Saarland University

Department of Language Science and Technology

Master's Thesis

# Thesis Title blabla blabla

*Author:*
Eva Richter

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

*Advisor:*
Dr. Christoph Aurnhammer

April 16, 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: April 16, 2024

SAARLAND UNIVERSITY

# *Abstract*

Faculty of Humanities
Department of Language Science and Technology

Master of Science

**Thesis Title**
**blabla**
**blabla**

by Eva RICHTER

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

# Chapter 1

# Introduction

Language Comprehension, behavioural vs electrophysiological/neurophysiological studies, reading times (processing effort), RI theory, Plausibility and Surprisal as predictors, single trial analysis etc say that predictions might not just depend on wehther surprisal or cloze is used or which type of LM is used to calculate surprisal but also on whether per-trial ratings are used

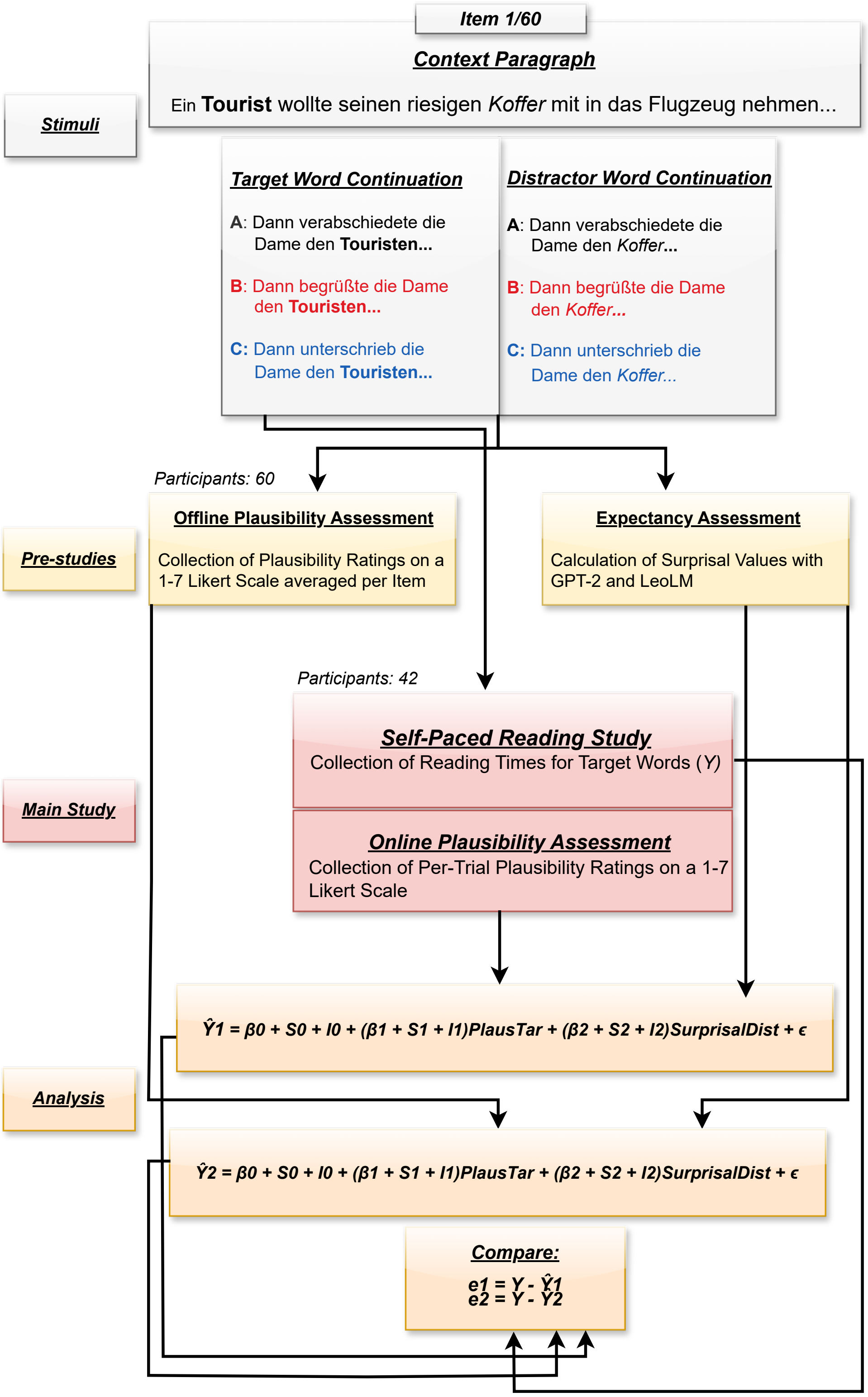
Item 1/60
Context Paragraph
Ein Tourist wollte seinen riesigen Koffer mit in das Flugzeug nehmen...
Stimuli
Target Word Continuation
A: Dann verabschiedete die Dame den Touristen...
B: Dann begrüßte die Dame den Touristen...
C: Dann unterschrieb die Dame den Touristen...
Distractor Word Continuation
A: Dann verabschiedete die Dame den Koffer...
B: Dann begrüßte die Dame den Koffer...
C: Dann unterschrieb die Dame den Koffer...
Participants: 60
Offline Plausibility Assessment
Collection of Plausibility Ratings on a 1-7 Likert Scale averaged per Item
Pre-studies
Expectancy Assessment
Calculation of Surprisal Values with GPT-2 and LeoLM
Participants: 42
Self-Paced Reading Study
Collection of Reading Times for Target Words (Y)
Main Study
Online Plausibility Assessment
Collection of Per-Trial Plausibility Ratings on a 1-7 Likert Scale
Ŷ1 = β0 + S0 + I0 + (β1 + S1 + I1)PlausTar + (β2 + S2 + I2)SurprisalDist + ϵ
Analysis
Ŷ2 = β0 + S0 + I0 + (β1 + S1 + I1)PlausTar + (β2 + S2 + I2)SurprisalDist + ϵ
Compare:
e1 = Y - Ŷ1
e2 = Y - Ŷ2

