Saarland University

Faculty of Humanities

Department of Language Science and Technology

Master's Thesis

Thesis Title
x
y

Author:
Eva Richter

Supervisors:
Prof. Dr. Matthew W. Crocker
Dr. Francesca Delogu

Advisor:
Dr. Christoph Aurnhammer

June 20, 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: E. Richter

Date: June 20, 2024

SAARLAND UNIVERSITY

Abstract

Faculty of Humanities
Department of Language Science and Technology

Master of Science

Thesis Title
x
y

by Eva RICHTER

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Psycholinguistic studies often assess how violations of variables like plausibility influence behavioural measures such as reading times (RTs) to draw conclusions about the processing difficulties they induce. However, the focus is usually not on the predictor variable itself, but rather on the impact it has on the outcome variable. Nonetheless, how well a specific effect can be captured depends on the operationalisation of the predictor variable as well. Typically, studies investigating the effects of plausibility rely on offline plausibility ratings collected in a pre-study, which are averaged over a group of participants and therefore do not correspond to the perceived plausibility of individual participants. While it has been argued that judgments averaged over a group of participants should be preferred to individual judgements as variation in judgment data represents mere noise rather than systematic differences (Featherston, 2007), others believe that this variation is structured instead of representing random noise and provides an important source of information (Verhagen et al., 2019). An interesting question, therefore, is whether plausibility ratings collected on each trial during an online self-paced reading experiment could capture RT effects more effectively, as they reflect individual participants' immediate perceptions of plausibility within the context of the ongoing reading task and can be used to account for individual processing differences. Therefore, the primary objective of this thesis is to investigate whether online plausibility ratings collected on each trial during a self-paced reading study or offline plausibility ratings averaged over a group of participants capture the effects in the observed RT data more accurately.

To address this question, the stimuli of the study by Aurnhammer et al. (2023), who examined the contrasting predictions of multi-stream models and Retrieval Integration (RI) theory, are used after a slight modification. Each stimulus consists of a context paragraph followed by three variations of a final sentence, varying in the plausibility of the target word $(A > B > C)$ and the low (A, C) or high (B) expectancy of the distractor word. After lowering the expectancy of the distractor word in Condition B, two pre-studies, collecting plausibility ratings and computing surprisal values with GPT-2 and LeoLM are conducted to assess the expectancy and plausibility levels across conditions. If the three plausibility levels of Conditions A, B and C are maintained, RTs should be graded for plausibility, similar to the the study by Aurnhammer et al. (2023). During the self-paced reading study participants are additionally asked to provide ratings on each trial, which are used in a subsequent regression-based analysis to model RTs as a function of target word plausibility and distractor word surprisal, allowing for comparison between RT predictions using online or offline plausibility ratings. Although this is not a primary research question, the study also allows for a comparison between GPT-2 and LeoLM distractor word surprisal as a predictor of RTs. However, if the manipulation effectively lowers distractor word expectancy, no RT effect due to distractor word expectancy may be observed. Figure 1.1 provides an overview of the design and steps involved in the current study.

The following 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-studies; 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.

better to put last part and figure in methodology chapter?

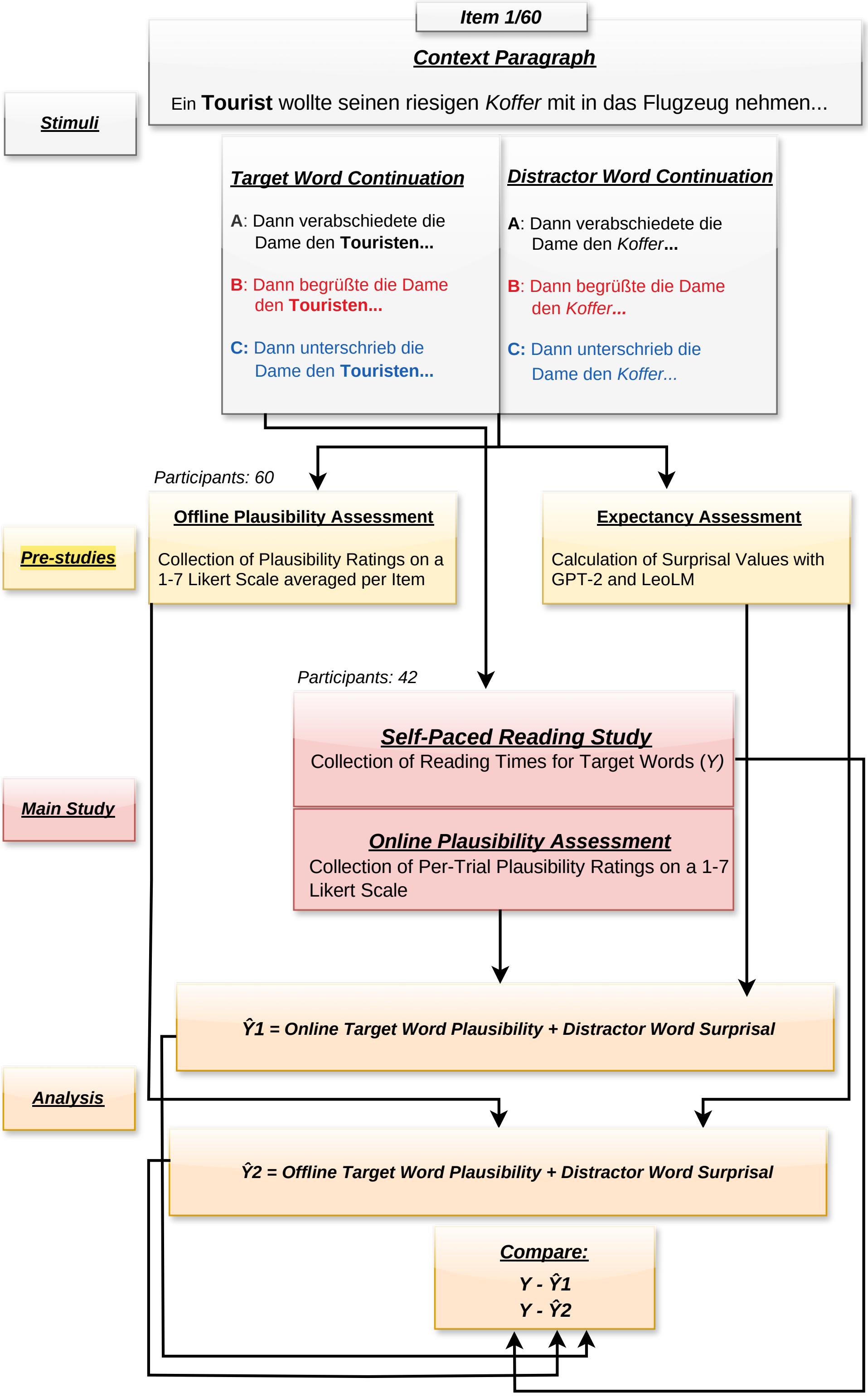
Item 1/60
Context Paragraph
Ein Tourist wollte seinen riesigen Koffer mit in das Flugzeug nehmen...
Stimuli
Target Word Continuation
A: Dann verabschiedete die Dame den Touristen...
B: Dann begrüßte die Dame den Touristen...
C: Dann unterschrieb die Dame den Touristen...
Distractor Word Continuation
A: Dann verabschiedete die Dame den Koffer...
B: Dann begrüßte die Dame den Koffer...
C: Dann unterschrieb die Dame den Koffer...
Participants: 60
Pre-studies
Offline Plausibility Assessment
Collection of Plausibility Ratings on a 1-7 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
Ŷ1 = Online Target Word Plausibility + Distractor Word Surprisal
Ŷ2 = Offline Target Word Plausibility + Distractor Word Surprisal
Compare:
Y - Ŷ1
Y - Ŷ2

FIGURE 1.1: Overview of the experimental design: Preparation of the stimuli, assessment of the stimuli manipulations in two pre-studies, implementation of a self-paced-reading study and linear mixed effects regression analysis using pre-study and single-trial plausibility together with distractor word surprisal to predict the observed reading time data.

Chapter 2

Background and Related Work

2.1 Theories of Language Comprehension

Language comprehension in the brain can be investigated and observed through behavioural and electrophysiological measures. Behavioural measures such as reading times (RTs) record the participant's behaviour in response to tasks such as lexical decision or self-paced reading. While these measurements are relatively straightforward to collect, they only provide an approximation of overall processing effort. In contrast, electrophysiological measures, specifically event-related potentials (ERPs), are more time-consuming and costly to collect, however, they offer high temporal resolution and information about the cognitive processes that underlie language comprehension. While this study focuses on using RTs to measure the effort involved in processing stimuli with varying degrees of plausibility and expectancy, this chapter also discusses different theories associated with the ERP components, as a future ERP study based on the current materials could investigate the mechanisms driving the observed processing effort and provide evidence for different models of language comprehension, such as multi-stream models and Retrieval-Integration (RI) theory.

The two most prominent ERP components studied in the context of human language comprehension are the N400, a negative-going voltage deflection peaking approximately 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 regarded as an index of semantic integration processes (Brown and Hagoort, 1993), after Kutas and Hillyard (1980) for the first time observed an increased N400 amplitude [1] in response to sentences ending with unexpected compared to expected words (e.g. "he spread his warm bread with **socks**" compared to "**butter**") (see Kutas and Federmeier, 2011, for an overview). The P600, on the other hand, was originally observed in response to unpreferred sentence structures (Osterhout and Holcomb, 1992) and syntactic violations (e.g. "the spoilt child **throw** the toys on the floor" compared to "**throws**)" (Hagoort et al., 1993) and therefore traditionally linked to syntactic processing. However, these interpretations are not compatible with several studies that 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

[1] Note that an N400 activity is elicited in response to all meaningful stimuli, however, its amplitude has been found to be increased for anomalous items relative to non-anomalous items.

the interpretation of the language-sensitive ERP components and resulted in multi-stream models (Kupferberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008) and Retrieval-Integration (RI) theory (Brouwer et al., 2012, 2017).

Multi-stream models propose that different cognitive mechanisms trigger either an N400 or a P600 effect, depending on whether an anomaly is repairable by the presence of a semantically attractive alternative. More precisely, these models rely on the assumption that there are at least two separate processing streams: A processing stream of a semantic nature, which does not detect an anomaly if it is repairable by the availability of an attractive alternative interpretation and a second processing stream, guided by syntactic principles, which detects the anomaly. This conflict between the (at least) two independent processing streams triggers a P600 effect. In contrast, they predict an N400 effect when no semantically attractive alternative is available, rendering the anomaly irreparable (Kim and Osterhout, 2005; Kupferberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008).

In recent years, the N400 has been reinterpreted as a measure of lexical retrieval. According to this perspective, both lexical and contextual priming contribute to the pre-activation of features associated with the upcoming words. The more consistent the features of an encountered word are with the features pre-activated in memory, the greater the ease of lexical retrieval of the word from long-term memory, which is reflected in a reduced N400 amplitude (Federmeier and Laszlo, 2009; Kutas and Federmeier, 2000, 2011). This assumption was adopted by Brouwer et al. (2012, 2017) as the *retrieval* part of the RI account, while the P600 component is associated with the semantic integration of an incoming word with the current interpretation of an 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, under this account, the P600 component demonstrates sensitivity 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. In contrast, the N400 component is not directly sensitive to plausibility under this account since the retrieval of both plausible and implausible words is facilitated when they are primed by the preceding context. In contrast, RI theory predicts that both the N400 and P600 components are sensitive to expectancy, as unexpected words are more challenging to retrieve and integrate, respectively, which is supported by the findings of (Aurnhammer et al., 2021). However, retrieval of unexpected words is facilitated and N400 amplitude is reduced when the target words are sufficiently primed, for example, by the context or lexical repetition (Brouwer et al., 2012, 2017).

In a context manipulation design contrasting the assumptions of multi-stream models and RI theory, Aurnhammer et al. (2023) found that both RTs and P600 amplitude pattern with plausibility. Given the absence of a semantically attractive alternative in the implausible condition, multi-stream models predict an N400 effect since the anomaly is irreparable. However, the findings provide evidence for RI theory, which predicts that both RTs and P600 correlate with plausibility due to the higher effort involved in the integration of less plausible words. These results (1) suggest a strong correlation between plausibility, RTs, and surprisal and (2) establish the P600 as a continuous, word-by-word index of integration effort.

2.2 Expectancy

Readers or listeners process language continuously and incrementally, more or less word by word (Tanenhaus et al., 1995), meaning that each word or unit of language is processed and integrated into the overall understanding of the sentence or discourse as it is encountered, rather than after perceiving the whole utterance. 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 (Kupferberg et al., 2020). Although the functions and even the names [2] referring to the concept of expectancy in language comprehension are not uncontroversial (for a discussion Van Petten and Luka, 2012; Kupferberg and Jaeger, 2016), 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), demonstrating increased N400 amplitudes for unexpected words compared to expected words, provide evidence that unexpected words require more processing effort than expected words.

The most widespread 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 specific 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 greater predictability of the missing word in this 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 (Brothers and Kupferberg, 2021; Aurnhammer et al., 2021) and reduced N400 amplitudes (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011) for words with higher cloze probabilities.

A more recent operationalisation of predictability is surprisal, which corresponds to the negative logarithm of cloze probability. Surprisal estimates are computed by LMs and differ mainly from cloze probabilities in that their predictions capture only the statistics of the language and - at least not explicitly - word meaning with respect to world knowledge. Precisely because of their subjective nature, which is also inherent in cognitive measures such as RTs, cloze probabilities have previously been found to be a better predictor of behavioural correlates of processing effort such as word RTs compared to LM-derived estimates (Smith and Levy, 2011; Brothers and Kupferberg, 2021). However, cloze probabilities are more time-consuming and expensive to collect and provide unreliable estimates for low-probability words, since huge sample sizes are needed to ensure that less expected continuations are at least produced once, i.e. don't result in a zero cloze probability despite being

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

plausible continuations (Smith and Levy, 2013). In contrast, surprisal estimates generate probability distributions across the entire vocabulary, capturing high and low-probability words to the same extent. Furthermore, Michaelov et al. (2022) point out that LM surprisal offers a precise method for assessing the degree to which linguistic input alone can predict measures of language comprehension because it isolates the specific impact of the linguistic input. Since there are contrasting findings as to whether cloze probability or surprisal estimates predict behavioural and electrophysiological human data more effectively (Smith and Levy, 2011; Shain et al., 2020; Michaelov et al., 2022), it can be concluded that there are different ways of operationalising the expectancy of a word in context, each of which has its strengths and weaknesses. As only surprisal estimates are used as predictor for RTs in the context of this thesis, the subsequent section will discuss only surprisal in greater detail.

2.2.1 Surprisal

Surprisal theory is an expectation-based processing theory that draws upon principles from information theory (Shannon, 1948) and has proven effective in explaining word-by-word processing difficulty (Hale, 2001; Levy, 2008). It is based on the assumption that each word carries a certain amount of information, predictive of the cognitive effort required to process the word. The processing effort is proportional to the surprisal of the word, which itself is inversely proportional to the expectancy of a word. Accordingly, higher cognitive effort is reflected in higher surprisal, corresponding to lower expectancy of a word in a given context. In more formal terms, 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 language models, the amount of cognitive effort which 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). Specifically, 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) as well as surprisal of parts-of-speech (POS) (Demberg and Keller, 2008; Frank and Bod, 2011). Accordingly, words (or POS) that carry more information, as indicated by higher surprisal values, are read more slowly compared to words that are less informative, corresponding to lower surprisal. Importantly, Monsalve et al. (2012) found that significant RT effects might be missed when spillover regions are not taken into account, i.e. when surprisal is only analysed in relation to the current item without considering its influence on the following item. In electrophysiological research, surprisal was 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 not been thoroughly studied in this context yet.

Finally, the accuracy of language models in predicting RT or EEG data depends not only on the characteristics of the training corpus and the analyzed elements (such as words versus POS) but also on the architecture of the language model itself. Various architectures have emerged and been investigated over time for estimating surprisal values in the context of sentence processing (e.g. Probabilistic Context-Free Grammars (PCFGs), N-gram models, Simple Recurrent Networks (SRNs)). The more recently developed Transformer model (Vaswani et al., 2017), has outperformed previous LMs in various NLP tasks and is being increasingly 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 RT (Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Wilcox et al., 2020) and EEG (Michaelov et al., 2021, 2023) data and have outperformed even improved types of RNNs in explaining both RT and N400 data from word-by-word reading experiments (Merkx and Frank, 2021), despite being cognitively less plausible (Michaelov et al., 2021).

However, contrasting findings have also been reported specifically for RTs. For instance Oh and Schuler (2022) observed that surprisal estimates derived from variants of the pre-trained GPT-2 LM, with more parameters and lower perplexity, are less predictive of self-paced RT data during naturalistic reading. Further investigations have shown that as the model size increases, the degree of underprediction, particularly at open-class words (nouns and adjectives), increases (Oh and Schuler, 2023) and that this relationship is most pronounced within the subset of least frequent words (Oh et al., 2024). In fact, these predictions may be accurate given that large LMs acquire substantial domain knowledge through training on large datasets and develop the ability to predict even rare words in later stages of training. However, Oh and Schuler (2023); Oh et al. (2024) assume that the fact that these models are trained with non-human learning objectives on vast amounts of text, which are not accessible to humans, may render them less suitable for cognitive modeling.

As mentioned, the operationalisation of expectancy as cloze probability has previously yielded better results for behavioural RT data than LM surprisal (Smith and Levy, 2011; Brothers and Kupferberg, 2021). More recently, however, sophisticated Transformer-based LMs such as GPT-3 have been found to be better predictors of RTs and N400 data than cloze probabilities (Hofmann et al., 2022; Michaelov et al., 2022, 2023), suggesting that the cognitive processes reflected in measures of language comprehension may be more shaped by the statistical properties of language than previously thought (Michaelov et al., 2022). Moreover, contrasting results across studies can often be attributed to differences in experimental designs, stimuli, and participant demographics. Specifically, the predictions underlying the behavioral and neural metrics of processing difficulty may interact differently with word expectancy operationalised as cloze probability or surprisal, especially considering the temporal dimension (Michaelov et al., 2022). At the same time, the results of Michaelov et al. (2022) are not inconsistent with previous studies in which cloze probability was shown to be a better predictor of processing difficulty, as Michaelov et al. (2022) only found surprisal estimates of the recent Transformer-based LMs to be a better predictor of processing difficulty, not the RNN-based surprisal estimates used by Smith and Levy (2011); Brothers and Kupferberg (2021).

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 reflect only statistical patterns of language, but not extralinguistic factors such as world knowledge, which have been shown to influence language processing beyond linguistic experience (e.g. Hagoort et al., 2004; Niewland et al., 2007). One exception is the model of language comprehension by Venhuizen et al. (2019), which instantiates a *comprehension-centric* notion of surprisal by incorporating both linguistic experience and world knowledge by deriving a Distributed Situation-state Space (DSS). While other studies (Frank et al., 2015; Michaelov and Bergen, 2020; Michaelov et al., 2022; Merkx and Frank, 2021) have found a correlation between purely linguistic surprisal and the amplitude of the N400 ERP component, *comprehension-centric* surprisal is predicted to reflect P600 amplitude (Venhuizen et al., 2019) under the RI account. Within this framework, *comprehension-centric* surprisal quantifies the likelihood of transitioning from one point in situation-state space to the next while P600 amplitude reflects the neurophysiological processing effort that is associated with this transition, i.e. longer transitions correspond to higher surprisal and lead to increased P600 amplitudes, reflecting the increased processing effort of integrating the meaning of this word with the meaning of the unfolding utterance representation. The word-by-word estimates produced by the neurcomputational model of incremental language comprehension of Brouwer et al. (2021) provide evidence that RTs and P600 amplitude increase in response to implausible compared to plausible target words, supporting a qualitative link between RTs, P600, and *comprehension-centric* surprisal. The findings of Aurnhammer et al. (2023) moreover show that this link also holds quantitatively, in that RTs and the P600 have been shown to index integration effort continuously.

2.3 Plausibility

Although plausibility is a widely used concept whose effects have been studied in many cognitive contexts (Rayner et al., 2004; Reder, 1982; 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 seems to be 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, 2010). Considering, for example, 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", the former is likely to be judged as more plausible since it aligns with most people's experience that bottles break rather than melt when dropped (Connell and Keane, 2004). Higher perceived plausibility is therefore linked to how much a situation or statement aligns with our 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 that is carried out in a pre-study. Participants are presented with experimental items one by one and are asked to provide a plausibility judgment for each item, usually on a Likert scale from 1 (indicating that an event is highly implausible) to 7 (indicating that an event is highly plausible) (Niewland et al., 2020; Haeuser and Kray, 2022; Delogu et al., 2021; Aurnhammer et al., 2023; Michaelov et al., 2023). Although Likert scales belong to the most widely used response formats for measuring attitudes and opinions in psycholinguistics (and other fields), the assignment of response choices such as "highly plausible", "less plausible", "implausible" to numerical values is often criticised, since it suggests equal differences between categories simply because the numerical differences are equal although the true differences between responses may not be equal (Knapp, 1990). Recently, LMs have been used to generate plausibility judgements on a scale from 1 to 7 as well. Amouyal et al. (2024) found a high correlation between GPT-4 and human plausibility judgements and concluded that LM-generated plausibility ratings are as effective as human judgments for coarse-grained ratings, but less reliable for fine-grained judgments.

As Niewland et al. (2020) note, it is challenging to distinguish plausibility effects from predictability effects, as less plausible stimuli are generally also less predictable. For example, in the sentence "the bottle rolled off the shelf and smashed/melted on the floor", the verb "smashed" renders the sentence not only more plausible than "melted", but "smashed" is also a more predictable continuation compared to "melted". One way to think about the differential notions of these related concepts is that predictability measures the likelihood of a specific word occurring at a particular sentence position based on the preceding context, while 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 not affected by any post-target continuation, which is usually included to capture spillover effects in RTs, while plausibility is not conditional in nature and therefore can vary depending on the post-target continuation. Most studies, however, 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 considered as an advantage in some contexts. Michaelov et al. (2023) point out that given the limited differentiation 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 plausibility from predictability effects, most studies have investigated their effects separately (Rayner et al., 2004; Warren et al., 2008). However, as pointed out by Matsuki et al. (2011), the mean cloze probability in these studies is not exactly zero and not identical for the plausible and implausible items. As they note, the challenge lies in creating stimuli with a cloze probability of zero or to use only items that are matched for cloze probability but differ in their plausibility rating. More recently, 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) revealed 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 results from ERP studies which found an earlier effect of predictability and plausibility either in the N400 time window (Niewland et al., 2007) or in even 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 findings of Brouwer et al. (2021), that the N400 component is primarily sensitive to expectancy and, at least not directly, sensitive to plausibility since there are no differences in N400 amplitude between words that render a sentence plausible or implausible, if they are equally primed. The later-emerging P600 component, on the other hand, is sensitive to plausibility violations, reflecting the increased effort required to integrate words that render a sentence implausible with the unfolding utterance representation. According to Brouwer et al. (2012, 2021), behavioral measures like RTs correlate with P600 amplitude on a word-by-word basis, suggesting that RTs are also sensitive to plausibility violations in a graded manner. This was confirmed by Aurnhammer et al. (2023) who observed graded RTs for plausibility, with less plausible items being read more slowly on average compared to more plausible items, reflecting the continuous integration effort involved in updating the unfolding utterance representation with words that render the sentences implausible.

