces at least some degree of implausibility, it appears that the anticipation of the rating task primarily leads to slower reading after processing the main verb in Condition C in the first self-paced reading study. Furthermore, RTs are generally higher across all regions and conditions, including the pre-critical ones, in the first self-paced reading study, suggesting that participants generally read the sentences more slowly due to the online rating task.