FIGURE 1: Overview of the steps involved in the current study: 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 the pre-study predictors to predict the reading times observed in the main study.

# Chapter 2

# Background and Related Work

## 2.1 Retrieval Integration Theory

## 2.2 Plausibility

How its similar and different to expectancy surprisal is different bc cloze presents accumulated expectancy at the end of the sentence and not word by word

This suggests that plausibility could be used as a heuristic for assessing the predictability of incoming information, as plausibility judgments take into account the context in which the information is presented. If the information fits well with the context and our expectations, it is more likely to be perceived as plausible. However, it is also possible that an event is unexpected but plausible. -> Michaelov 2023, p. 35

The evaluation of plausibility takes into account the context in which the information is presented. If the information aligns well with the context and our expectations, it is more likely to be perceived as plausible. (txt)

## 2.3 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) and potentially information from other sources such as visual context (Ferreira and Tanenhaus, 2007) 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 actually encountered words with the preceding context depends on the degree to which they align with these expectations. (Kupferberg et al., 2020). The function and even the name of prediction in language comprehension are not uncontroversial (Van Petten and B. J. Luka, 2012; Kupferberg and Jaeger, 2016, for a discussion). Some researchers argue that predicting upcoming words would be inefficient for the language processing system, as these predictions would be more often inaccurate than accurate (Forster, 1981) while others make more nuanced distinctions between the terms *expectation*, *prediction* and *anticipation* (Van Petten and B. J. Luka, 2012). However, evidence from behavioural studies (Schwanenflugel and Shoben, 1985; Stanovich and West, 1983) and ERP studies (Kutas and Hillyard, 1980, 1984), which demonstrate higher reaction times and especially increased N400 amplitudes for unexpected words compared to expected words in a context, shows that context information actually serves to pre-activate features of likely upcoming words such that processing of expected words is facilitated Kupferberg and Jaeger (2016).