2.4 Individual Variability in Language Comprehension

Many psycholinguistic studies have explored individual differences in language processing (Boudewyn, 2015, for an overview), as these provide valuable insights into the functioning of the language system and a better understanding of language disorders and difficulties. Often studies investigated individual differences in general cognitive processes, such as working memory (Nakano et al., 2010) and cognitive control (Boudewyn et al., 2012), in language proficiency, such as second-language learning (McLaughlin et al., 2004, 2010) or individual differences related to age (Federmeier et al., 2010) or gender (Payne and Lynn, 2011). Another 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. Troyer and Kutas (2018) recorded an EEG during which participants with varying levels of HP knowledge read sentences related to HP and about general topics that ended either with contextually supported or unsupported words. N400 amplitudes were reduced to supported endings for both types of sentences, but varied depending on participants' HP knowledge only for HP-related sentences, with larger effects observed in more knowledgeable individuals compared to smaller effects in less knowledgeable individuals, providing evidence that N400 context effects vary based on individual knowledge levels. This study, along with the previously mentioned studies, take into account variability between subjects, which can be systematic and useful to identify meaningful subgroups of language users, but they do not take into account variability across single trials.

Focusing on average changes in behavioural or neural measures implies that experimental manipulations affecting cognitive processing remain stable over the duration of an experiment. According to this approach, any variation or fluctuation that may occur in repeated measurements of the same individual is considered 'noise' - random variations that may mask the true effects of the experimental manipulation (Payne and Federmeier, 2017). Although intraindividual variability, measuring variability within individuals across trials, has been studied less extensively, there is evidence that trial-to-trial variability can provide distinct insights into cognitive processing beyond average measurements, for example as an indicator of impaired cognitive functions in individuals with different disorders (Hutsch, 2000; Dinstein et al., 2015). In behavioural psycholinguistics, most studies taking into account trial-level variability concentrated on how the effects of experimental manipulations affect the shape of the underlying RT distributions, demonstrating that language processing system does not always react consistently to linguistic difficulty across all trials of an experiment (Payne and Federmeier, 2017).

Single-trial measurements can also be operationalised as a predictor variable to gain a more granular understanding of how specific factors influence the (behavioural or neural) response of individuals on each trial. Returning to the study by Troyer and Kutas (2018), participants were categorized based on high or low HP knowledge. However, the study did not take into account which (and how many) facts each individual knew. Therefore, they could only assume that N400 amplitudes of individuals with higher HP knowledge were decreased for supported words because they likely knew more facts. To investigate whether the observed pattern was a result of the proportion of facts an individual knew, or if those with greater HP knowledge also knew more facts, and whether this was reflected in the 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 if they had known the fact or not. Based on this single-trial design they showed that the proportion of trials which participants knew was highly correlated with their HP domain knowledge and served as a strong predictor of their N400s to supported endings of HP sentences. Crucially, the results from Troyer et al. (2020) were consistent with those of Troyer and Kutas (2018), showing that domain knowledge correlates with a decrease in N400 amplitude to contextually supported sentences, suggesting that the N400 effects were not influenced by task effects based on the participant reports. Moreover, HP knowledge had the greatest influence when participants did not know a fact, suggesting that domain knowledge especially has an impact when retrieval is difficult. Therefore 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 on 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), while N400s to unsupported words did not vary based on participants' reports, indicating that domain knowledge affects the information brought to mind during language processing in a broader sense.

The findings of both Troyer et al. (2020) and Troyer and Kutas (2020) show that single-trial analyses are useful to establish a direct link 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, which would be missed in aggregated data analyses.

Although plausibility is a commonly manipulated variable in many studies, no (I did not find any?) study has investigated the effects of plausibility at the trial-level. As discussed in Sections 2.3 and 2.2, human judgment 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 align with the perception of a specific individual. Since different participants may respond to the same item differently, the range of variation increases. Whether this variation is simply noise that should be ignored by averaging the judgments, or whether it contains useful information that can explain individual variation in measures of language comprehension, has not been widely investigated in (psycho)linguistic research. Featherston (2007) states that differences in individuals' judgments are noise inherent in the judgment process. Specifically, he argues that variation in judgements primarily arises from the inherent noisiness of individuals' judgements rather than from systematic differences (reflecting individual variations in grammars). Therefore, comparing individuals' judgments increases the error variance because each individual introduces their own noise, and the variability in each judgment can differ, even in opposite directions. Instead, the mean judgments of a group of individuals should be considered, since "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). An alternative perspective is that "variability is structured rather than random" (Foulkes, 2006). Verhagen et al. (2019) examined variation not only across participants but also across items, time, and methods using 7-point Likert scale or Magnitude Estimation scale judgment data and concluded that in metalinguistic judgements variation 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, variation in actual judgements is attributed to characteristics of language use and linguistic representations and represents valuable information instead of just noise. However, this would require a thorough investigation to identify and control for all relevant factors that distinguish real from non-real ratings, which might not always be possible. Verhagen et al. (2019) point out that this does not mean that there is no unexplained variance in the data when considering individual judgements, but rather that analyzing this variance can reveal meaningful information. This suggests that the choice between considering or ignoring individual variation in judgments (or any data in general) may depend on the specific goals of the study. If the primary interest is in identifying patterns and factors that drive differences in language comprehension, individual judgments may be more insightful given their granularity. However, if the goal is to find the best operationalisation in terms of minimizing unexplained variance, this might not necessarily be the case.

Chapter 3

Research Questions

The extent to which a behavioural or neural effect is captured in a (psycholinguistic) experiment depends on many factors, including the appropriate operationalisation of the predictor variable that is expected to influence language processing. In most cases, however, the predictor variable is not studied as such but is instead just linked to the outcome variable, reflecting the processing effort associated with it. Several studies have examined the effects of plausibility violations on sentence processing in various contexts (Haeuser and Kray, 2022; Matsuki et al., 2011; Brothers et al., 2020; DeLong et al., 2014; Aurnhammer et al., 2023), relying on offline plausibility ratings collected in a pre-study. Since these offline plausibility ratings are averaged for each item across all participants, they may not necessarily reflect the perceived plausibility of individual participants. Moreover, offline plausibility ratings (similar to cloze responses) can be distorted by "conscious reflection and other strategic effects" (Smith and Levy, 2011) since they are collected in a rather unnatural, untimed task. Since plausibility ratings collected on a by-trial basis provide a more precise measure of the perceived plausibility or difficulty of each individual on every trial, this raises the question of how effectively they can predict the processing effort implicit in behavioral measures such as reading times. Thus, the first research question of this thesis 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?

Previous studies employing single-trial analyses collected participants' trial-level responses during (Troyer et al., 2020) or after (Troyer and Kutas, 2020) EEG experiments and showed that they are strong predictors of their ERPs. Similarly, this thesis aims to collect participants' responses (ratings) on a by-trial basis, but with a different focus: the primary goal is to assess whether a more fine-grained, trial-level operationalisation of plausibility is a stronger predictor of processing effort, as reflected in reading times, compared to offline plausibility ratings averaged over a group of participants. Given that single-trial plausibility ratings are collected from the same participants performing the reading task, individual differences in perception and processing are directly accounted for, which may lead to more accurate reading time predictions. Conversely, offline plausibility ratings might capture effects in reading times more accurately as they are less affected by variability and random fluctuations in participants' ratings during the reading task, thus offering a more general and stable measure of plausibility.

To explore this question, the stimuli from the study by Aurnhammer et al. (2023), designed to test the conflicting predictions of multi-stream models and RI theory, are used for the current study. The items differ in the plausibility of the target word $(A > B > C)$ and the low (A, C) or high (B) expectancy of the distractor word in the final sentence following a context paragraph. Since this study aims to investigate only the graded reading time effects of plausibility, the expectancy of the distractor word in Condition B is lowered by modifying the main verb. 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 a lower distractor word expectancy?

If the main verb in Condition B was chosen such that it renders the final sentence given the preceding context in Condition B less plausible than in Condition A but more plausible than in Condition C, reading times should be graded, reflecting the increased effort of integrating less plausible content.

Additionally, two Transformer-based LMs, differing primarily in their size of parameters and training data, were used to compute surprisal values for the target and distractor words in the final sentence to assess whether distractor word expectancy in Condition B (as well as Conditions A and C) is lower than target word expectancy after modifying the main verb. While this is not directly tied to a specific research question, it may be interesting to observe whether including GPT-2 or LeoLM surprisal as a predictor affects the reading time estimates. Given that Aurnhammer et al. (2023) did not observe a significant modulation of RTs due to distractor cloze probability (even though it was high in Condition B), this suggests that distractor word expectancy will not significantly predict RTs in the current study as well. Nonetheless, there may be differences in RT predictions depending on the choice of LM, given the difference in size and the operationalisation of expectancy as surprisal rather than cloze probability in this study.

Furthermore, the current design and stimuli provide the opportunity to investigate two additional research questions following the study by Aurnhammer et al. (2023), which are discussed in more detail in Chapter 8. Since the distractor word expectancy was lowered in Condition B, the distractor word no longer serves as a semantically attractive alternative. Consequently, (1) the predictions of multi-stream models and RI theory diverge not only in Conditions C, but also in Condition B, and (2) the negativity observed in Condition B around 250-400 ms post-stimulus onset might disappear in an EEG study based on the current stimuli, if indeed it was triggered by the unfulfilled expectation of the distractor word.

Chapter 4

Materials and Methodology

This section describes the structure of the stimuli on which the pre-studies and the self-paced reading experiments are based, the architecture of the language models used to compute surprisal to assess target and distractor word expectancy and the linear mixed effects regression re-estimation technique used to analyse the RT data.

4.1 Stimuli

The stimuli used for the following pre-studies and 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 stimuli from Nieuwland and van Berkum (2005) and in some cases developed completely new items. The 60 best items selected by Aurnhammer et al. (2023) based on the results of a cloze task were also selected for the current pre-studies and the main experiment after slight modification (see Appendix A for the full list of German stimuli).
In the original design of Nieuwland and van Berkum (2005), a context paragraph is followed by either a coherent continuation containing a plausible target word ("the woman told the tourist") or a continuation rendered implausible by an implausible target word ("the woman told the suitcase").

Aurnhammer et al. (2023) changed the target manipulation design of Nieuwland and van Berkum (2005) to a context manipulation design to directly test the contrasting predictions of multi-stream models and RI theory. In this context manipulation design, each item consists of a context paragraph followed by a manipulated final sentence. The main verb of the final sentence was chosen in such a way that it renders the target word of the final sentence 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"). In this way, Aurnhammer et al. (2023) could test whether RTs and P600 amplitude are graded for plausibility, and, as predicted by RI theory, continuously index integration effort. Additionally, the main verb of the final sentence in Condition B was chosen in such a way that the expectancy of a distractor word is higher than the expectancy of the target word. That means, in the final sentence "the lady weighed the tourist", the distractor word "suitcase" is globally available as a semantically attractive alternative, although it never appears in target position. This design allowed them to test the predictions of multi-stream models and RI theory, that are similar (although for different reasons) in Condition B and diverge in Condition C, where the distractor word does not serve as an attractive alternative that could explain a P600 effect.

In the 60 items selected from Aurnhammer et al. (2023) the target word is the same across conditions to minimize potential effects due to word length or word frequency. To make sure that the entire main verb can be integrated with the preceding context prior to reading the target word, no separable verbs (e.g. "*Dann **bereitete** der Mann das Essen **zu***") were used. Moreover, 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 only arises when reading the target word and not based on the combination of the preceding main verb and agent already (e.g. "*Dann zerschnitt die Dame*"). However, this cannot always be entirely achieved, given that, especially in Condition C, the main verb itself often introduces some degree of implausibility.

The context paragraph is the same for each of the three conditions of an item. In addition, the context paragraph repeats the target and distractor words three or four times each to prime the target word's meaning 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 word should facilitate retrieval and thus no N400 effect should be observed across conditions when conducting an EEG study (Brouwer et al., 2012, 2017), which however is not implemented in this case. The final sentence varies across conditions only regarding the main verb which is what renders the sentence plausible (Condition A: "the lady *dismissed* the tourist"), medium plausible (Condition A: "the lady *welcomed* the tourist"), or implausible (Condition A: "the lady *dismissed* the tourist"). In the final sentence, each target word is followed by an additional clause ("[...] and then he went to the gate.") to capture spillover effects in RTs. In contrast to the materials used by Aurnhammer et al. (2023), the main verb in Condition B was changed for the current study in such a way that the ambiguity in Condition B was removed, while maintaining graded plausibility across conditions. For example, by changing "the lady *weighed* the tourist" to "the lady *welcomed* the tourist" the expectancy of the distractor word "suitcase" is lowered and the target word "tourist" expectancy is now higher instead while still being less plausible compared to the baseline Condition A. Hence, the stimuli used in the current study differ only concerning the main verb in Condition B from the stimuli created by Aurnhammer et al. (2023). Figure 4.1 shows an item in the three conditions in the current study compared to an item in the study of Aurnhammer et al. (2023).

If the three plausibility levels for the target word are successfully maintained, the subsequent self-paced reading study should, analogous to the results of Aurnhammer et al. (2023), result in a graded RT~~s~~ effect for the three Conditions ($A < B < C$), reflecting increased integration effort as plausibility decreases. To assess whether the items in Condition B have been successfully manipulated in terms of plausibility and expectancy of the target and distractor words, two norming studies are conducted before the self-paced reading study.

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

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

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

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 weighted 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: Item 1 in the current study compared to item 1 in the study by Aurnhammer et al. (2023). Both are transliterated from German. 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, distractor words are highlighted in boldface.

4.2 Language Model Architectures

The LMs used for computing surprisal in a pre-study of this thesis belong to the Transformer model family. As discussed in Chapter 2.2.1, Transformer models have outperformed traditional LM architectures in several NLP tasks and are increasingly being investigated for modeling human sentence processing in the field of psycholinguistics. Although Transformers are less cognitively plausible than RNNs, they have recently been shown to be better predictors of behavioural and neural measures (Merkx and Frank, 2021; Michaelov et al., 2021).

The Transformer architecture as introduced by Vaswani et al. (2017), consists of an encoder that processes an input sequence and generates a fixed-length vector representation, and a decoder that generates an output sequence token by token based only on the context vector generated by the encoder. However, the LMs used to calculate surprisal for the current study, a German GPT-2 version (Schweter, 2020) and LeoLM (Plüster, 2023), are both decoder-only architectures, which are based solely on the decoder component of the Transformer architecture. Both GPT-2 and LeoLM use a causal attention mechanism that allows each token in the generated sequence to attend only to the preceding and the current tokens, but not to future tokens. This is crucial for calculating surprisal because it ensures that predictions are made based on the available context up to the current word, mirroring human sequential language processing.

Although the LMs used to compute surprisal for the current study both belong to the family of Transformer-based LMs, they differ in terms of their size and training data. The first model 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], 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 LLM built on Llama-2, which is larger than GPT-2 in terms of the number of parameters and resources 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 predominantly English texts (Touvron et al., 2023). For the current study, the LeoLM version with 13 billion parameters was chosen. To improve their proficiency in the German language, LeoLMs are initialized 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

[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].

trained on two smaller datasets, comprised of Wikipedia [5] and news (Tagesschau) [6] articles, resulting in a training dataset of approximately 600 gigabytes.

As the overview of the parameters of both models in Table 4.1 indicates, LeoLM is a more powerful model in terms of its size and the resources required for training and inference. Since the models' general capabilities were not evaluated using a metric such as perplexity for the present study, their performance cannot be compared in this context. However, given the notable difference in size, LeoLM would very likely outperform (i.e. achieve lower perplexity) GPT-2 in a performance test.

	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 language model architectures used to compute Surprisal in a pre-study: GPT-2 and LeoLM.

One disadvantage that is worth noting relates to the Byte-Pair Encoding (BPE) (Sennrich et al., 2016) tokenisation model that is used in both GPT-2 and LeoLM. BPE is a data compression technique that was adapted by Sennrich et al. (2016) for word segmentation and works by iteratively merging the most frequent pairs of consecutive characters in the input text. It is a commonly used tokeniser in Transformer models because it reduces the model's vocabulary size while maintaining its expressive power and enables the model to generate accurate predictions for rare or Out-of-Vocabulary words. However, the use of LLMs employing techniques like BPE is not uncontroversial from a cognitive modeling perspective (Nair and Resnik, 2023). More specifically, the problem is that the surprisal values for orthographic words are calculated as the sum of the surprisal values of their subwords ($P(w) = P(sw_1) + ... + P(sw_n)$). However, since subword tokenisation by LLMs is based on the frequency of character combinations, it differs from morphological subword decomposition in human processing. BPE-derived units that occur either as single words or as part of compound words are assigned the same token id. Thus, the surprisal values of compound words include those of their constituent subwords, and they therefore receive higher surprisal estimates by default than parts of compound words that exist independently, even if the actual expectancy of the single word is higher. Although Nair and Resnik (2023) found no disadvantage in the aggregate ability to predict RTs using BPE tokenisation compared to morphological segmentation, the former should be used with more caution, as it

[5] https://dumps.wikimedia.org. [Accessed: 2024-04-16].

[6] https://huggingface.co/datasets/bjoernp/tagesschau-2018-2023. [Accessed: 2024-04-16].

is less psychologically plausible. To avoid this problem, some studies only included items that were not split by the tokeniser (Michaelov et al., 2023). However, this approach is not suitable 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, suggesting that a small proportion of the RT data consists of much longer RTs than a larger proportion of the data. To normalize their distributions, the RT data was log-transformed. Applying a logarithmic transformation compresses the scale of the data, especially for larger values, therefore reducing higher RTs to a greater extent than lower RTs.

The log-transformed RTs were then analysed with the same linear mixed effects regression re-estimation technique as 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 consisted 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 from Item 1, the critical regions are underlined: "Dann verabschiedete die Dame den (*Pre-critical*) Touristen (*Critical*) und (*Spillover*) danach (*Post-spillover*) ging er zum Gate").

The plausibility ratings collected and the surprisal values computed by the LMs in the pre-studies were used as numerical predictors in linear mixed effects regression models to investigate the extent to which they account for the variability in the observed RT data across all regions. 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 a linear mixed effects regression model based on plausibility and surprisal that are included as predictors. How closely these estimates reflect the observed RTs can be assessed through the residual or prediction error, which is calculated by subtracting the observed RTs from the predicted RTs ($y - \hat{y}$). The smaller the residuals, i.e. the more closely the predictions of the model match the actual RTs. More specifically, the predictors included in the model are target word plausibility and distractor word surprisal, which, as demonstrated in Table 5.2 of Chapter 5 are independent of each other. 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.

Before fitting the model, 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 corresponds to the mean of the outcome variable when all predictors are set to zero, facilitating the interpretation and comparison of the coefficients in the regression model. 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 less plausible items. By inverting the predictor, the coefficients for plausibility should be positive, indicating that as plausibility decreases, RTs increase.

Next, the data from both self-paced reading studies were re-estimated using a separate linear mixed effects regression model for each of the four critical regions, including variations of target word plausibility and distractor word surprisal. 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 between groups (e.g. subjects and items) and individual-level variability (Jaeger, 2008). Since plausibility and surprisal are the predictors of primary interest, they were included as fixed effects in the model. Random slopes and intercepts were added for subjects and items to account for individual differences in RTs that are not explained by the fixed effects predictors. For example, subjects may vary in terms of their reading speed or comprehension abilities and items may introduce varying degrees of difficulty depending on word frequency or length. Thus, the full model specification is

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

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 associated with the predictor variables 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-effects predictors or the random effects, capturing the difference between the observed and the predicted values.

First, the single-trial plausibility ratings collected during the self-paced reading study and GPT-2 surprisal were included as predictors in the models. The same models were then fitted with LeoLM surprisal followed by GPT-2 surprisal, allowing for a comparison of the RT predictions using surprisal estimates from different LMs. Next, the average plausibility ratings per item that were collected in a pre-study study were included as a predictor, first along with GPT-2 surprisal and subsequently along with LeoLM surprisal. This should reveal whether single-trial plausibility ratings collected during the self-paced reading study or average plausibility ratings collected in the pre-study predict the RT data more accurately. The RT data of the second self-paced reading study was only estimated using the average plausibility ratings from the pre-study together with GPT-2 and then LeoLM surprisal, as no single-trial plausibility ratings were collected during this study. Coefficients, z-values and p-values from all sets of models were reported. Since separate analyses were run for the four critical regions, treating each of them as a distinct set of hypotheses, it was not necessary to correct the p-values to account for multiple comparisons.

Finally, a likelihood ratio test (LRT) was performed to objectively compare the goodness-of-fit of a simple model, containing only the plausibility predictor which explains more variance in the outcome variable (together with distractor word surprisal), and a complex model, containing both single-trial and average pre-test plausibility ratings (as well as 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 equals the number of additional parameters in the complex model (compared to the simple 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 will be rejected, suggesting that the predictor that is included in the complex, but not the simple model, provides additional information that improves the model's ability to explain the observed data.

Chapter 5

Pre-studies

Before the main experiment, two norming studies were carried out to test whether the plausibility and expectancy manipulation of the final sentence in Condition B was successful. First, a plausibility rating study was conducted to ensure that plausibility is graded across conditions ($A > B > C$). Secondly, surprisal values were computed using the language models GPT-2 and LeoLM to assess whether the expectancy of the distractor word in Condition B was lowered. The plausibility norming study, as well as the subsequent self-paced reading studies, 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 their plausibility. Participants were asked to rate the plausibility of the final sentence of each item in the context of its preceding context paragraph on a seven-point Likert scale, with 7 indicating the highest plausibility and 1 indicating the lowest plausibility. Plausibility ratings were collected for both target and distractor words in all three conditions to assess if the main verbs in Condition B were chosen in a way that results in medium plausibility for Condition B, while plausibility for Condition A should still be high and plausibility for Condition C still low. Consequently, each of the items 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 per item (360 variations in total). In case the manipulation was successful, Condition A should receive higher plausibility ratings on average compared to Condition B and especially compared to Condition C. Given that more plausible items are usually also more expected, the same pattern may be observed in the distractor condition.

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 gave their consent to participate in the study by agreeing to a consent form. 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. To ensure that each participant read each item in only one condition

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

and that all conditions received an equal number of ratings from each participant, the remaining 60 participants were assigned to six different lists. Participants in each list read the same items, which differed from those in the other lists. Since each participant rated a total of 60 items, ten in each of the six conditions, 3600 ratings were collected in total. In addition to the 60 critical items, each participant was presented with 12 filler items, which, similar to the critical items, were plausible, medium plausible or implausible. The purpose of the filler items was to make sure that participants read the texts carefully. Therefore, the filler items contained instructions in the middle of the context paragraph asking 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. Before starting the actual experiments, participants were presented with three practice items. Consequently, each participant encountered a total of 75 items, comprising critical items, filler items, and practice items.