Michaelov 2020: One well-known correlate of N400 amplitude is the cloze probability (Taylor, 1953; Bloom and Fischler, 1980) of a word—the probability that it will be offered to fill a specific gap in a sentence by a given sample of individuals in a norming study. All else being equal, higher-cloze completions elicit lower N400 amplitudes (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011). Additionally, even when matched for cloze, words semantically related to the highest-cloze completion elicit lower-amplitude N400s than unrelated words (Kutas, 1993; Federmeier and Kutas, 1999; Ito et al., 2016).

say that there are different operalisations from human judgements based or LMs and they dont capture all notions of expectancy, will focus here on surprisal since thats used in my thesi

However, probabilistic data derived from the cloze test are expensive to collect, and they tend to provide unreliable estimates for low-probability words under realistic sample sizes (Shain et al., 2022). Note that this aspect is problematic not only from a methodological point of view but also from a theoretical one, since differences in processing cost associated with low-probability words are crucial in disentangling between the linear and the logarithmic accounts of the relationship between predictability and processing cost. As a way to offset this complication, researchers have sometimes employed predictability ratings from a normative group of participants, where the extent to which a word could be anticipated from the previous context is evaluated on a Likert scale 3rd option: surprisal estimates -> de varda Computational estimates of word predictability have the undeniable advantage of generating probability distributions over the whole vocabulary, and thus are particularly suited to model the low-probability tail of the distribution. Computational estimates are an interesting option also from a methodological perspective, since they can account for human performance without requiring human annotation; each of the three options has its own strengths -> see de Varda for comparison of the 3 SEE ESPECIALLY DE VARDA 2023 PAGE 2-3 for comparison cloze and surprisal Diese Operationalisierung (siehe auch Kapitel 2) gilt als objektiver, ist kostengünstiger zu erheben und generiert Wahrscheinlichkeitsverteilungen über den gesamten Wortschatz, während Lückentext-Wahrscheinlichkeiten den Nachteil haben, unzuverlässige Schätzungen für Wörter mit geringer Wahrscheinlichkeit zu liefern (Smith und Levy 2011). Welche der beiden Metriken geeigneter ist, ist unklar und die Ergebnisse unterscheiden sich (Smith und levy, shain, 2022 Hofman, michaelov 2022)., each might be better for different purpose.

### 2.3.1 Language Model 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, reading times 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 values. Monsalve et al. (2012) observed that given a sufficiently large and general training corpus, word surprisal is a better predictor of RTs than POS-based surprisal. However, word surprisal effects might not be captured when the training corpus is relatively small or specific, but syntactic patterns might still be modeled due to the higher frequency of occurrence of each possible POS (compared to words) in the training corpus. Crucially, a significant surprisal effect can be missed in any case when spillover regions are disregarded, i.e. when surprisal is only analysed in relation to the current item without considering its influence on the following item (Monsalve et al., 2012).

In neurophysiological 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). Typically, higher N400 amplitudes are observed in response to stimuli with higher surprisal values, indicating a greater degree of semantic incongruity or unexpectedness. Conversely, the P600 ERP component has yet to be studied in this context.