Participants who met the criteria of being a native German speaker, aged between 18 and 32 years, and without any language-related disorders or literacy difficulties were directed to the PCIbex platform for the experiment. Upon giving consent to participate in the experiment, participants read the instructions and examples of plausible, medium plausible, and implausible items. The instructions asked participants to rate the plausibility of the last sentence in the context of the previous paragraph on a scale of 1 ("implausible") to 7 ("plausible") and, when asked, with a specific number (1 or 7), regardless of its actual plausibility. Subsequently, participants were presented with three practice tasks to familiarize themselves with the task before the start of the experiment. During the study, the context paragraph was presented along with the final sentence and the seven-point Likert scale. In contrast to the plausibility rating study from Aurnhammer et al. (2023), the continuation of the final sentence ("and after that he left the store") was not excluded in the current study, since the final sentence continuation has to be included in the case of the single-trial plausibility ratings collected during the self-paced reading study to capture spillover effects in RTs. This is crucial for consistency reason as the post-target material can alter the plausibility ratings.

5.1.2 Results

On average, participants rated 99% (mean = 99.16%, SD = 2.52, range = 91.66%-100%) of the items that contained attention checks correctly, i.e., with the number 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.

Based on the average plausibility ratings, it appears that target word plausibility is graded $(A > B > C)$, indicating that the main verb in the final sentence of Condition B was chosen such that participants rated Condition B as medium plausible on average. Additionally, the results for Conditions A and C are similar to those of Aurnhammer et al. (2023), with Condition A being rated as plausible and Condition C as implausible on average. The larger standard deviation and range in Condition B, in contrast to Conditions A and C, indicate greater variability in participants' plausibility ratings, suggesting that participants also tended to assign high or low plausibility ratings to items in Condition B. This is not unexpected given the challenge of rating items of medium plausibility on a scale of 1 to 7, compared to rating items that are either obviously plausible or implausible.

Regarding the distractor word, the average plausibility ratings are also graded ($A > B > C$). However, the differences between the average ratings per condition are small, as all of them fall in the lower (implausible) range of the scale. This doesn't seem surprising, as replacing the target word with the distractor word in the final sentence renders the item even in Conditions A and B less plausible ("Then **dismissed** (A)/**welcomed** (B) the lady the suitcase"). Even though 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 in Condition B, considering that less plausible stimuli are also 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 found for Condition C. This could be attributed 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 instance, the sentence "Then **signed** the woman the suitcase (distractor)" is somewhat more plausible than "Then **signed** the woman the tourist (target)", as 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 objective was to achieve a graded effect for target word plausibility. Moreover, both target and distractor words received similarly low plausibility ratings in Condition C on average.

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 density plot on the top left shows the distribution of plausibility ratings for the target word, averaged per item across participants. Conditions A and C show a unimodal distribution, peaking at plausibility levels 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 suggests that while there may be instances where participants assigned high plausibility ratings to an item in Condition C or low plausibility ratings to an item in Condition A, on average, 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 that correspond to a medium level of plausibility. However, the average ratings are more spread out across the entire spectrum in Condition B compared to Conditions A and C. Since even average ratings fall in the very upper (plausible) or lower (implausible) range, this suggests that there is less consensus among participants regarding the plausibility of items in Condition B. As mentioned previously, this is not surprising, since assessing nuances of medium plausibility is inherently more challenging than assessing clearly plausible or implausible items. On the top right, Figure 5.1 shows the distribution of plausibility ratings for the distractor word, averaged per item across participants. On average, plausibility ratings for Condition C are lower and more concentrated towards the implausible end of the scale compared to Conditions A and B. In contrast, ratings in Conditions A and B display a broader distribution across the spectrum, confirming that while most items in Conditions A and C were rated as implausible on average, there is greater variability among participants' ratings.

Based on the results of this plausibility study and those of Aurnhammer et al. (2023), the single-trial plausibility ratings that will be collected during a self-paced reading study will most likely follow the same pattern. In addition, the self-paced reading study will show whether the RTs for the target word are graded for 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 at eliminating the ambiguity in Condition B by reducing the expectancy of the distractor word. Since Conditions A and C are adopted from Aurnhammer et al. (2023) without any changes, the expectancy of the target words should also be higher than the expectancy of the distractor words in these conditions, even though a different metric was used in the current study to assess their expectancy. Aurnhammer et al. (2023) determined the expectancy of the target and distracter words based on cloze probabilities, a human-based operationalisation of expectancy. In the current study, surprisal, an LM-derived operationalisation of expectancy, was used to estimate the expectancy of target and distractor words across conditions. If the distractor word expectancy in Condition B was sucessfully 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 unchanged Conditions A and C), given that surprisal is inversely proportional to expectancy.

To calculate surprisal values for the target and distractor words across conditions, two different transformer-based LMs were used: a pre-trained German GPT-2 model (Schweter, 2020) and LeoLM (Plüster, 2023), a German Foundation LM built on Llama-2. The sentence materials used as input to the LMs are the same as those used for the plausibility rating study.[2] First, the stimuli were preprocessed using a regular expression, which 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 identifies non-alphanumeric characters as separate tokens instead of considering them as part of the previous or following word, which is necessary because both GPT-2 and LeoLM use the Byte-Pair Encoding (BPE) (Sennrich et al., 2016) tokenisation model.

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

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 original stimulus "*Dann verabschiedete die Dame den Touristen und danach ging er zum Gate*" was truncated after "*Dann verabschiedete die Dame den* ***Touristen***".

Subsequently, the models processed the input sequence and generated logits, which were then transformed into probabilities using the softmax function. Each probability represented the likelihood of the corresponding token being the next token in the sequence. Then surprisal values for the tokens were computed by applying the $-log_2$ function to the probability estimates, measuring how surprising the token is, given the context provided by 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. Similarly, the surprisal values of the combined tokens were summed to obtain a single surprisal value for each word.

5.2.2 Results

Table 5.1 shows the descriptive statistics of the surprisal values computed by GPT-2 and LeoLM. The average surprisal values computed with GPT-2 are higher for the distractor than for the target word across all conditions. 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 main verb in Condition B was effectively manipulated in a way that resulted in a higher average expectancy of the target word compared to the distractor word 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 using LeoLM. 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 4.2). However, this is not the case for the surprisal values computed with GPT-2.

Moreover, the average surprisal values of both models 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 compared to the target word in Conditions A and B, which matches 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, Conditions A and B are almost identical, with Condition A having a slightly higher average surprisal value than Condition B $(C > A > B)$. LeoLM's average surprisal values for the distractor word deviate even further from the target word pattern $(A > B > C)$. While the goal of achieving a lower expectancy for the distractor word compared to the target word was accomplished, it's important to note that the expectancy, operationalised as surprisal, is not equally low across conditions, which makes it difficult to separate plausibility from predictability effects on RTs, an issue that was discussed in Chapter 2.3.

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 calculated by GPT-2 across all conditions, except for the distractor word in Condition A.

		Plausibility			**Surprisal $_{\text{(GPT-2)}}$**			**Surprisal $_{\text{(LeoLM)}}$**		
	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-studies 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 discerned from the densities (and the dashed lines) based on the surprisal estimates computed by both LMs for the target word. The high density of surprisal values observed for Condition A in the lower (implausible) range is consistent with the summary statistics in Table 5.1, confirming that the LMs, particularly LeoLM, predict mostly low surprisal values, reflecting high expectancy. As the target words in Conditions B and C are generally less predictable compared to those in Condition A, the surprisal values exhibit greater variability. This is demonstrated by the relatively even density curves in Conditions B and C, which span a wider range of values. The variability of the surprisal values appears to be even higher in the case of the distractor word due to the additionally rather low expectancy of the distractor word compared to the target word. Thus, the gradation per condition is only recognisable by the dashed lines, representing the average surprisal value per condition. In the case of GPT-2 in particular, the density curves of the surprisal values across conditions overlap almost completely, while the surprisal values computed by LeoLM, differ primarily in the density of high surprisal values for Condition A compared to Conditions B and C.

Table 5.2 presents the 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 between target word plausibility and target word LeoLM surprisal (r = -0.51) and a weaker negative correlation between target word plausibility and target

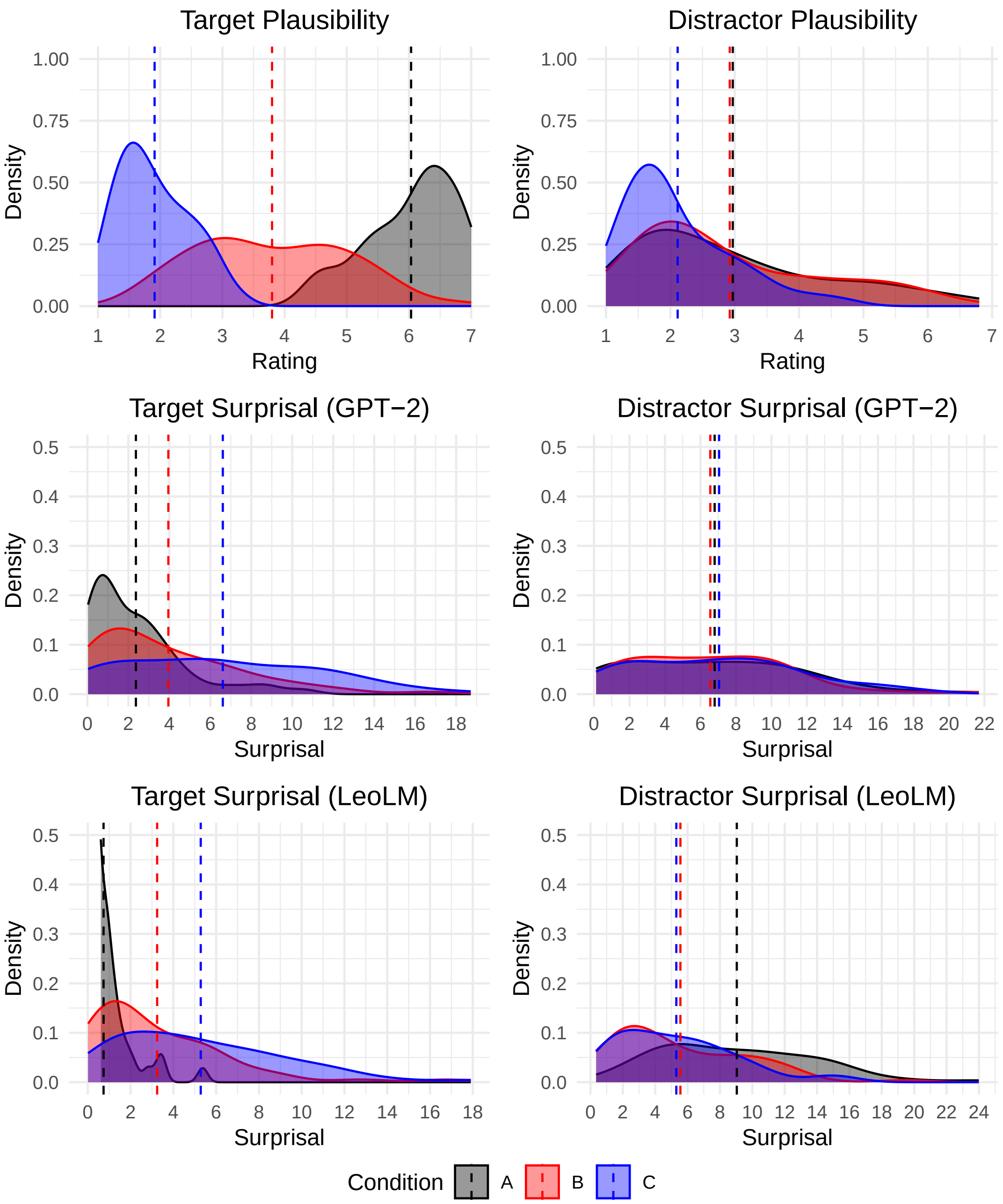

FIGURE 5.1: Densities for the results of the plausibility rating study that collected seven-point scale plausibility ratings and the surprisal values calculated by GPT-2 and LeoLM for the target and distractor words.

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. Moreover, 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), as both follow the pattern $A > B > C$. 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 reading time 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. Hence, 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, for example, in terms of multicollinearity.

		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 and (GPT-2 and LeoLM) Surprisal of the target and distractor words.

The observed correlation between plausibility and LeoLM surprisal is lower compared to the correlation between plausibility and expectancy operationalised as cloze probability that was observed by Aurnhammer et al. (2023). These differences could be attributed either to the human-derived nature of cloze probabilities, which might align more closely with human plausibility judgements compared to the LM-based surprisal estimates, the calculation of the surprisal values itself, which is not unproblematic as it relies on subword units created during tokenisation (Nair and Resnik, 2023) or be caused by other factors.

Ultimately, this may not impact the analyses of 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-study 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 semantically attractive alternative in Condition B has been removed, no significant RTs modulations 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 main 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 particularly 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 average plausibility ratings collected in a pre-study. The materials for the rating task were identical to those used in the plausibility pre-study. However, in the self-paced reading study, RTs were recorded only for the target word, not the distractor word, resulting in a total of 180 items being analyzed.

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). Following exclusion criteria based on inattentive reading, demonstrated by low response accuracy on comprehension questions (less than 80% correct), the data from three participants were excluded from the subsequent statistical analyses. 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 were not previously exposed 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 received compensation of £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 distinct lists and read different materials based on their assignment, which ensured that each participant read each of the 60 items only in 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, 60 of which were critical and 45 of which 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 180 item variations in total. Therefore, only three of the six lists contained unique items, whereas the remaining three lists included the same items arranged in a different order. Specifically, for half of the lists, the order of the blocks was reversed compared to the other half, and the items were randomized within each of the three blocks. Consequently, each item was read by precisely 14 different participants – seven from a forward-presented list and seven from a backward-presented list – in exactly one of the three conditions, which differed from the conditions of the items that were assigned to the other lists and read by the same number of different participants.

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 themselves with the task before the start of the main 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 centre of a blank page, indicating the position in which the words of the final sentence were subsequently presented. After that, participants read the final sentence word by word by pressing the *Space* bar 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. Following the last word of the final sentence, participants were presented with a seven-point Likert scale, 7 indicating a very plausible sentence and 1 indicating an implausible sentence, based on which they were prompted to rate the plausibility of the final sentence given the context paragraph they had seen before. This structure differs from the plausibility pre-study, where the seven-point Likert scale was presented on the same page as the context paragraph and the final sentence, allowing participants to re-read the entire item as many times as necessary before giving a rating. However, 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, requiring participants to more or less remember the content in order to give a rating. Even though participants could re-read 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. Since the participants' response accuracy on the comprehension questions is presumably reflective of their attention during reading, this served as a criterion to exclude the data of participants with too low overall response accuracy (below 80%) from all statistical analyses. Additionally, participants were encouraged to take a brief break between each of the three blocks.

6.3 Analysis

The items were analyzed 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 resulting answers to 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 subsections.

6.4.1 Comprehension Questions

All participants answered comprehension questions on approximately half of the experimental items (46% of all trials) and on two-fifths of all filler items. The descriptive statistics for response accuracy and reaction time on the comprehension questions were calculated across subjects. The average 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 compared to questions related to medium plausible and especially implausible items.

	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 Single-trial Plausibility Ratings

During the self-paced reading study, a plausibility rating was 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-study due to the different number of participants (60 in the pre-study and 42 in the self-paced reading study). Additionally, plausibility ratings were only collected for the target word during the self-paced reading study, resulting in 2520 trials (60 items read by 42 participants), whereas in the plausibility rating study conducted as a pre-test plausibility ratings for both target and distractor words were collected, resulting in 3600 trials (60 items read by 60 people), 1800 of which contained plausibility ratings for the target word. However, since the ratings from the plausibility pre-study study were averaged per item across participants, this results in only 180 average plausibility ratings, one per item and condition.

Another difference to the plausibility pre-study 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-study 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 since 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 displayed in Figure 6.1 (right). The plausibility ratings collected online during the self-paced reading study for the target word follow a graded pattern across conditions $(A > B > C)$, similar to the results from the plausibility pre-study. However, the differences between the average plausibility values per condition calculated from the single-trial plausibility ratings are smaller compared to those from the plausibility pre-study.

		Average 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 during a pre-study (left) and online during a self-paced reading study (right) on a seven-point scale for the target word.

The correlations between single-trial and average pre-test target word plausibility as well as 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 average pre-test target word plausibility (r = 0.72). The correlation between the average pre-test plausibility and the single-trial plausibility ratings averaged per item across subjects is even higher (r = 0.92), although it is not explicitly displayed in the table. This indicates

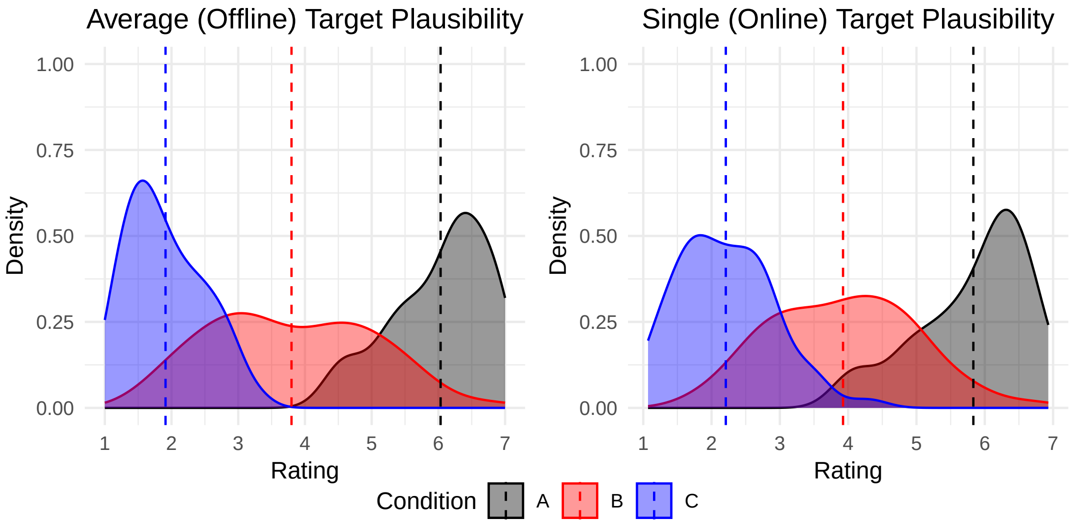

FIGURE 6.1: Densities for the results of two studies, that collected plausibility ratings offline during a pre-study (left) and online during a self-paced reading study (right) on a seven-point scale for the target word.

a strong relationship between the plausibility ratings collected in both studies. In terms of target word surprisal, the strongest (negative) correlation is observed between LeoLM surprisal and average target word plausibility, followed by GPT-2 surprisal and average 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 average target word plausibility (*r* = - 0.01) as well as 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 average target word plausibility (*r* = 0.28) are higher. Even though the latter can still be considered as rather weak, the correlation

		Plausibility (Tar)		**Surprisal (Tar)**		**Surprisal (Dist)**	
		Average	Single	GPT-2	LeoLM	GPT-2	LeoLM
Plausibility (Tar)	Average	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 average plausibility ratings collected in a pre-test, single-trial plausibility ratings collected online during a self-paced reading study for the target word and GPT-2 and LeoLM Surprisal for the target and distractor word.

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 average target word plausibility are positive, indicating that as plausibility increases, surprisal also increases. This is the case because the surprisal values calculated by LeoLM for the distractor word follow, perhaps unexpectedly, the same pattern as average and single-trial plausibility (A > B > C). Potential reasons for this are discussed in Chapter 5.2.2.

6.4.3 Reading Times

The observed log-transformed RTs per condition for 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.

The observed RT data indicates that on average the RTs in the Pre-critical region are generally lower compared to the RTs in other regions and are nearly identical across conditions. However, upon closer inspection, they already appear to follow the pattern $A < B < C$ in the Pre-critical region, corresponding to the three levels of plausibility associated with Conditions A, B and C. In the critical region, i.e. upon reading the target word, the RTs increase in all conditions, but to a greater extent in Conditions B and C compared to Condition A. In the subsequent Spillover and Post-spillover regions, the RTs diverge even further, as they continue to increase in Conditions B and C but decrease in Condition A. Since the log-transformation compresses RTs in higher value ranges more strongly, it may appear that the RTs for Condition C decrease in the Post-spillover region compared to the Spillover region, but the raw RTs (see Appendix B.3) show a continued increase in this region for Condition C as well. Generally, the average RTs are graded for plausibility in all regions: Items in Condition A, which correspond to a high level of plausibility were read the fastest on average, while items in Condition C, corresponding to a low level plausibility, were read the slowest on average. However, this gradation is not strongly pronounced, as the average RTs for Conditions B and C are very similar in all four critical regions, particularly in the Post-spillover region, and differ to a greater extent from the average RTs in Condition A. While the gradation is most evident in the Spillover region, the difference in RTs between Conditions B and C is less distinct compared to the difference in RTs between Conditions A and B in this region as well. This suggests that although a plausibility effect is present, it is possible 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 anticipated.

The average RTs per region are similar compared to the observed RTs in the self-paced reading study that was conducted by Aurnhammer et al. (2023) on the Pre-critical region. However, they don't increase as much on the Spillover and Post-spillover regions and only form a weak contrast between Conditions B and C.

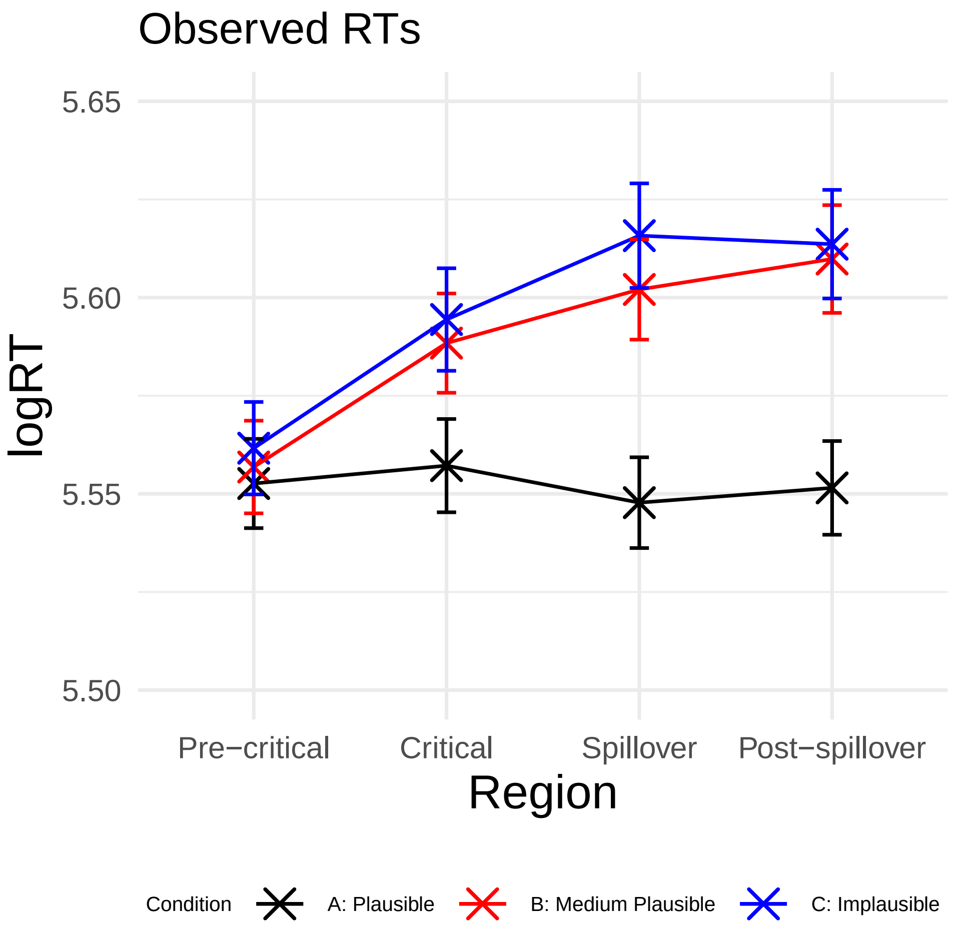

FIGURE 6.2: Log-Reading Times from the first self-paced reading study per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions. The error bars show the standard error calculated from the per-subject per-condition averages.

A linear mixed effects model in which target word plausibility and distractor word surprisal were used as predictors of RTs was fitted separately for each critical region to isolate the influence of each predictor variable on the RTs in each critical region. Both average target word plausibility collected in a pre-test as well as 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 displayed in Figure 6.3 and the corresponding residuals, i.e. the differences between the observed RT data and the RTs predicted by the models, are shown in Figure 6.4.

The estimated RTs and 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 used in each model and the conditions and regions analysed. The models that combined average target word plausibility with either GPT-2 distractor word surprisal or LeoLM distractor word surprisal appear to capture the effects structure in the

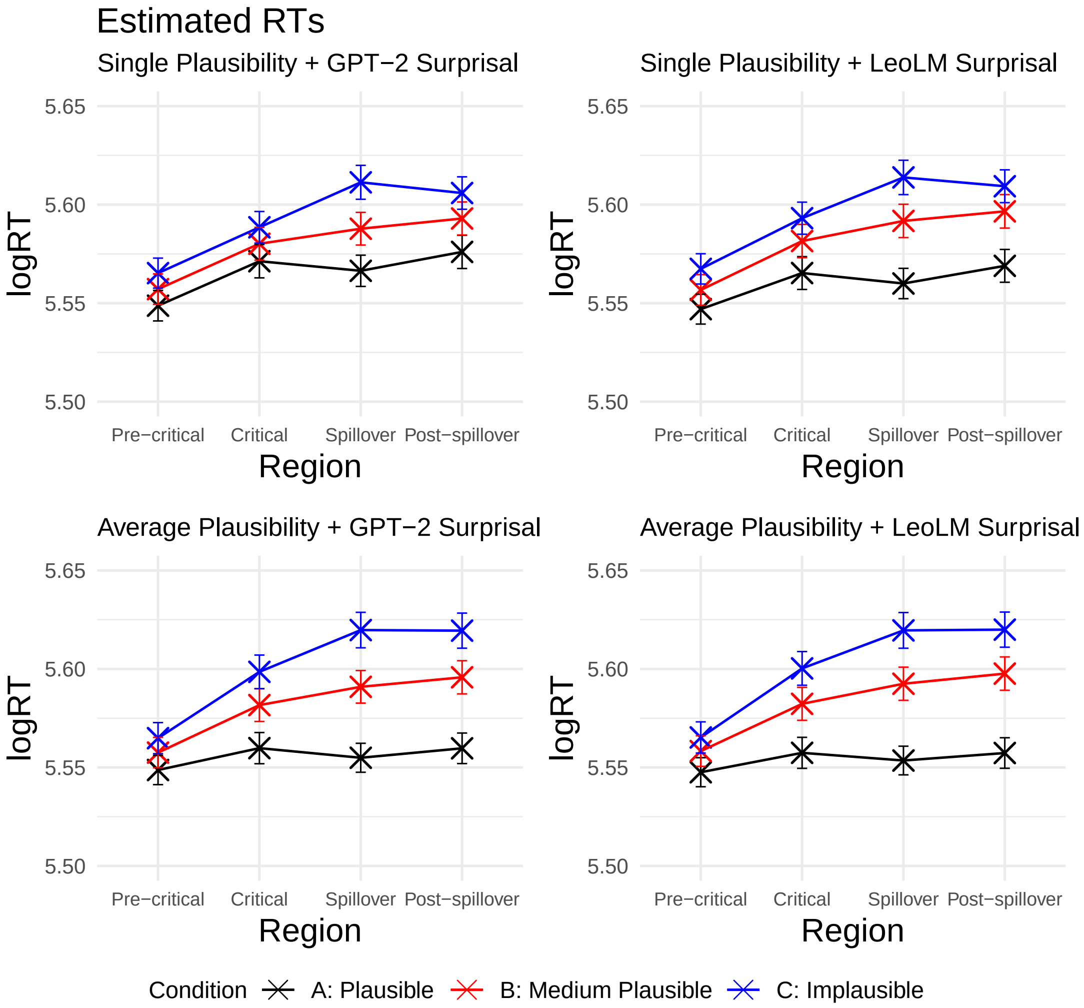

FIGURE 6.3: Estimated log-Reading Times using the predictors single-trial Plausibility and GPT-2 Surprisal (top left), single-trial Plausibility and LeoLM Surprisal (top right), average Plausibility and GPT-2 Surprisal (bottom left) and average Plausibility and LeoLM Surprisal (bottom right) per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

observed RT data better than the two models that used single-trial target word plausibility, as shown by the smaller residual errors across regions and conditions. The residual errors in the models fitted with average target word plausibility are particularly small for Conditions A and C, but larger for Condition B. Since both models underestimate the RTs in Condition B more strongly, the gradation of the estimated RTs appears more pronounced than it is the case in the observed RTs. Whether GPT-2 distractor word surprisal or LeoLM distractor word surprisal is used in combination with average target word plausibility does not seem have a substantial impact on the prediction accuracy. Inspection of the exact residual errors reveals that the models incorporating GPT-2 distractor word surprisal yield slightly more accurate predictions for RTs in Condition C, whereas the models employing LeoLM distractor word surprisal explain the RT data in Conditions A and B slightly

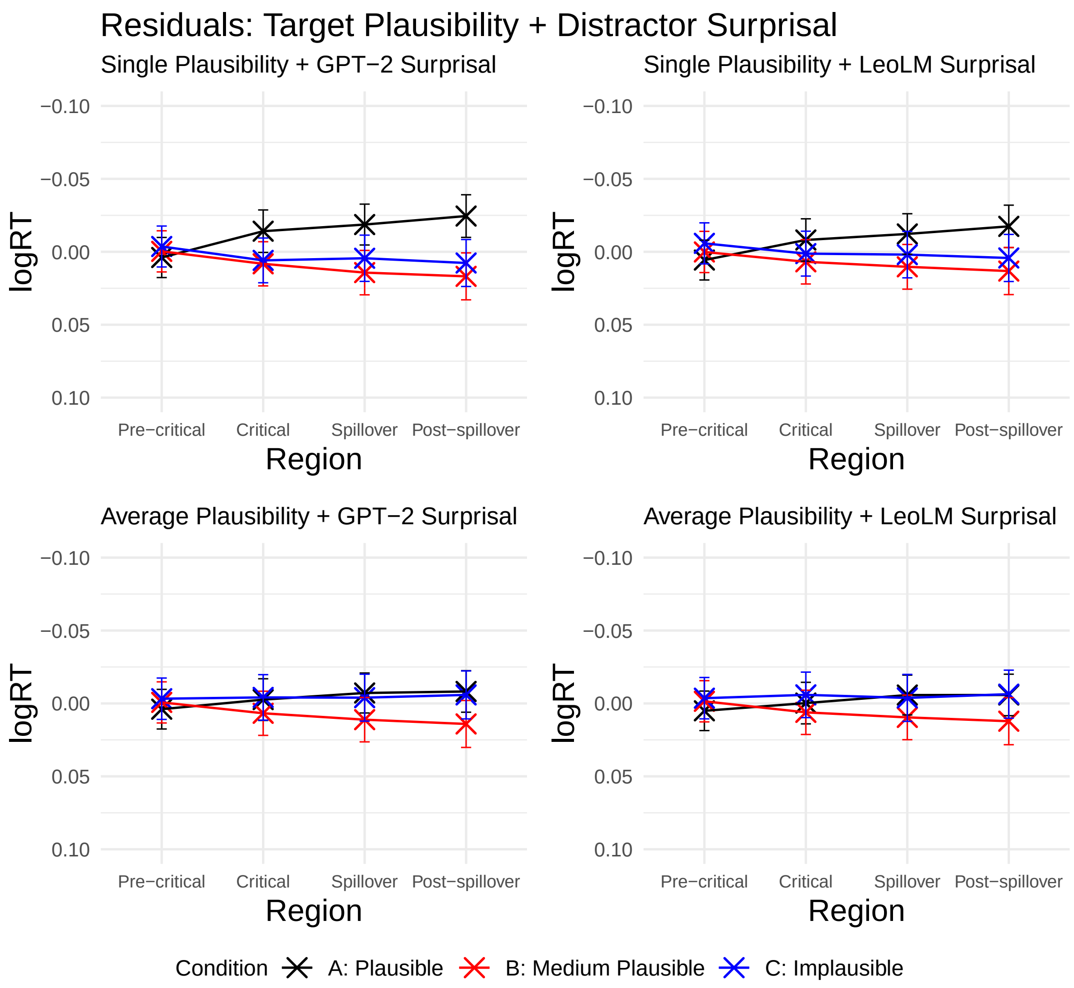

FIGURE 6.4: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

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 the larger residual error. Specifically, the model containing 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, however, the predictions for the RTs in Condition A improve, resulting in a reduced residual error.

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

Figure 6.5 shows the model coefficients, added to their intercept for single-trial and average target word plausibility as well as for GPT-2 and LeoLM distractor word surprisal. The coefficients for both single-trial plausibility and average plausibility are positive in all regions, indicating that lower plausibility predicts slower RTs. In contrast, the coefficients for GPT-2 surprisal and LeoLM surprisal are negative in all regions, suggesting that less surprising (more predictable) words are read slower. Negative surprisal coefficients might seem counterintuitive, as lower surprisal (higher predictability) would typically be expected to result in faster RTs. However, this result may 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 more plausible on average and plausibility and expectancy usually align) than items in Conditions B and C. Thus, both lower plausibility and lower surprisal (higher

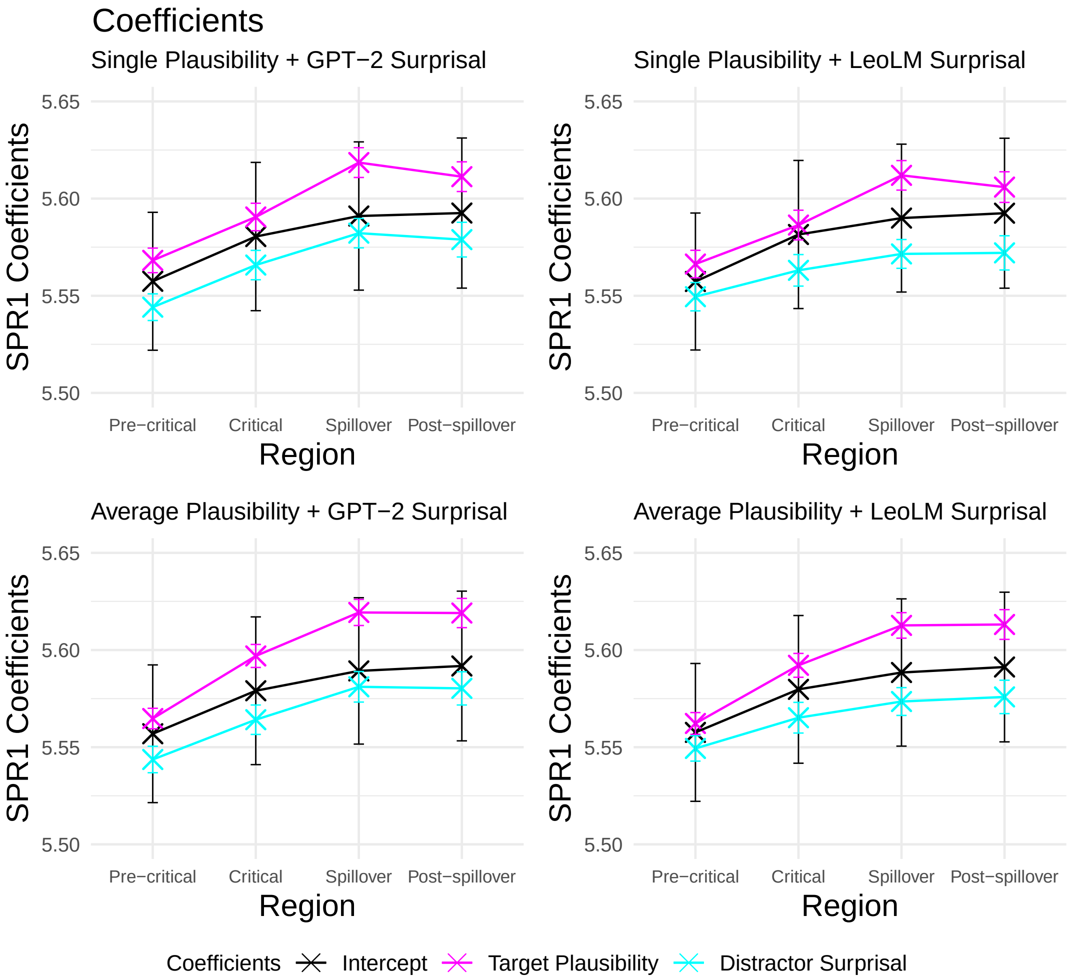

FIGURE 6.5: Coefficients, added to their intercept, from the models fitted with four different predictor combinations. Error bars indicate the standard error of the coefficients in the fitted statistical models.

predictability) simultaneously predict slower RTs. Figure 6.6 displays the z-values associated with each fixed effect coefficient and whether a predictor was significant on the respective region or not. The exact p-values are reported in Table 8.1.

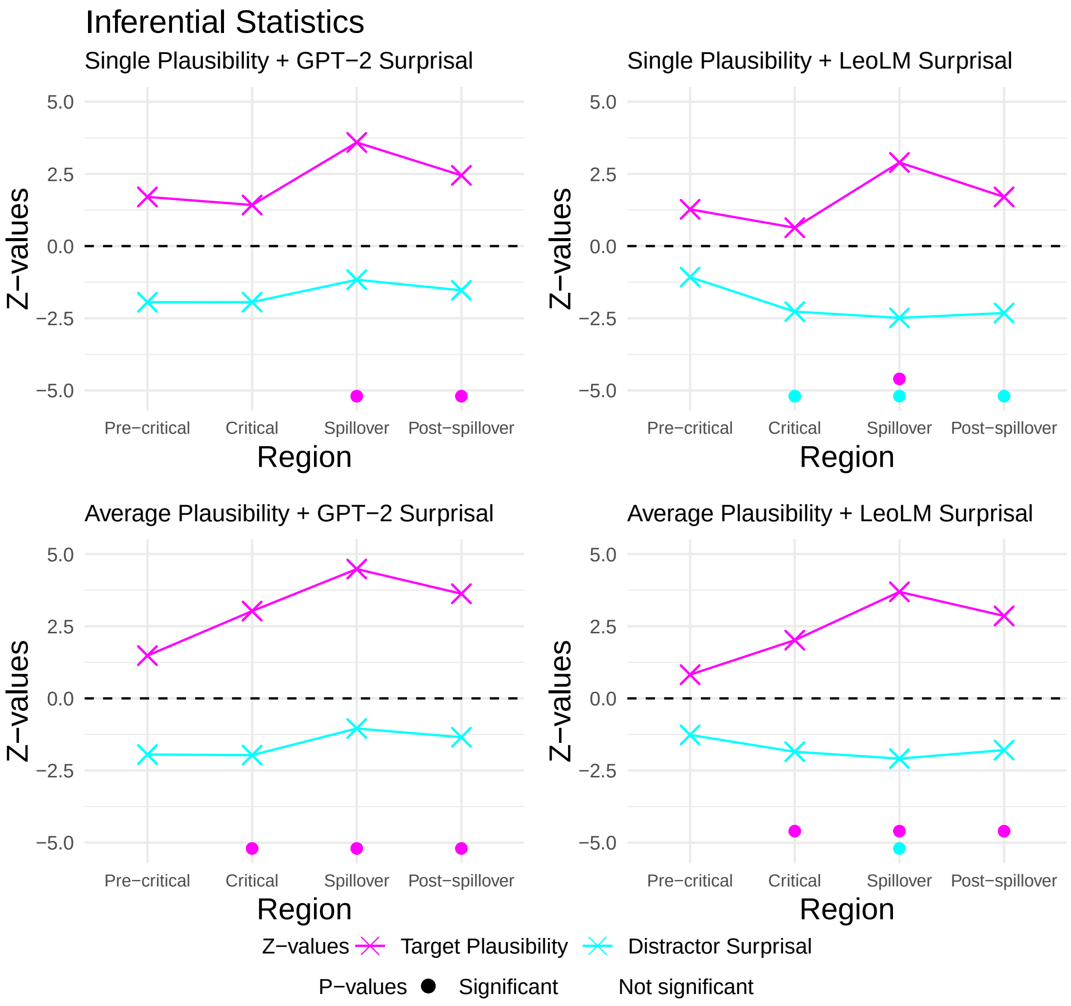

FIGURE 6.6: Effect sizes (z-values) and p-values from the models fitted with four different predictor combinations.

The z-values determine how many standard deviations the estimated coefficients deviate from zero, i.e. from the mean, serving as a measure of the statistical significance of the predictor variables' effects on the dependent variable (RTs). 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 is 0.05 [1] in the case of this study.

As the z-values and p-values associated with the coefficients for single-trial target word plausibility and GPT-2 distractor word surprisal indicate, target 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.

plausibility is a significant predictor of RTs at the Spillover and Post-spillover regions, while it is not significant on the Pre-critical and Critical regions. Distractor word surprisal does not significantly predict RTs on any region. In contrast, in the model fitted with single-trial target word plausibility and LeoLM distractor word surprisal, target word plausibility significantly predicts RT only on the Spillover region, while distractor word surprisal is significant on the Critical, Spillover and Post-spillover regions. In the two models in which average target word plausibility was included as a predictor, target word plausibility is a significant predictor of RTs on the Critical, Spillover and Post-spillover regions. Conversely, GPT-2 distractor word surprisal is again not significant on any region, whereas LeoLM distractor word surprisal significantly influences RTs on 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 8.1).

As shown in Figure 6.2, the observed RTs in Condition C are, on average, already higher than the RTs in Condition B and especially Condition A in the Pre-critical region. This is likely due to the processing of the preceding main verb, which may already introduce varying levels of plausibility depending on the condition before reaching the target word. To determine whether the RTs in Condition C differ on the Critical region (i.e. the target word) and on the following regions due to the plausibility of the target word itself or due to variations in the contexts leading up to it, the same models were re-fitted, including Pre-critical RT as a third predictor. By including Pre-critical RT in the analysis, the effects of the target word's plausibility and expectancy can be isolated from the influence of the preceding context, so that the plausibility and surprisal predictors now account for any additional RT variations beyond the differences in RTs observed in the Pre-critical region. Before fitting the models, the Pre-critical RT predictor was standardized but not log-transformed to ensure it remained distinct from the log-transformed dependent variable, which also includes the Pre-critical RT values. The resulting coefficients are displayed in Figure 6.7 and the z- and p-values in Figure 6.8.

As shown by the z- and p-values in Figure 6.8, Pre-critical RT is a significant predictor of RTs across all models including different predictor combinations, i.e. when single-trial or average 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, beyond what is explained by the Pre-critical RT predictor, while surprisal is still not significant on any region. When using single-trial plausibility and LeoLM surprisal as predictors, plausibility still significantly predicts RTs on the Spillover region, but no longer on the Post-spillover region, whereas surprisal is still significant on the Spillover and Post-spillover regions, but no longer on the Critical region. Regarding the two models including average target word plausibility, plausibility is still significant on 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 effect on RTs.