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. Hale (2001) used a *probabilistic Earley parser* based on a phrase structure grammar (PSG) to generate reading time predictions. PSGs rely on the hierarchical syntactic structures of sentences, while linear representations such as $n$-gram models (also known as Markov models) and Recurrent Neural Networks (RNNs) depend only on the sentences' sequential structure. Since $n$-gram models estimate word probabilities by considering only the previous $n-1$ words, their limited access to the preceding context makes them cognitively implausible, despite their accuracy in certain cases (Frank et al., 2015). In contrast, RNNs are sensitive to all the sentence's previous words instead of just the previous $n-1$ (Elman, 1990). Since this architecture reflects incremental word-by-word processing, which is believed to occur during human language comprehension, RNNs are considered highly plausible cognitive models of temporal processing and are widely used in psycholinguistics (Monsalve et al., 2012; Rabovsky et al., 2018; Michaelov and Bergen, 2020). In practice, RNNs have difficulty in capturing input sequences of increasing length (Hochreiter, 1991), analogous to humans having "distance-based memory costs" (Keller, 2010). This view is consistent with results, which show that sequential models, which do not rely on language-specific assumptions, perform better in fitting RT (Frank and Bod, 2011) and EEG data (Frank et al., 2015) compared to hierarchical PSGs. However, these findings were not replicated by, for example, Monsalve et al. (2012) and found to be "premature" by Fossum and Levy (2012), who argued that perplexity [1] and not syntactic capacity determines the ability of LMs to predict RTs. More recent work (Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019; Merkx and Frank, 2021; Wilcox et al., 2020) confirmed this finding, while also finding differences among model architectures after controlling for perplexity. Moreover, discrepancies in results like in the case of Frank and Bod (2011) and Monsalve et al. (2012) could also be a consequence of different language comprehension contexts – self-paced reading of isolated sentences with carefully controlled materials as opposed to reading of more naturalistic texts.

Another recently developed language model architecture, known as the Transformer network (Vaswani et al., 2017), has outperformed previous LMs in various Natural Language Processing (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). Unlike RNNs, the Transformer architecture doesn't rely on recurrence for processing sequential information. Instead, self-attention layers allow them to access all parts of the previous input directly, which means that processing is not incremental over time as it is thought to occur in humans. Despite their lower cognitive plausibility, larger, more sophisticated Transformer-based LMs[2] 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. Moreover, Merkx and Frank (2021) showed that Transformer-based LMs also outperform even improved types of RNNs in explaining both RT and N400 data from word-by-word reading experiments. They explain these results by suggesting that the self-attention patterns of Transformer-based language models are reflective of cue-based retrieval theory (Parker et al., 2017), rather than incremental language processing.

However, contrasting findings have also been reported specifically for reading times. For instance, Arehalli et al. (2022) and Hahn et al. (2022) found that large neural LMs tend to underpredict RTs of targeted constructions. Similarly, Oh and Schuler (2022) observed that surprisal estimates derived from variants of the pre-trained

[1] Perplexity quantifies the predictive power of a LM. Lower perplexity typically corresponds to better model performance, as it indicates that the model is less surprised and can better predict the next word in a sequence.

[2] This applies not only to Transformer-based LMs but also to other types of LMs (see Frank et al., 2015; Aurnhammer and Frank, 2019)

GPT-2 LM, with more parameters and lower perplexity, are less predictive of self-paced reading times during naturalistic reading, which contradicts the findings of Goodkind and Bicknell (2018) and Wilcox et al. (2020). Further investigations have shown that as 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 modelling.

In terms of ERP components, (Michaelov et al., 2021), similar to Merkx and Frank (2021), found evidence that Transformer-based models are more accurate in predicting N400 amplitude, despite being less cognitively plausible than RNNs. Assuming N400 amplitude is influenced by both word predictability and the semantic relatedness of preceding words (Kutas and Federmeier, 2011), Michaelov et al. (2021) further demonstrate that Transformer-based LMs appear to better capture semantic facilitation effects (see also Misra et al., 2020) involved in human language comprehension. Given that Transformer-based LMs have a perfect memory of the entire context window, they are able to leverage semantic relationships effectively, whereas RNN-based LMs cannot store individual previous words explicitly. This suggests that RNNs may be more effective in modelilng aspects of language comprehension related to limited working memory, while Transformer-based models may excel in modelling semantic facilitation effects, which may also explain the conflicting findings regarding reading times (Merkx and Frank, 2021; Oh and Schuler, 2022).