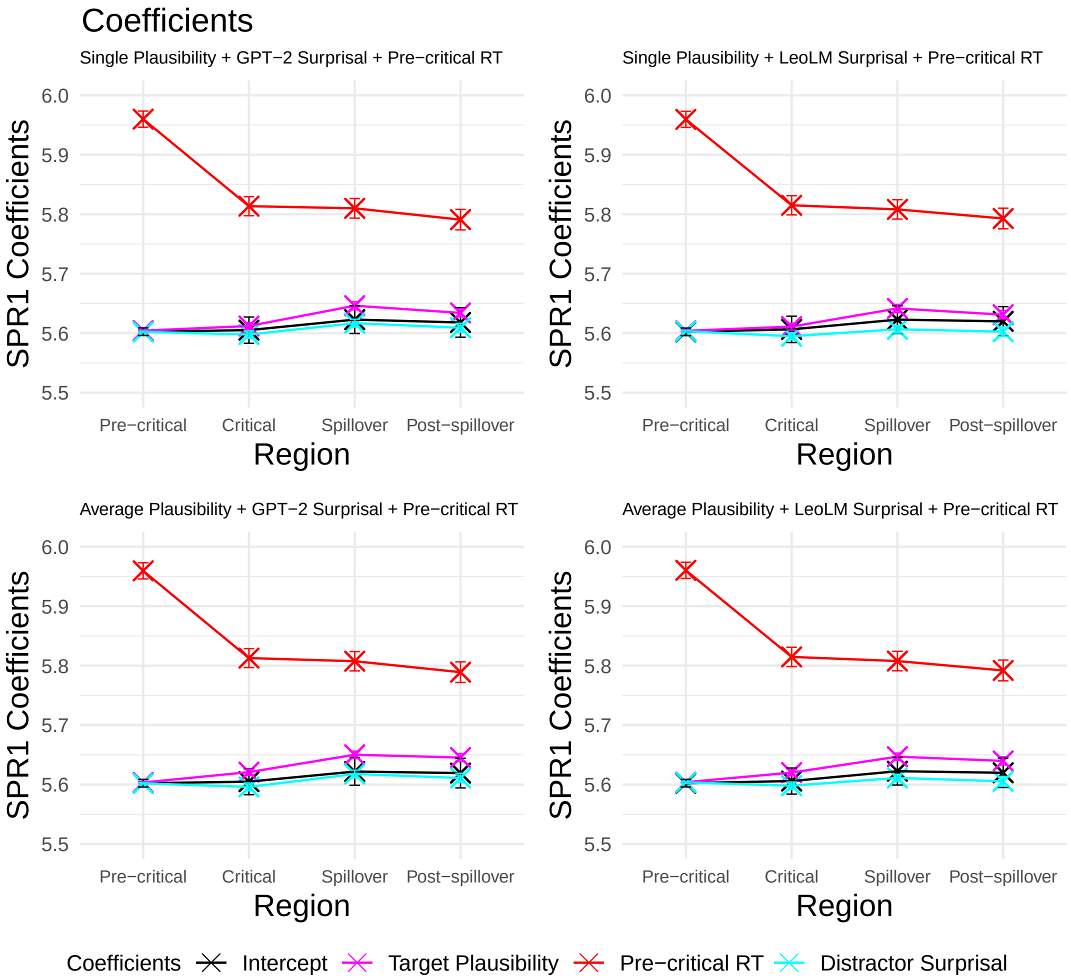

FIGURE 6.7: Coefficients, added to their intercept, from the models including Pre-critical Reading Time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

6.4.4 Model Comparison

Since the average target word plausibility ratings collected in the pre-study proved to be a better predictor of RTs, 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 average and single-trial target word plausibility, significantly improves model fit. To compare the goodness-of-fit between a complex model, which in this case includes average and single-trial target word plausibility, and a simpler model, including only average target word plausibility, a likelihood ratio test (LRT) was performed. The LRT determines whether the complex model (alternative hypothesis), significantly improves the fit to the data compared to the simpler 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 simpler model. If the test statistic exceeds the critical value from the chi-squared distribution

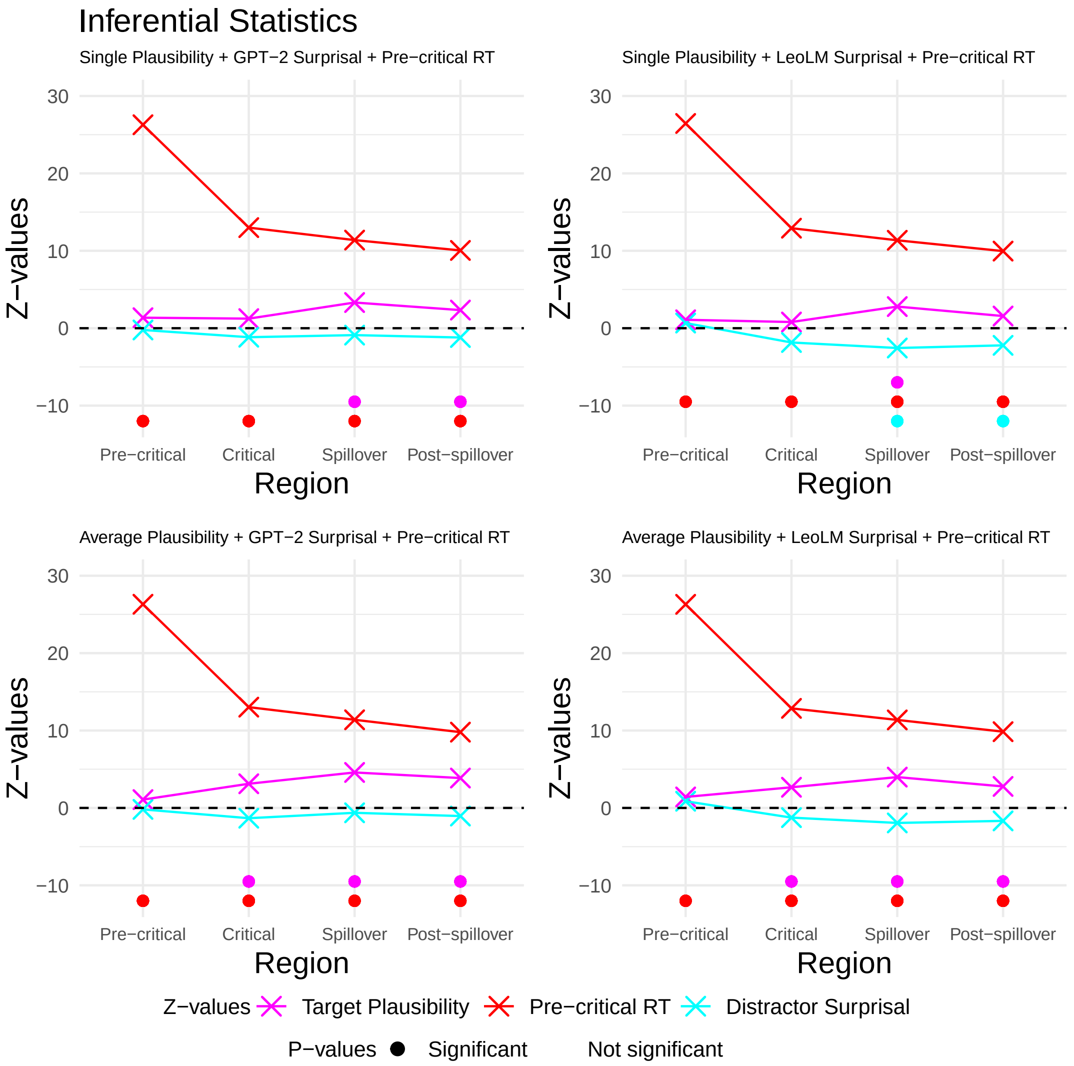

FIGURE 6.8: Effect sizes (z-values) and p-values including Pre-critical Reading Time as a predictor.

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 average and single-trial target word plausibility are mentioned in Table 6.4, GPT-2 distractor word surprisal was also used as a predictor in the simple and complex model. The LRT was also performed using LeoLM surprisal instead of GPT-2 surprisal; however, since the results did not significantly differ, only the results obtained with GPT-2 are reported in Table 6.4. 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 findings suggest that while single-trial plausibility ratings alone yield a less accurate estimate of RTs compared to average plausibility ratings, they capture additional variability that improves the overall model fit when

		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	0.000224
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	1.105e-05

TABLE 6.4: Results from the Likelihood-Ratio Test via ANOVA for model comparison.

combined with average plausibility ratings. The reason why the BIC value contrasts with other metrics – shown by the lower BIC value in all regions for the simpler model, suggesting a better fit to the RT data than the complex model – is not entirely clear. It may be due to do with 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.

Since the LRT indicated that including both single-trial and average target word plausibility as predictors significantly improves the model fit, a separate model including both plausibility predictors as well as GPT-2 surprisal was fitted for each critical region. The resulting estimated RTs and residuals are presented in Figure 6.9. The same models were also fitted using LeoLM surprisal instead of GPT-2 surprisal. However, 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 included. Moreover, the estimated RTs and residuals obtained from the complex model show no observable differences compared to the estimates and residuals obtained when predicting RTs based on average target word plausibility and GPT-2 or LeoLM distractor word surprisal alone (see Figures 6.3 (bottom left) and 6.4 (bottom right)). However, Figure 6.10 displays the coefficients added to their intercept (left) and z- and p-values (right), for the models including single-trial and average 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 surprisal or LeoLM surprisal as the third predictor. The coefficients for both GPT-2 and LeoLM distractor word surprisal are still negative across all regions, indicating that lower surprisal (higher predictability) predicts slower reading. Similarly, the model coefficients for average target word plausibility are still positive across all regions, suggesting that lower plausibility predicts slower RTs, except in the Pre-critical region of the model fitted with LeoLM 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 on the Pre-critical and on the Spillover region and negative on the Critical and the Post-spillover regions, indicating that it predicts slower or faster RTs depending on the region of interest.

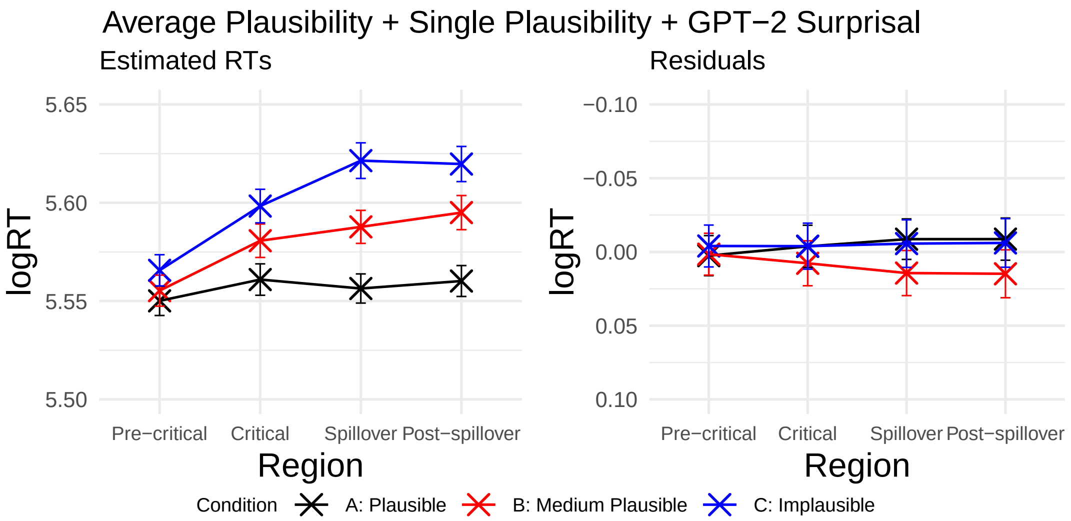

FIGURE 6.9: Estimated log-Reading Times (left) and residuals (right) obtained from the models including average (offline) Plausibility, single-trial (online) Plausibility and GPT-2 Surprisal as predictors per condition on the four critical regions.

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

6.5 Discussion

The results of the pre-studies (see Chapters 5.1.2 and 5.2.2) showed that the expectancy of the target word was lowered in Condition B, resulting in a lower average expectancy of the target word compared to the distractor word in all conditions [2], while the three levels of plausibility $(A > B > C)$ were maintained. The results of the self-paced reading study further demonstrated that RTs gradually increased with decreasing plausibility levels, such that items were read slower on average the less plausible they were (see Figure 6.2). However, in contrast to the observed RTs in the study of Aurnhammer et al. (2023), which formed a consistent pattern with Conditions A, B and C, the observed RTs collected in the current

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

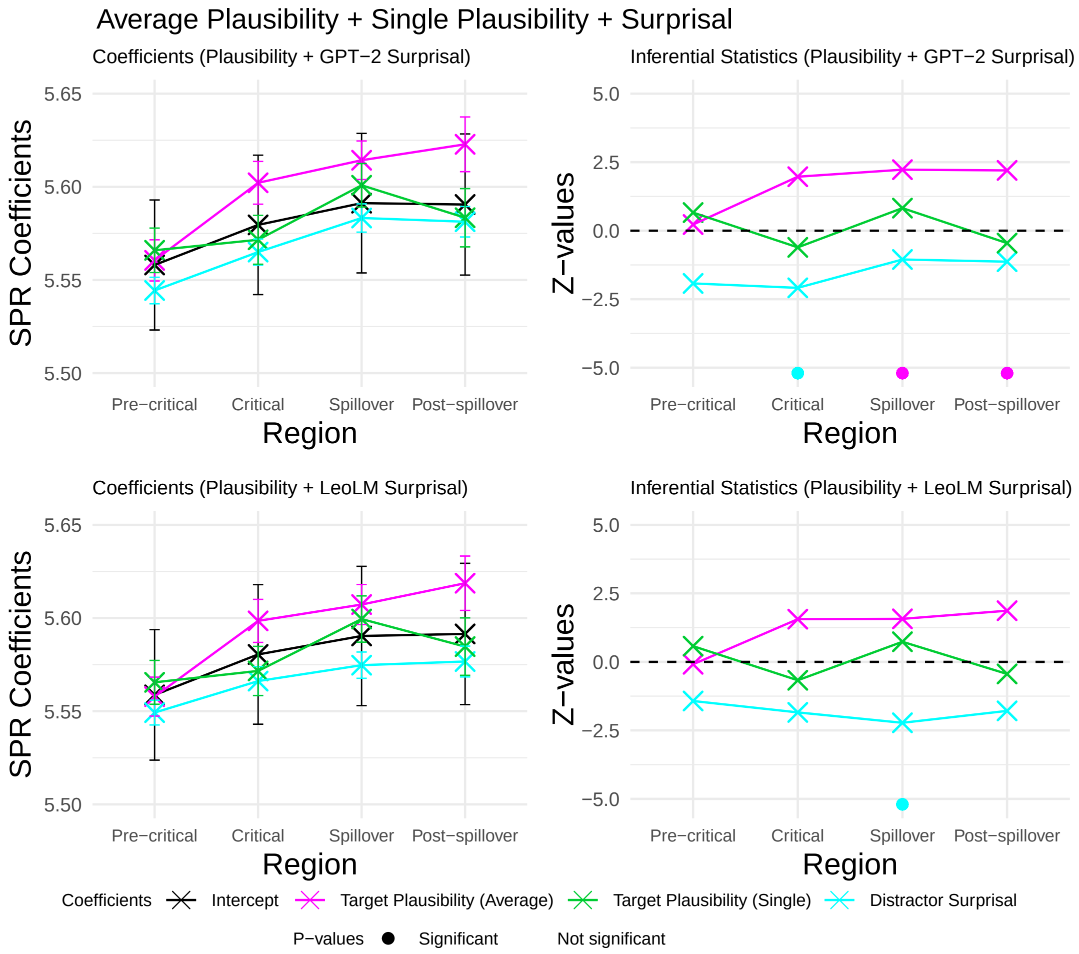

FIGURE 6.10: Coefficients, added to their intercept, (left) and z- and p-values (right) obtained from the models including average (offline) Plausibility, single-trial (online) Plausibility and GPT-2 Surprisal (top) or LeoLM Surprisal (bottom) as predictors per condition on the four critical regions.

self-paced reading study did not increase as strongly after the Critical region, i.e., after reading the target word, and formed only a weak contrast between Conditions B and C. A possible explanation for the similarity between the RTs in Conditions B and C could be that the item manipulation led to lower plausibility in Condition B, resulting in relatively high RTs. However, the results of the plausibility pre-study and the online plausibility rating task indicated that plausibility is evenly graded across 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. It may therefore have been the combination of the self-paced reading study and the 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 identify other potential factors that might have contributed to the observed RT pattern in the self-paced reading study, several post hoc analyses were carried out, which are described in Appendix B.3 - B.6.

The linear mixed effects regression analysis demonstrated that both single-trial plausibility ratings and average plausibility ratings from the pre-study continuously predict RTs. However, single-trial plausibility was a significant predictor of RTs either only in the Spillover region or in both the Spillover and Post-spillover regions, depending on whether GPT-2 or LeoLM surprisal was used as the second predictor, whereas average 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 average plausibility ratings from the pre-study 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 is reasonable to assume that single-trial ratings would be a better predictor of RTs due to their ability to capture individual differences and variations, allowing for a more granular analysis. One possible reason could be that precisely because of this granularity single-trial ratings might be disproportionately affected by outliers, which can distort the overall analysis. Even after the removal of outliers based on extremely low or high reading or reaction times, the single-trial plausibility ratings can fluctuate from trial to trial 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 average 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 fluctuations. This robustness to variability may be what ultimately makes them more effective predictors of RTs.

Further analyses (see Appendix B.1), revealed that the single-trial plausibility ratings averaged per item predicted the RTs better than the single-trial plausibility ratings but worse than the average plausibility ratings collected in a pre-study. Surprisingly, this shows that the average plausibility ratings from the pre-study provide a better fit to the RT data than the plausibility ratings collected online beyond the higher predictive power of the average itself, despite the fact that the former were collected from a different set of participants than the RTs. The reason for this might be that the combination of self-paced reading and the plausibility rating task may have not just distorted the RTs, but also the single-trial plausibility ratings, and therefore the average of these ratings, albeit to a lesser extent due to its relative stability. An important difference between the online and offline plausibility rating task was that during 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 average plausibility ratings per condition from the online study were more similar to each other (see Table 6.2) than those of the plausibility pre-study, may be that participants could provide ratings only based on the content they remembered, which made them more hesitant to select extreme ratings (very plausible or very implausible) during 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 make an observable difference when combined with the more effective average target word plausibility predictor. Moreover, the results showed no significant effects of GPT-2 distractor word surprisal on RTs when combined with either single-trial or average target word plausibility. This was anticipated, as GPT-2 surprisal was on average higher, i.e. expectancy 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 findings of Aurnhammer et al. (2023) did not reveal a significant effect of distractor word cloze probability on RTs, even though distractor word expectancy was higher than target word expectancy in Condition B, suggesting that RTs may not be affected by the presence of unfulfilled expectations.

Interestingly, LeoLM distractor word surprisal significantly predicted RTs in the Spillover region when used together with average target word plausibility and was additionally significant in the Critical and Post-spillover regions when used alongside single-trial target word plausibility to predict RTs. After including Pre-critical RT as a predictor to account for pre-critical RT differences, LeoLM surprisal was no longer significant when used together with average target word surprisal, however, it 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 may be related to the observed pattern for the LeoLM distractor word surprisal (see Section 5.2.2), which is the same as the pattern observed for target word plausibility ($A > B > C$), indicating that distractor word surprisal is simultaneously more plausible and less expected. This results in a higher correlation with single-trial target word plausibility (r = 0.16) and especially with average target word plausibility (r = 0.28) than between GPT-2 surprisal and single-trial and average target word plausibility, in the case of which it is almost identical to zero (see Table 6.3). As mentioned previously, this correlation should ideally be closer to zero. However, given that it is not excessively high there should not be any multicollinearity issue.

Finally, the results of a LRT showed that the complex model, which included both single-trial and average target word plausibility (along with distractor word surprisal), provided a better fit to the RT data than the simple model, which included only average target word plausibility from the pre-test and distractor word surprisal, in all regions except the Spillover region (see Table 6.4). Surprisingly, the complex model did not capture the RT structure across conditions more accurately, as indicated by slightly larger residual errors (see Figure 6.9). The difference in predictions accuracy is small and cannot be detected through visual inspection, but examination of the exact 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 indicates that although the single-trial plausibility ratings account for variation in RTs beyond the variation explained by the average pre-test plausibility ratings, this improvement is not specific to the RTs of Conditions A, B, C. 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 captures the RTs grouped by Conditions A, B, and C better.

Chapter 7

Self-Paced Reading Study II

Additional analyses (see Appendix B.5) investigating the RT data from the first self-paced reading study revealed that when grouping RTs based on the assigned plausibility ratings rather than based on Conditions A, B and C into three equally sized groups, items with a medium level of plausibility (i.e. assigned a plausibility rating of 3, 4 or 5 on a seven-point Likert scale) exhibited on average higher RTs compared to items rated as very plausible (assigned a plausibility rating of 5, 6 or 7) or implausible (assigned a plausibility rating of 1, 2 or 3) on average. A potential explanation is that items perceived as medium plausible are inherently more challenging to rate than highly plausible or implausible items and since participants anticipate the upcoming rating task, the increased processing difficulty of medium plausible items may be reflected in increased RTs on the final sentence before reaching the rating task itself. 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 combining the online plausibility rating task with the self-paced reading study may have distorted the RTs. More precisely, RTs observed for items in Condition B might have been higher and the RT pattern $A < B < C$ less pronounced than expected because Condition B mostly contains items of medium plausibility, which are more challenging 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, in which the plausibility rating task was removed, was conducted. Otherwise, the materials were identical to those used in the first self-paced reading study.

7.1 Participants

Similar to the first self-paced reading study, participants were recruited via Prolific Academic Ltd. for a second self-paced reading study, conducted on the PCIbex platform. A total of forty-five participants were initially 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 compared to the threshold in the first self-paced reading study (less than 80% correct). Consequently, in the current self-paced reading study data from five participants, who answered less than 80% (but more than 70%) of the questions correctly, were included. 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 not been previously exposed to the study materials, participants who participated in the first self-paced reading study, the plausibility rating pre-study 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 exactly 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 assess the plausibility of the word-by-word presented final sentence given the preceding context paragraph. Instead, they moved directly to the comprehension question, in case there was any, or to the next item in case there wasn't a question.

7.3 Analysis

The items were analysed based on the same linear mixed effects regression re-estimation technique (see also Aurnhammer et al., 2021, 2023) that was applied in Chapter 6. The procedure and all further 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 (in case 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 resulting answers to 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 nearly half of the experimental items (46% of all trials) and on two fifth 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 the 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 then 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). Thus, in the current study the mean accuracy in Condition C is lower and the mean reaction time higher than in Conditions A and B, indicating that the items in Condition C were more difficult to process than the items in Condition A or B. The mean response accuracies across conditions were slightly lower in the current study than in the first study, likely 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 also 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 presented in Figure 7.1. Already on 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 increase more strongly compared to 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 show only a slight increase and RTs in Condition A decrease again. Thus, RTs pattern with the three levels of Conditions A, B and C, reflecting high, medium and low levels of plausibility, on the Spillover and even more so on the Post-spillover region. The RTs observed in the current study reveal a more pronounced gradation of RTs for plausibility compared to the RTs 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 since 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 across 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 the observed RTs in the first self-paced reading study. They are also lower compared to the observed RTs in the self-paced reading study conducted by Aurnhammer et al. (2023), but otherwise show a similar pattern.

Figure 7.1 displays the estimated RTs using average target word plausibility and either GPT-2 distractor word surprisal (left) or LeoLM distractor word surprisal (right). Figure 7.2 illustrates the corresponding residuals, representing the differences between the observed RTs and those predicted by the linear mixed effects models. Visual inspection shows no differences between the estimates and residuals of the models in which GPT-2 distractor word surprisal was included compared to those fitted with LeoLM distractor word surprisal. In both cases, the residual error is low across all regions and conditions, especially in Conditions A and C, indicating

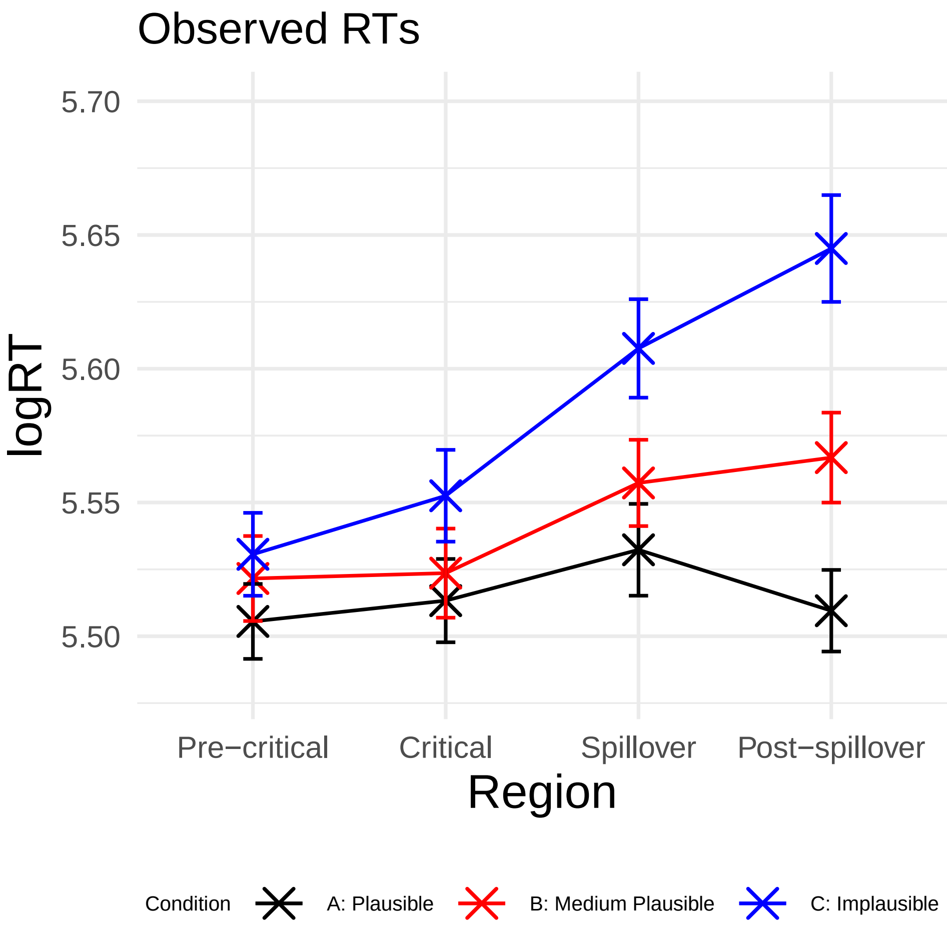

FIGURE 7.1: Log-Reading Times from the second self-paced reading study per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions. The error bars show the standard error calculated from the per-subject per-condition averages.

that the models accurately capture the observed RT data's structure. Similar to the first self-paced reading study, the model estimates in Condition B are slightly less accurate, evidenced by a higher residual error. However, while the first study showed a positive residual error (indicating overestimated RTs), the residual error in the second self-paced reading study is negative (indicating underestimated RTs) in Condition B.

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

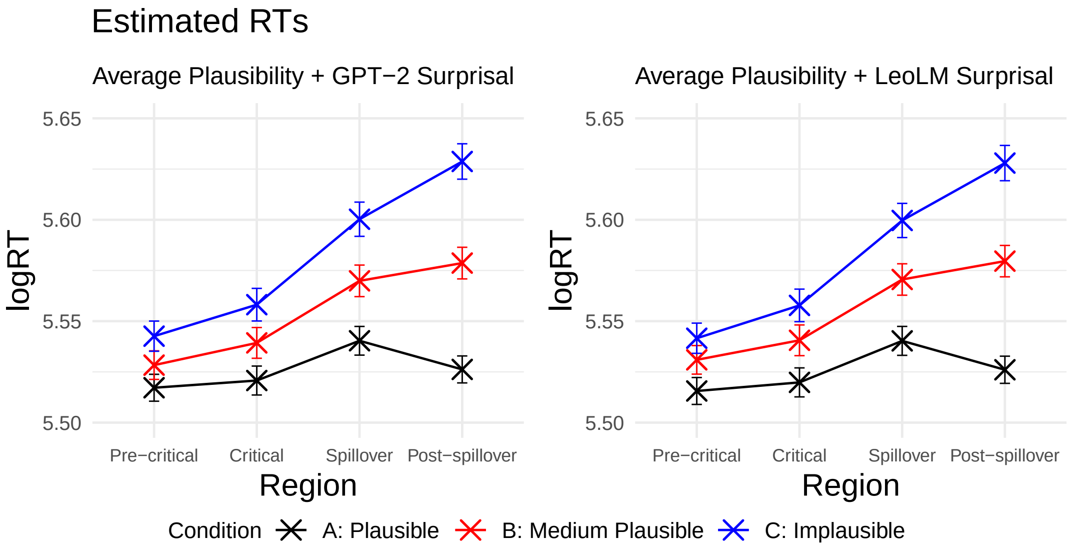

FIGURE 7.2: Estimated log-Reading Times using the predictors average Plausibility and GPT-2 Surprisal (left) and average Plausibility and LeoLM Surprisal (right) per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

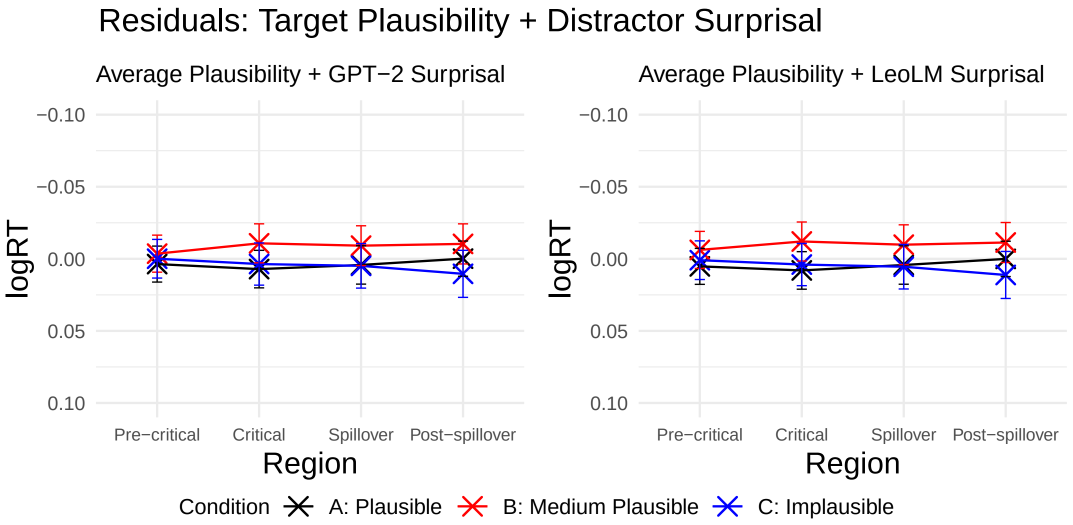

FIGURE 7.3: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

The relatively large z-values and small p-values (< 0.05) associated with target word plausibility indicate that plausibility is more significant in both cases, the model including GPT-2 distractor word suprisal as a second predictor and the model including LeoLM distractor word surprisal instead. In the first case, target word plausibility is statistically significant on 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 on any of the four regions, as indicated by small z-values and large p-values (> 0.05). The z-values obtained for average target word plausibility in the

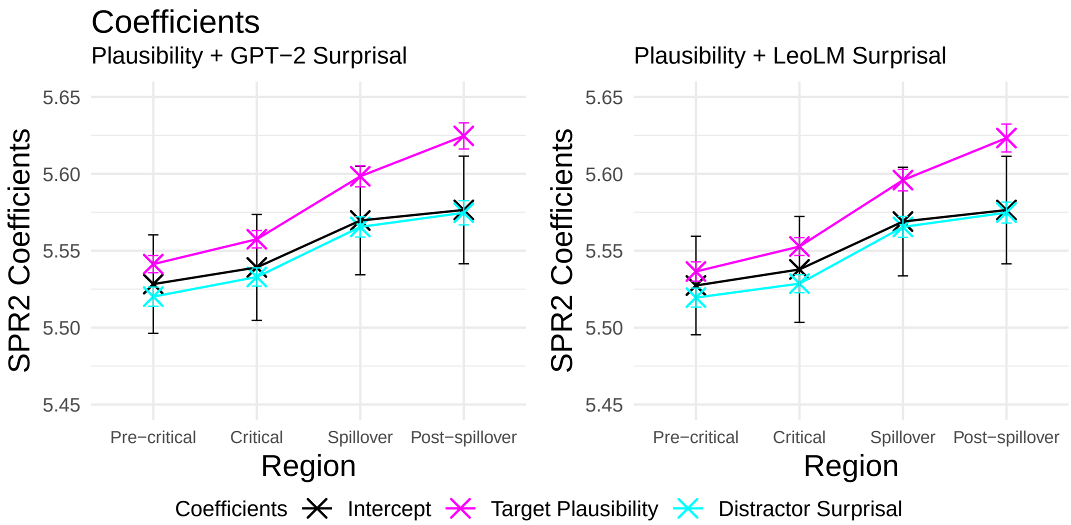

FIGURE 7.4: Coefficients, added to their intercept, for average Plausibility and GPT-2 Surprisal (left) and average Plausibility and LeoLM Surprisal (right). Error bars indicate the standard error of the coefficients in the fitted statistical models.

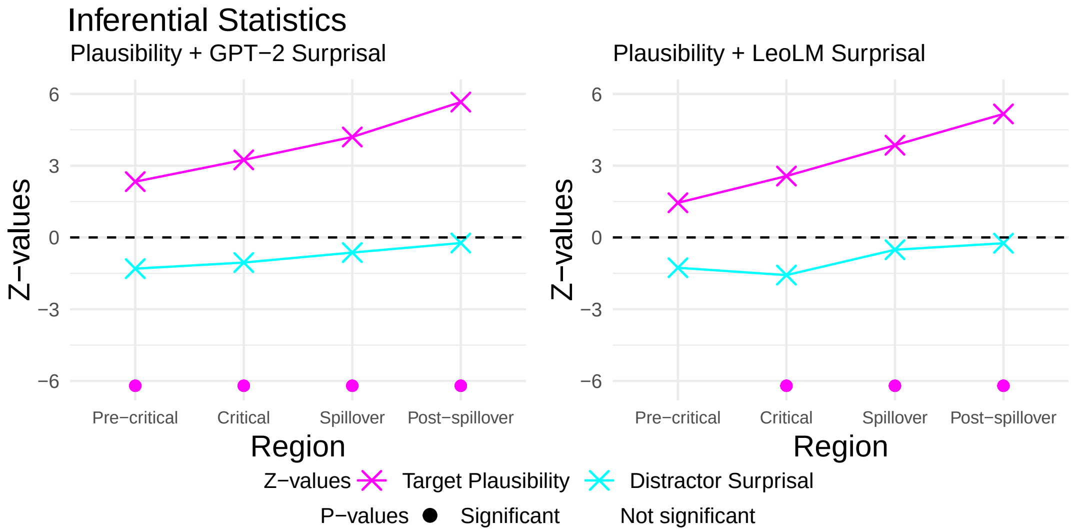

FIGURE 7.5: Effect sizes (z-values) and p-values.

second self-paced reading study are larger than those obtained in the first study across all regions (see Figure 6.6), which suggests 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 of 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 p-values (see Tables 8.1 and B.2) moreover reveals that plausibility has a more significant effect on RTs across all regions, whereas surprisal has a less significant effect on RTs in the second self-paced reading study compared to the first study. While average target word plausibility is a significant predictor of RTs on the Critical, Spillover and Post-spillover regions in the first self-paced reading study, average target word plausibility is significant on 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 either in the first or in the second self-paced reading study, while LeoLM distractor word surprisal significantly predicted 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 average target word plausibility, but it is not significant on any region in the second self-paced reading study

Reading time differences across conditions in the Pre-critical region were even more pronounced than in the first self-paced reading study, with Condition C being read slower on average than Conditions B and A. Therefore, 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 average 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 exact corresponding p-values are reported in Table B.2. The resulting z- and p-values indicate that RTs are significantly predicted by Pre-critical RT across all regions when using target word plausibility and GPT-2 distractor surprisal as well as when using LeoLM distractor surprisal instead. When plausibility and GPT-2 surprisal are used as predictors, plausibility remains a significant predictor of RTs across all regions except in the Pre-critical region. Conversely, when plausibility and LeoLM surprisal are included, plausibility is still significant on the Spillover and Post-spillover regions but no longer on the Critical region. GPT-2 and LeoLM surprisal still have no significant effect on RTs in any region.

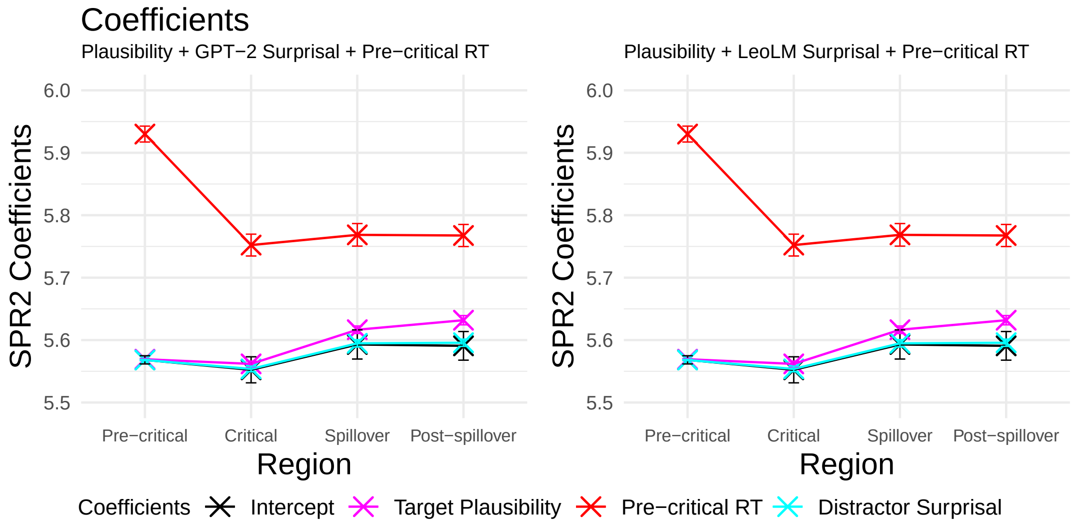

FIGURE 7.6: Coefficients, added to their intercept, from the models including Pre-critical Reading Time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

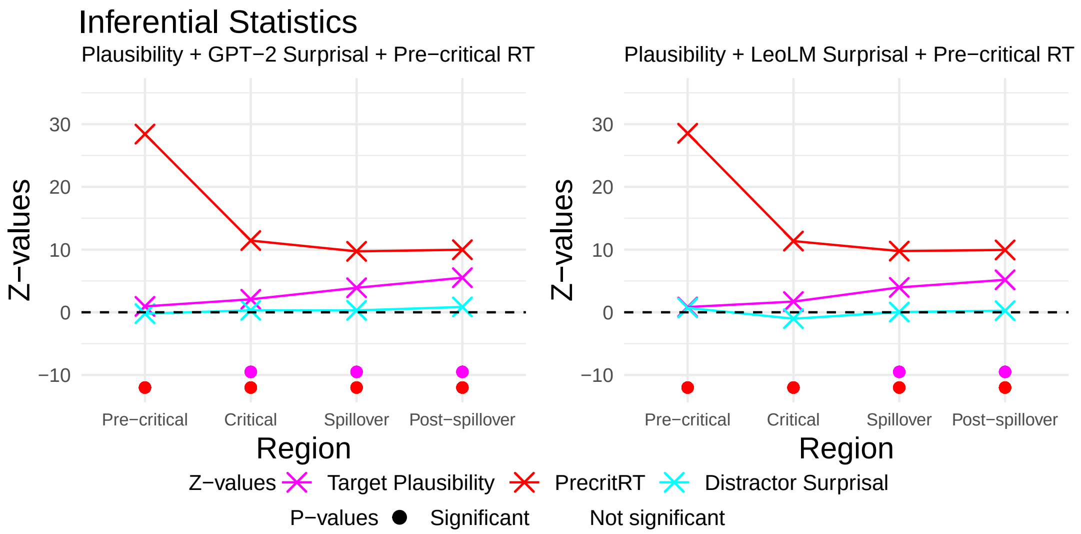

FIGURE 7.7: Effect sizes (z-values) and p-values from the models including Pre-critical Reading Time as a predictor.

7.5 Discussion

The results of the second self-paced reading study show that RTs pattern with the three levels of plausibility associated with Conditions A, B and C (see Figure 7.1). However, the RT pattern $A < B < C$ is more pronounced compared to the observed RT pattern in the first self-paced reading study (see Figure 6.2). Specifically, the difference in RTs between Conditions B and C in the second study is more pronounced than in the first study, resulting in a more evenly graded pattern. Consequently, the results of the second study align with those of Aurnhammer et al. (2023), both of which indicate that RTs gradually increase as plausibility decreases, reflecting the increased effort of integrating less plausible content. Aurnhammer et al. (2023) also found a graded P600 effect for plausibility in a subsequent EEG experiment, where P600 amplitude increased continuously with decreasing plausibility, providing evidence that both RTs and P600 amplitude quantitatively index integration difficulty, which is consistent with the predictions of RI theory (Brouwer et al., 2012). Given that the experimental design employed in the current study is identical (with the exception of the lowered expectancy of the distractor word, which, however, does not influence the presence of a P600 effect), RI theory predicts a graded P600 pattern for the current materials as well. However, multi-stream models would no longer predict a P600 effect in Condition B, but rather an N400 effect, given that the ambiguity was removed by lowering the distractor word expectancy, such that it no longer serves as a semantically attractive alternative.

Furthermore, the observed RT pattern in the second self-paced 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 lead to the observed RT pattern in the first self-paced reading study, which was only graded to a minimal extent due to the similarity of RTs in Conditions B and C. The reason for this seems to be that including 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. As shown in Figure B.10 items rated as medium plausible, irrespective of their condition, have higher RTs in the critical regions due to the increased difficulty in assessing a medium level of plausibility in comparison to very high or low plausibility levels. Moreover, Figure B.11 illustrates the impact of the rating task during self-paced reading on the pre-critical regions. RTs increased strongly on the third pre-critical region in Condition C due to the implausibility of the main verb, which possibly attenuated the implausibility 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 after the exclusion of the online plausibility rating task, the sharp increase in the RTs observed in Condition C disappeared. 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 ratings task, regardless of their condition. Additionally, it would have been interesting to measure 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 on the rating task itself. An examination of the reaction times on the subsequent task, i.e. the comprehension questions, from the first study did not indicate that reaction times were the highest either for Condition B when grouped by conditions or for Group 2 when grouped by plausibility ratings.

As the single-trial ratings averaged per item have been demonstrated to be a more accurate predictor of the RT data than the single-trial ratings, but not compared to the average plausibility ratings from the pre-study (see Figure B.2), this indicates (1) that average plausibility ratings predict the RT data generally more accurately due to the higher stability and robustness of the average and (2) that the combination of the plausibility rating task and the self-paced reading experiment distorted the RTs and potentially also the plausibility ratings. This could be attributed to the fact that participants cannot revisit the context paragraph or the final sentence when they reach the rating task, and thus are only able to provide their ratings based on their memory.

Finally, the linear mixed effects regression analysis showed that plausibility continuously predicts RTs. The estimates (see Figure 7.2) and residuals (see Figure 7.3) of the models including GPT-2 or LeoLM distractor word surprisal alongside average target word plausibility from the pre-study do not show any 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 models containing LeoLM, and additionally in the Critical region in the model including GPT-2 surprisal (see Figure 7.7). In contrast, neither GPT-2 nor LeoLM surprisal is significant in any of the four regions, while it was significant in the Spillover region in the first self-paced reading study for unknown reasons. It is possible that this difference is also a consequence of the distortion of RTs resulting from the combination with the plausibility rating task in the first self-paced reading experiment.

Chapter 8

General Discussion

	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 based on the materials of Aurnhammer et al. (2023) (a) and the materials used in the current study (b) for the N400 and P600 ERP components.

Chapter 9

Conclusion

Acknowledgements

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 Kupferberg, 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 Kupferberg, 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.

Connell, L. and Keane, M. T. (2010). A Model of Plausibility. *Cognitive Science*, 30(1):95–120.

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.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Dinstein, I., Heeger, D. J., and Behrmann, M. (2015). Neural variability: friend or foe? *Trends in Cognitive Sciences*, 19(6):322–328.

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.

Federmeier, K. D. and Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *The psychology of learning and motivation*, 51:1–44.

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. and Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

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.

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

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.

Hutsch, D. F. (2000). Intraindividual variability in cognitive performance in older adults: comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Trends in Neurosciences*, 29(8):474–480.

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.

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

Kupferberg, 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.

Kupferberg, 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. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

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. 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.

Niewland, 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.

Niewland, 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.

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.

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

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.

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

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.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, 89(3):250–280.

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 Average Online Plausibility Ratings

To determine whether the average plausibility ratings collected in a pre-study 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-study 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 used to predict the RTs.

Figure B.1 shows the estimated RTs based on single-trial target word plausibility (top), average target word plausibility collected online (middle) or average 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 average 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 average offline plausibility ratings. This suggests that average plausibility ratings are generally a better predictor of RTs, because they provide more stability and are less susceptible to noise and variability than single-trial plausibility ratings. At the same time, the average pre-test plausibility ratings capture the effects structure in the observed RT data better than the average online plausibility ratings (even though the pre-test ratings were collected from different participants than the RT data), which suggests that the RTs and/or the online plausibility ratings might have been affected by the combination of the reading and rating task (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 average online plausibility. For the better-performing models fitted with average offline plausibility, there is no discernable difference in prediction accuracy between the use of GPT-2 or LeoLM surprisal as the second predictor.

The model coefficients, added to their intercept, for single-trial, average online and 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 average online plausibility are more similar to those for average offline plausibility than to those for single-trial plausibility in the models that include GPT-2 surprisal. Both average online and average 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 average online plausibility fall between those for average offline plausibility and single-trial online plausibility. In this scenario, average online plausibility is significant in the Spillover and Post-spillover regions, while average offline plausibility is found to be 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 average online plausibility and only on the Spillover region in the model with average 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 in Figure B.6. The exact p-values are reported in Table B.1. The resulting coefficients and z-values of average online plausibility are similar to those of average 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 average 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 average online plausibility are similar to those of single-trial plausibility. In both cases, Pre-critical RT is significant across all regions, single-trial and average online plausibility are significant in the Spillover region, while LeoLM surprisal is significant in the Spillover and Post-spillover regions. In the model incorporating average offline plausibility, plausibility is additionally significant in the Critical region, whereas LeoLM surprisal does not significantly predict RTs in any region.

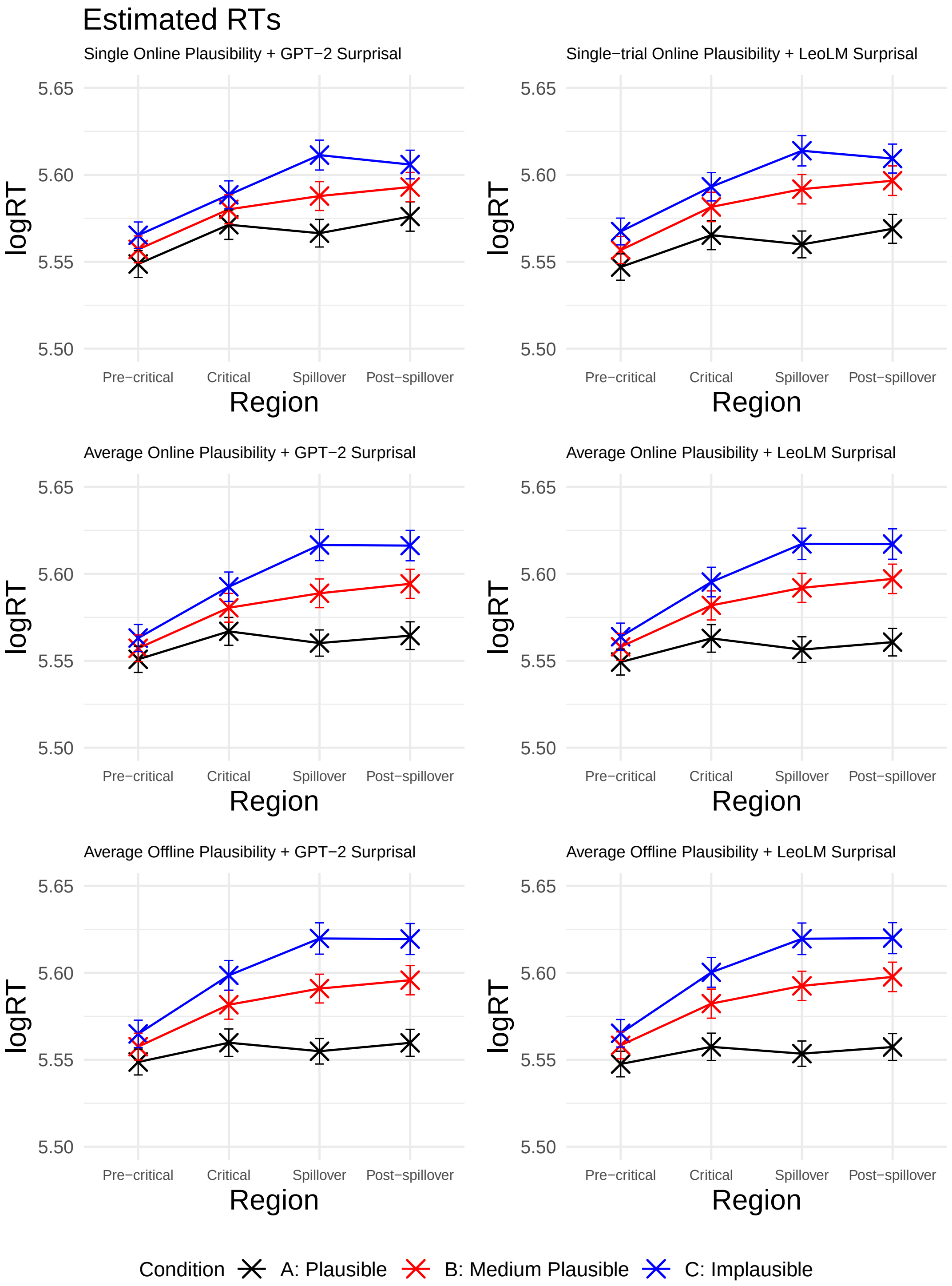

FIGURE B.1: Estimated log-Reading Times using combinations of single-trial (online), average (online), average (offline) Plausibility and GPT-2 (left) and LeoLM Surprisal (right) as predictors per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

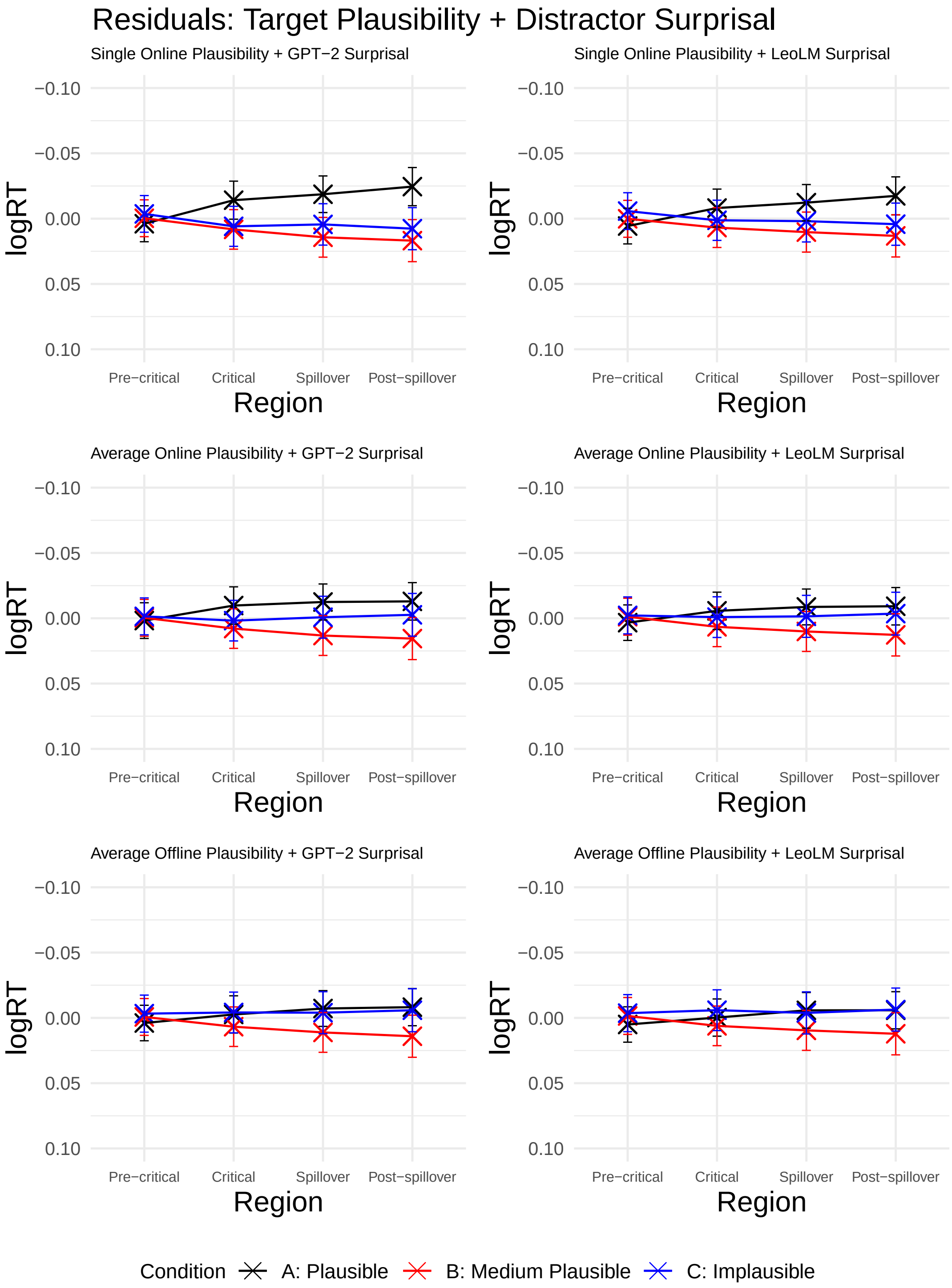

FIGURE B.2: Residual error per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions.

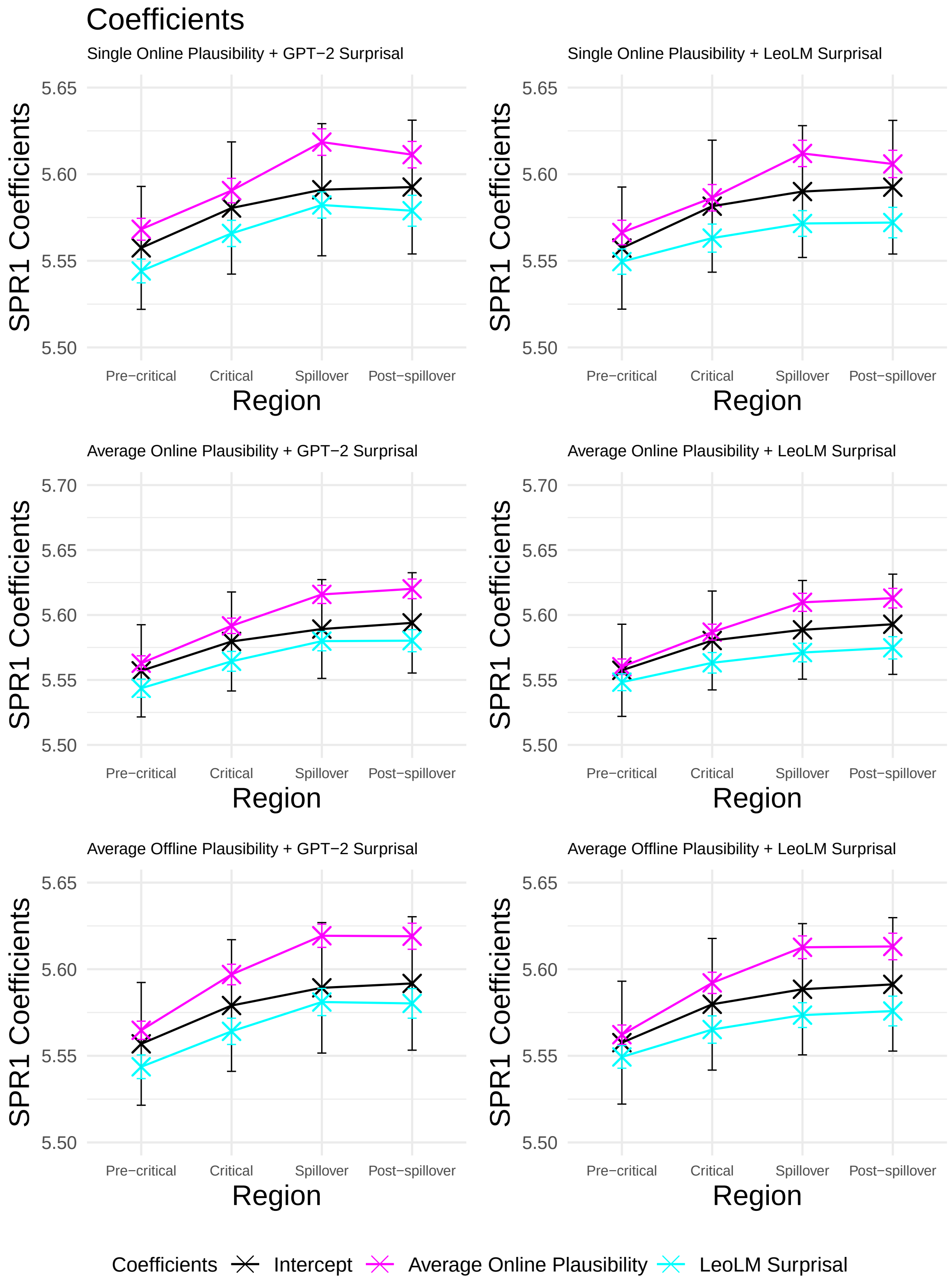

FIGURE B.3: Coefficients, added to their intercept, for the six predictor combinations used for fitting the models. Error bars indicate the standard error of the coefficients in the fitted statistical models.

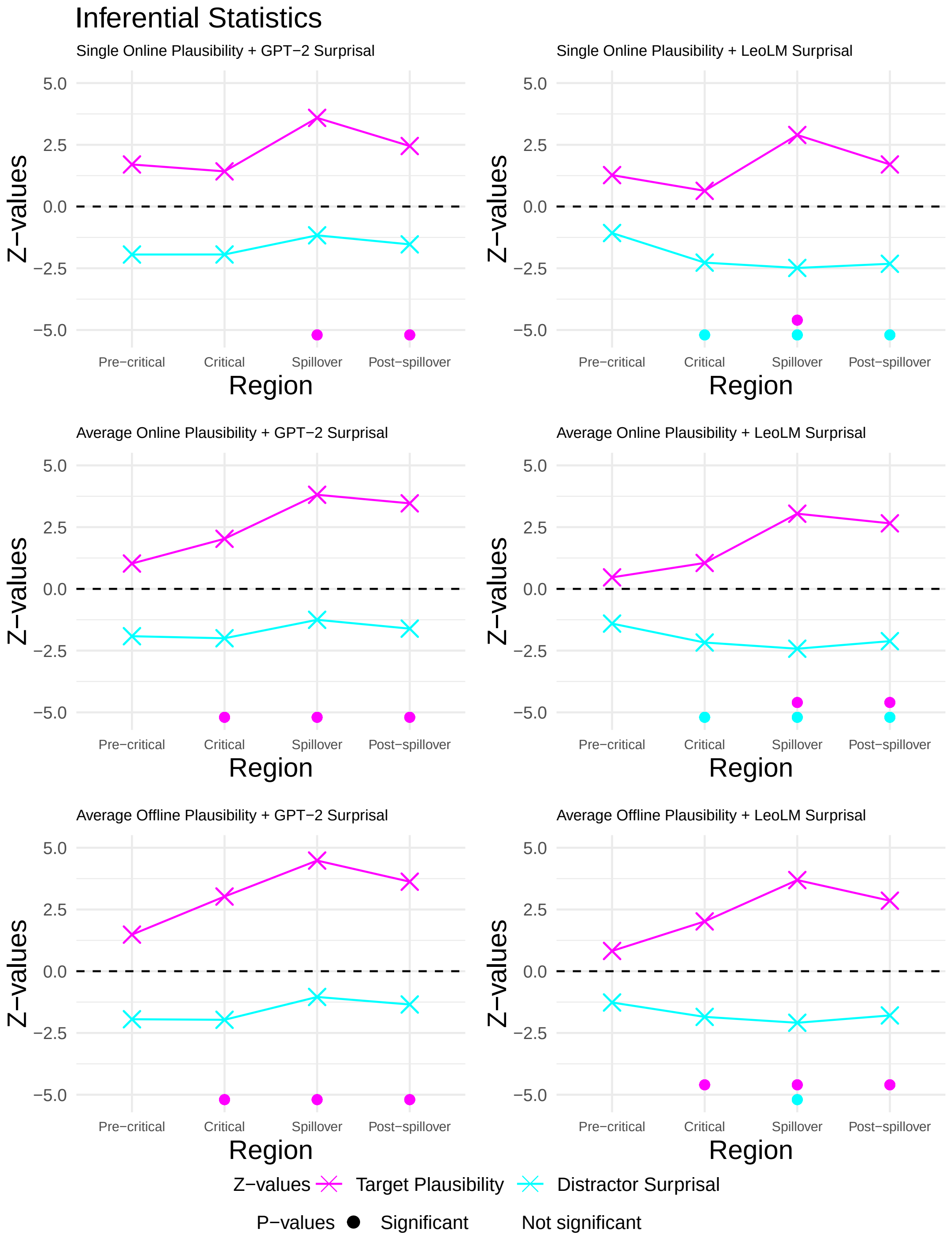

FIGURE B.4: Effect sizes (z-values) and p-values for the six predictor combinations used for fitting the models.

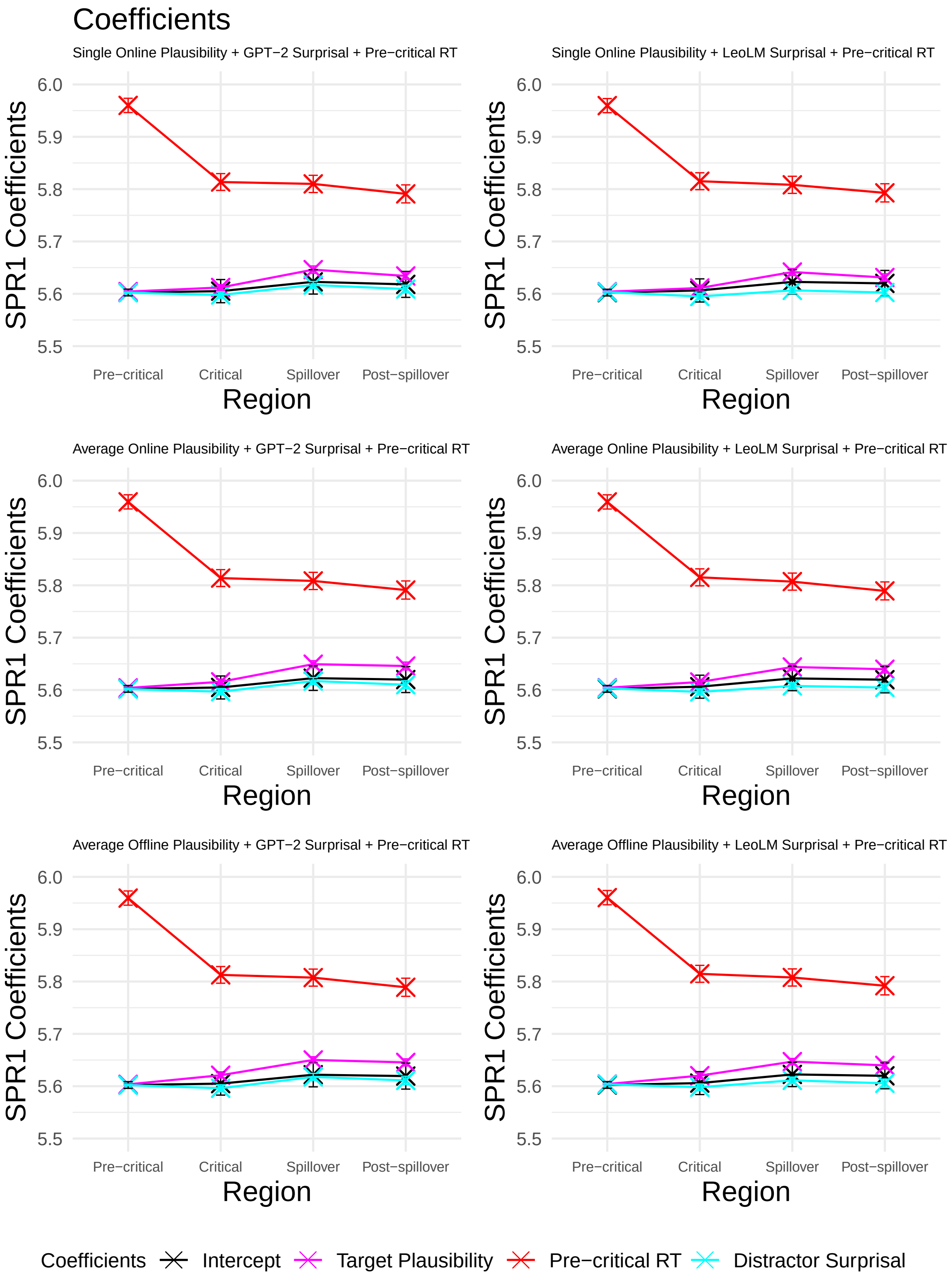

FIGURE B.5: Coefficients, added to their intercept, for the six predictor combinations used for fitting the models, including Pre-critical Reading Time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

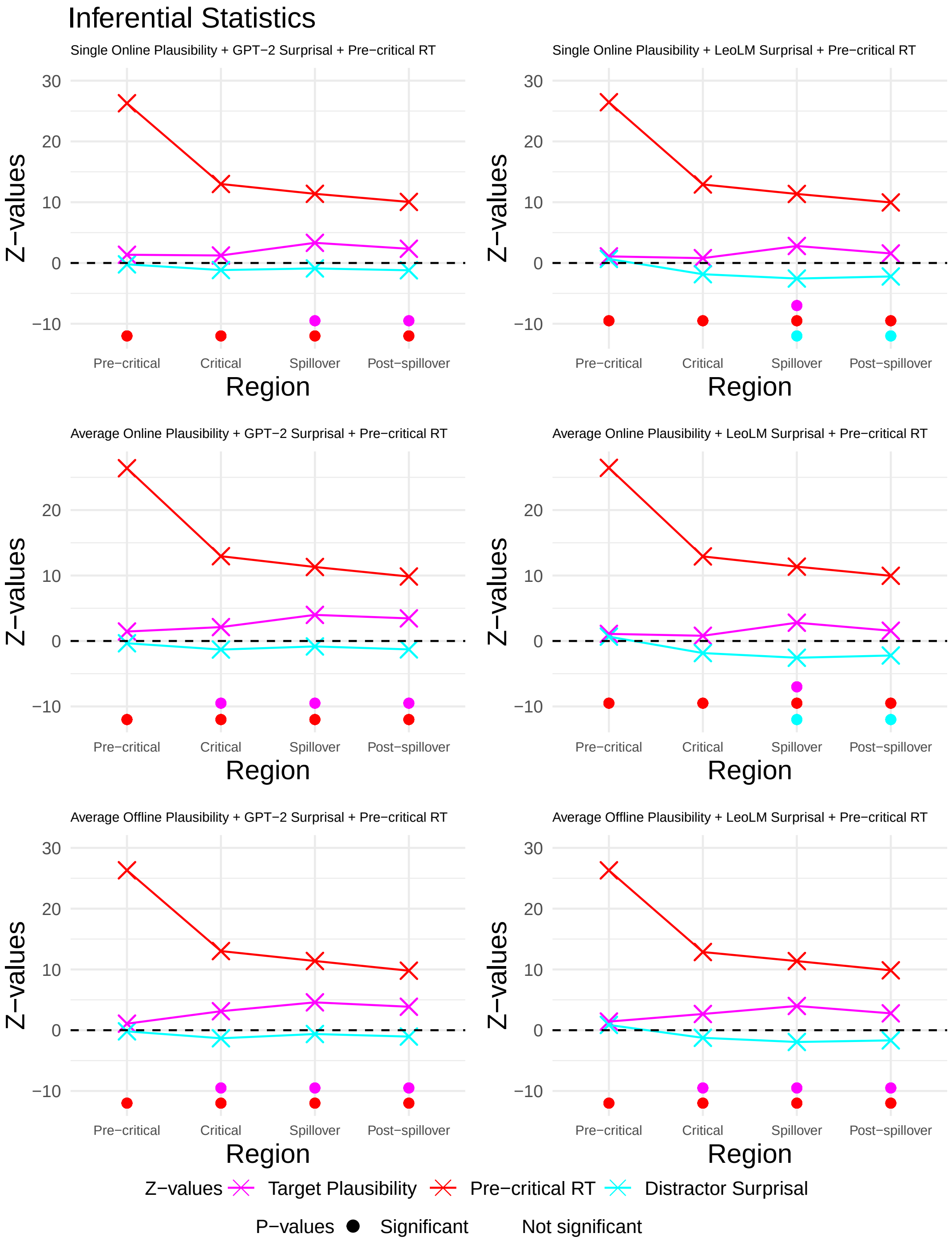

FIGURE B.6: Effect sizes (z-values) and p-values including Pre-critical Reading Time as a predictor.

B.2 P-values

Table B.1 presents the p-values from the first self-paced reading study, which includes the predictors single-trial, average online and offline plausibility, GPT-2 and LeoLM surprisal across all critical regions. Furthermore, the p-values from the aforementioned models including Pre-critical RT as a predictor 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	<2e-16	<2e-16	4.28e-16	1.41e-13
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	<2e-16	<2e-16	5.9e-16	2.06e-13
P-Value	Avg. Online Plaus.	0.309	0.0477	0.000433	0.000969
	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	0.00029	0.00115
	GPT-2 Surprisal	0.728	0.202	0.413	0.208
	Pre-critical RT	<2e-16	<2e-16	4.98e-16	2.18e-13
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	<2e-16	<2e-16	5.9e-16	2.06e-13
P-Value	Avg. Offline Plaus.	0.139	0.00309	4.55e-05	0.00054
	GPT-2 Surprisal	0.0605	0.0572	0.303	0.186
P-Value	Avg. Offline Plaus.	0.413	0.0449	0.000425	0.00532
	LeoLM Surprisal	0.214	0.0714	0.0415	0.0789
P-Value	Avg. Offline Plaus.	0.288	0.00224	4.41e-05	0.000572
	GPT-2 Surprisal	0.851	0.194	0.526	0.357
	Pre-critical RT	<2e-16	<2e-16	4.31e-16	1.5e-13
P-Value	Avg. Offline Plaus.	0.159	0.00916	0.000207	0.00709
	LeoLM Surprisal	0.393	0.225	0.0563	0.102
	Pre-critical RT	<2e-16	<2e-16	4.66e-16	3.8e-13

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 shows the p-values of the predictors employed in the complex model, which includes single-trial and average pre-test 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.3 presents the p-values from the second self-paced reading study for the predictors average pre-test 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 average pre-test plausibility are reported.

		Pre-critical	Critical	Spillover	Post-spillover
P-Value	Average 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	Average 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.2: P-values on each critical region for the complex model fitted with single-trial Plausibility, average 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	Average Plausibility	0.0231	0.00165	0.000164	1.05e-06
	GPT-2 Surprisal	0.196	0.295	0.529	0.817
P-Value	Average Plausibility	0.153	0.0138	0.000376	7.29e-06
	LeoLM Surprisal	0.208	0.120	0.609	0.809
P-Value	Average Plausibility	0.350	0.0426	0.000321	2.09e-06
	GPT-2 Surprisal	0.832	0.753	0.762	0.394
	Pre-critical RT	<2e-16	4.05e-16	7.25e-13	1.10e-13
P-Value	Average Plausibility	0.406	0.0919	0.000217	5.61e-06
	LeoLM Surprisal	0.482	0.312	0.967	0.802
	Pre-critical RT	<2e-16	3.88e-16	6.08e-13	1.07e-13

TABLE B.3: P-values on each critical region for the model fitted with average 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 Raw 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 post-hoc 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 Reading Times filtered 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

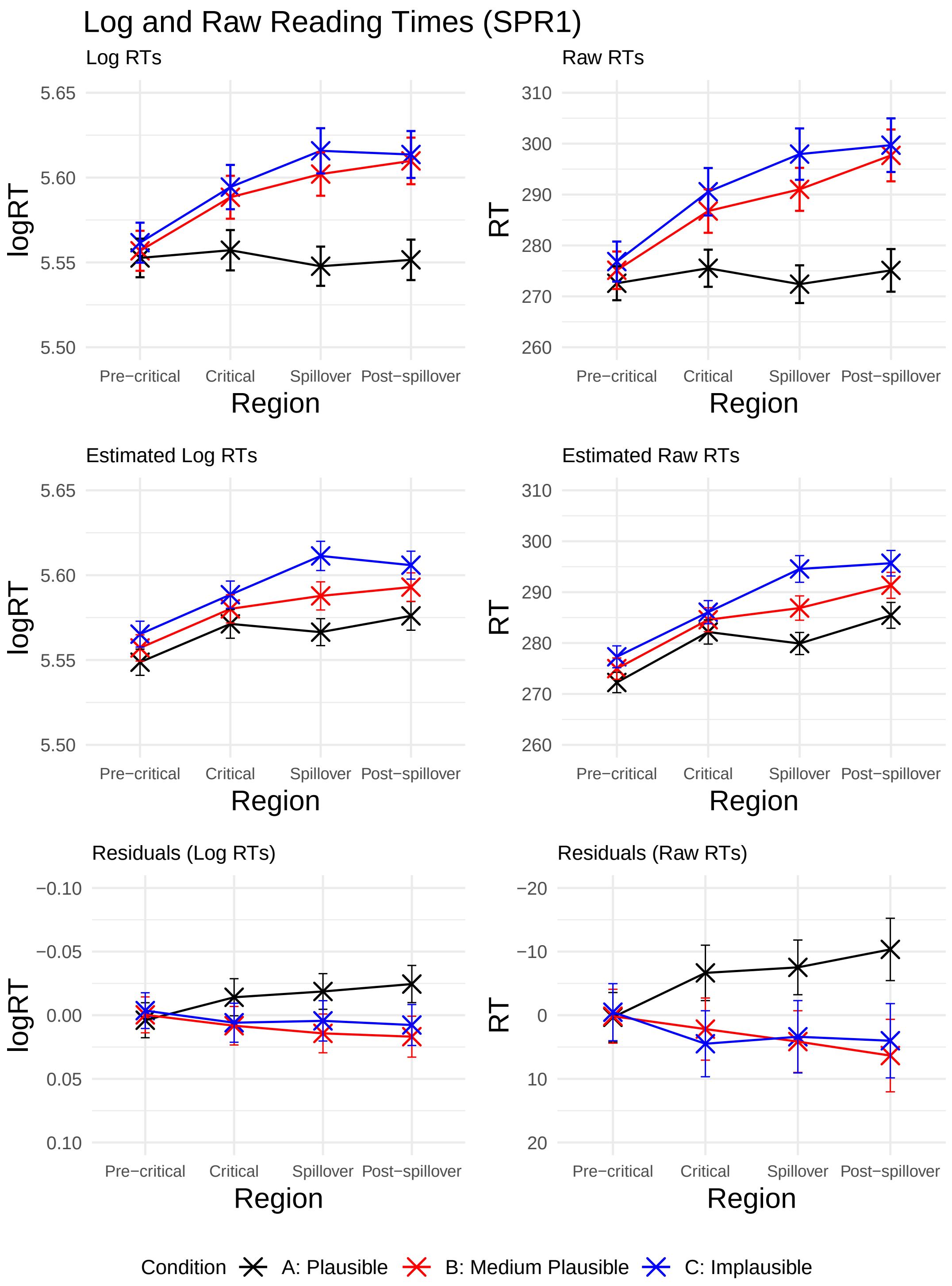

FIGURE B.7: Log Reading Times and raw Reading Times (top), estimated Reading Times using single-trial Plausibility and GPT-2 Surprisal (middle) and residuals (bottom) from the first SPR study per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate the standard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

C. This observation 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 ratings 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 average plausibility ratings collected in the pre-study. 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 average 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 average 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 average ratings per item are more robust and therefore correspond more reliably to the three plausibility levels of Conditions A, B, and C (which probably also makes them better predictors of RTs). The pattern of the observed RTs filtered based on the average 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 average plausibility ratings than when RTs are not filtered or are filtered by single-trial plausibility ratings.

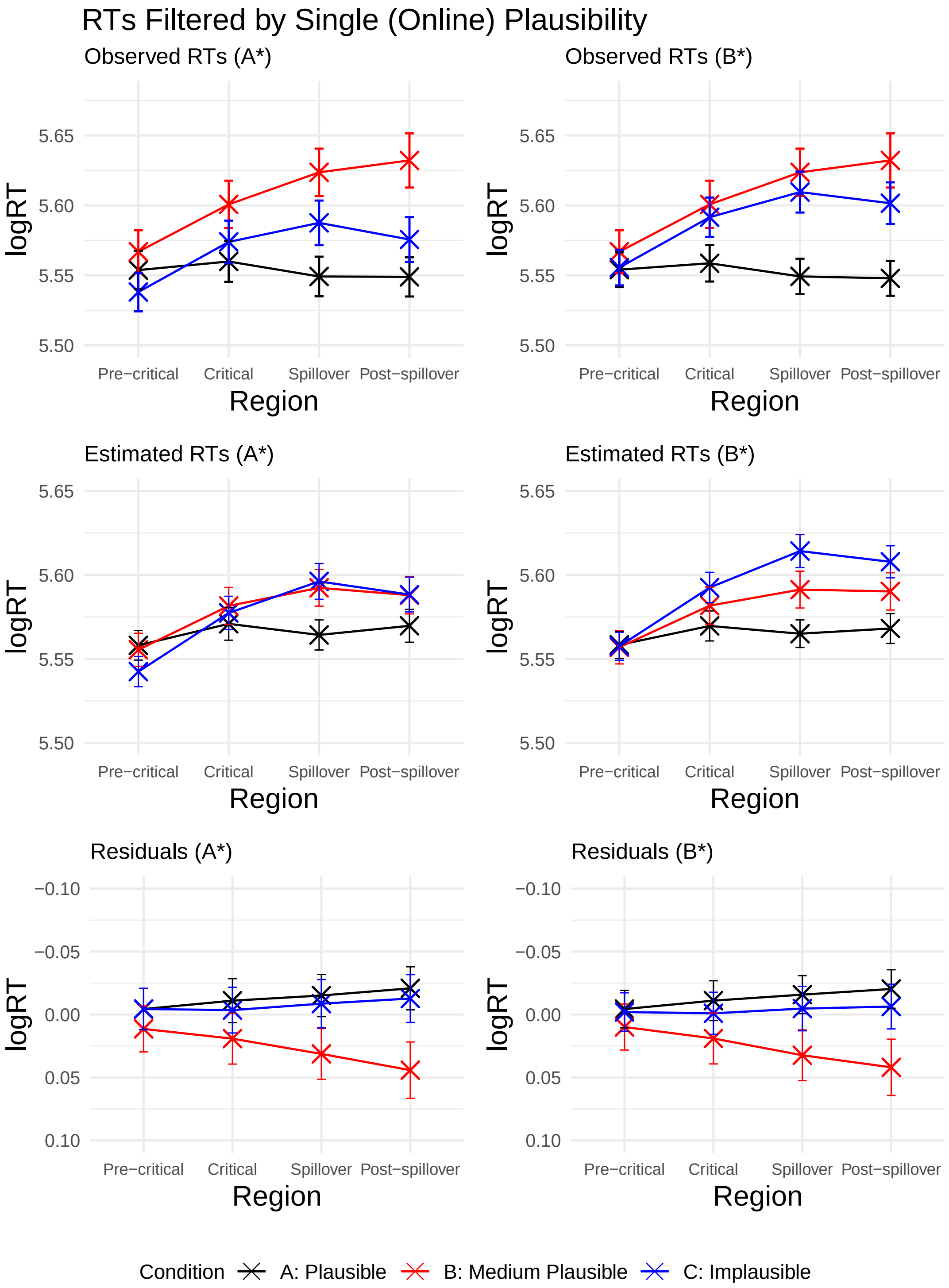

FIGURE B.8: Observed Reading Times (top), estimated Reading Times (middle) and residuals (bottom) per condition across all critical regions, filtered based on single-trial Plausibility. This includes only Reading Times of items rated within range **A*** (**Condition A: 6, 7; Condition B: 3, 4, 5; Condition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

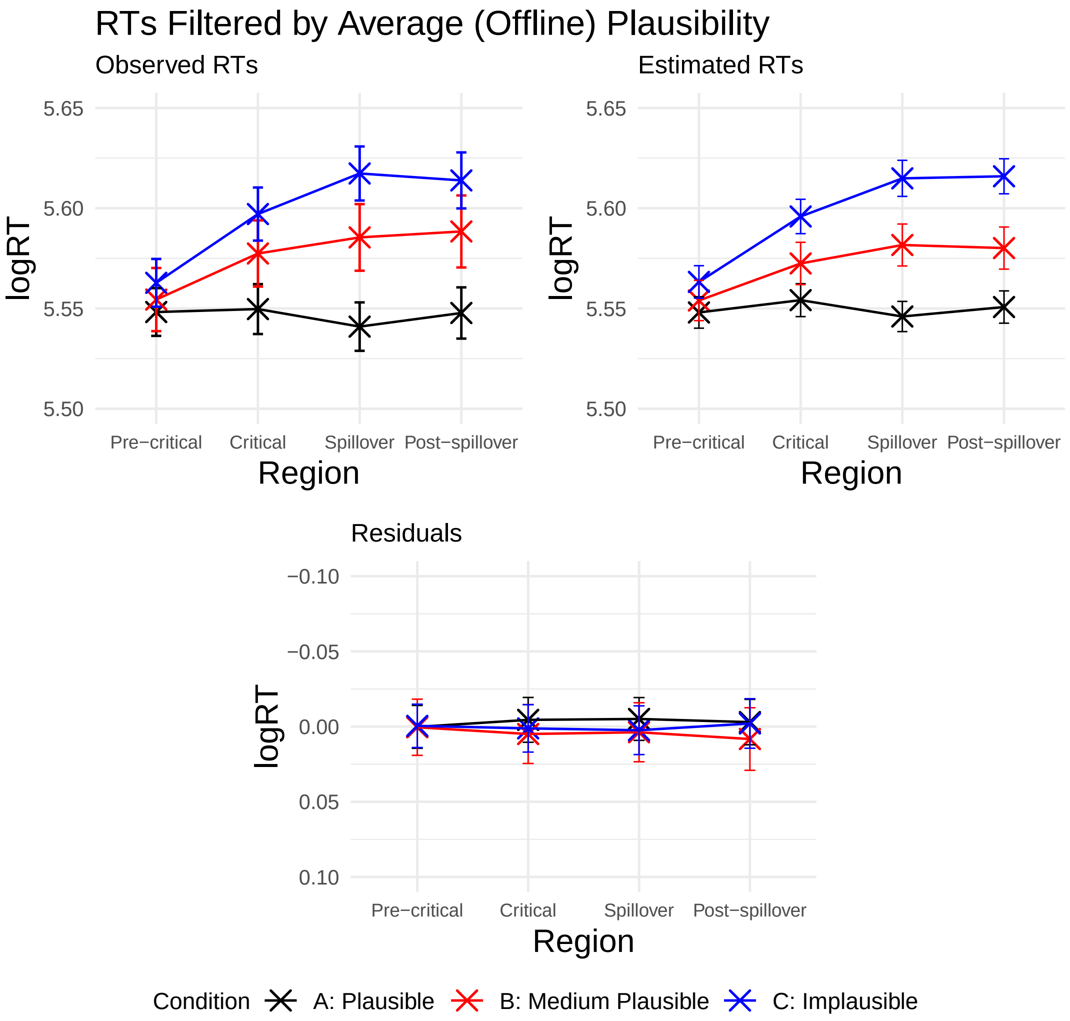

FIGURE B.9: Observed Reading Times (left), estimated Reading Times (middle) and residuals (right) per condition across all critical regions, filtered based on average pre-test Plausibility. This includes only Reading Times of items rated within the range 5-7 (Condition A), 3-5 (Condition B), 1-3 (Condition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

B.5 Reading Times grouped 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 average 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-sized 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 average plausibility ratings from the pre-study, 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 average 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 average plausibility ratings. The observed RTs grouped by the average 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 average plausibility ratings from the pre-study 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.

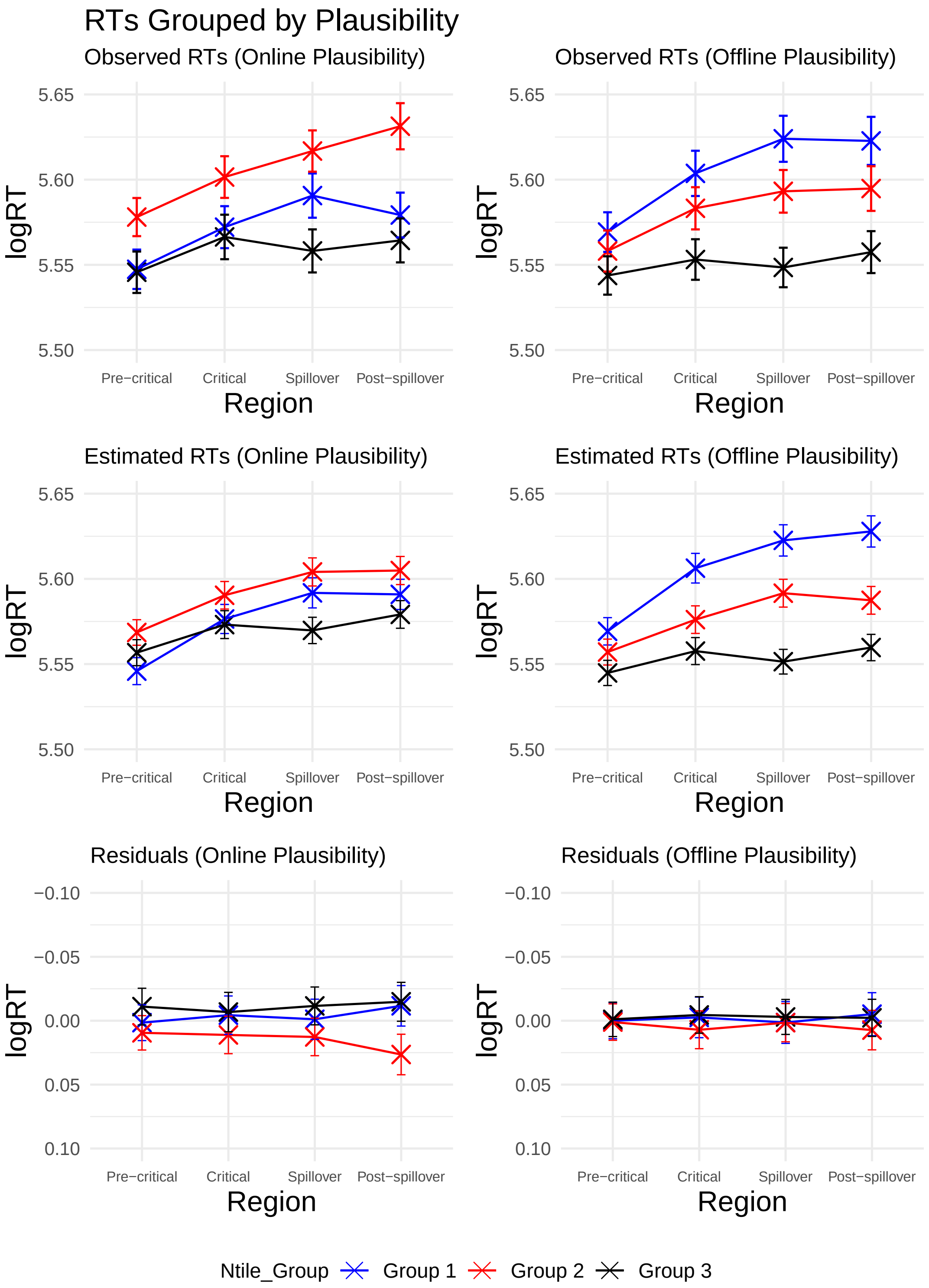

FIGURE B.10: Observed Reading Times (top), estimated Reading Times (middle) and residuals (bottom) grouped by single-trial online plausibility ratings (left) and by average offline plausibility ratings (right).

B.6 Pre-critical Reading Times

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).

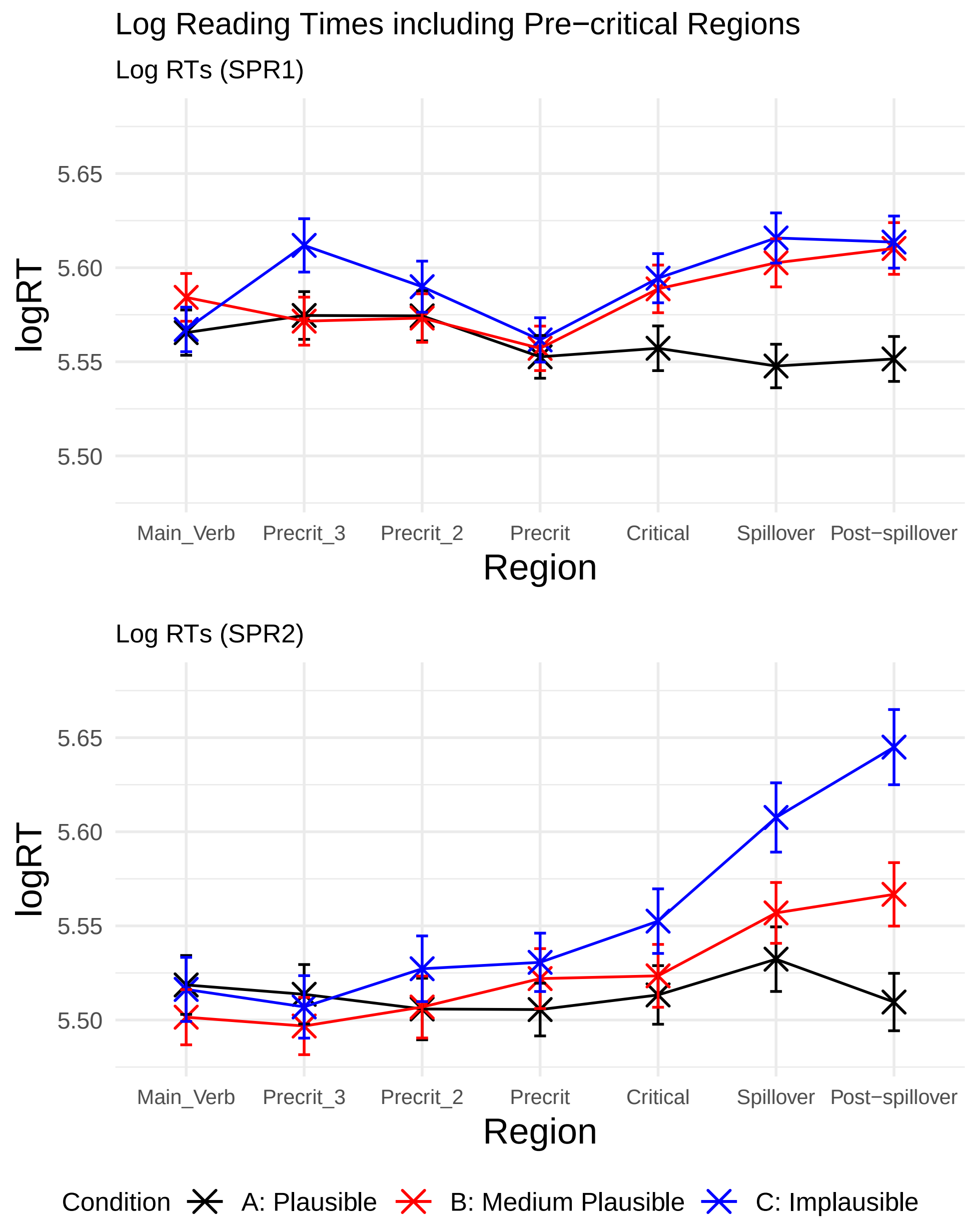

FIGURE B.11: Log-Reading Times from the first and the second SPR study per condition on the main verb, Pre-critical3, Pre-critical2, Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate the standard error computed from the per-subject per-condition averages.

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 implausibility 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 marked increase in RTs in Condition C at the main verb or 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.