However, all LMs, regardless of their architecture and size, share one universal feature, namely that their estimates are based solely on statistical language patterns, without taking into account world knowledge (Venhuizen et al., 2019) or extralinguistic influences such as gestures and images (Kutas and Hillyard, 1980), as humans do. In contrast, cloze probabilities, representing a human-derived measure of expectancy, have the advantage of taking into account not only participants' knowledge of language, but also their knowledge of the world. For this reason, they have traditionally been regarded as the preferred and most effective (Smith and Kevy, 2011) method of measuring word predictability. It's therefore somewhat unexpected that contemporary LMs such as GPT-3 have been found to be a better predictor of N400 amplitude than cloze probabilities (Michaelov et al., 2023), suggesting that the cognitive processes reflected in neural [3] measures may be more shaped by the statistical properties of language than previously thought.

### 2.3.2 Comprehension-centric Surprisal

write about predictions RI theory makes in terms of RTs, N400 and P600 for expectancy (and plausibility)

[3] And probably also behavioural measures

## 2.4 Single-Trial Analysis

## Chapter 3

# Research Questions

bj

# Chapter 4

# Materials and Methodology

This section explains the structure of the stimuli on which the pre-studies and the self-paced reading experiment are based. It also describes the language models used to compute the surprisal values to assess target and distractor word expectancy in the second pre-study as well as the linear mixed effects regression re-estimation technique used to analyse the reading time 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 <u>tourist</u>") or a continuation rendered implausible by an implausible target word ("the woman told the <u>suitcase</u>").

Aurnhammer et al. (2023) modified 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 <u>tourist</u>"), intermediately plausible (Condition B: "the lady *weighed* the <u>tourist</u>") or implausible (Condition C: "the lady *signed* the <u>tourist</u>"). In this way, Aurnhammer et al. (2023) could test whether reading times and P600 amplitude are graded for plausibility, as predicted by RI theory, which assumes that P600 amplitude continuously indexes the effort required to integrate the meaning of the recently encountered word with the meaning of the unfolding utterance representation. Additionally, the main verb of the final sentence in Condition B was chosen in such a way that the expectancy of another, so-called distractor word, is higher than the expectancy of the target word. That means, in the final sentence "the lady weighed the <u>tourist</u>", the distractor word <u>"suitcase"</u> is globally available as a semantically attractive alternative, although it never appears in target position. Precisely because of this unfulfilled expectation, multi-stream models predict a P600 effect for Condition B relative to baseline and an N400 effect for the implausible Condition C, for which no such semantically attractive alternative is available. RI theory predicts a graded P600 effect, as integration difficulty increases with decreasing plausibility. As discussed, Aurnhammer et al. (2023) found a graded P600 amplitude relative to baseline and no N400 effect, and hence evidence for the assumptions of RI theory.

Similar to Aurnhammer et al. (2023), the 60 items selected for the present studies consist of a context paragraph and a manipulated final sentence. The target word is kept the same across conditions to minimize potential effects due to word length or word frequency. For similar reasons, the same case, typically the accusative, is used in all conditions within each item. In addition no separable verbs (e.g. "Dann **bereitete** der Mann das Essen **zu**") were used to make sure that the entire main verb can be integrated with the preceding context prior to reading the target word. Another constraint was to avoid reflexive verbs, 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 ("Dann verabschiedete die Dame"). However, this cannot always entirely achieved, given that, especially in Condition C, the main verb itself often already introduces some degree of implausibility.

The context paragraph is again 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 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. Similarly, the final sentence varies across conditions only regarding the main verb which also renders the sentence plausible (Condition A: "the lady *dismissed* the <u>tourist</u>"), intermediately plausible (Condition A: "the lady *welcomed* the <u>tourist</u>"), or implausible (Condition A: "the lady *dismissed* the <u>tourist</u>"). 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 reading times. In Aurnhammer et al. (2023), the final sentence of Condition B is additionally ambiguous due to the availability of a semantically attractive alternative. In contrast, 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 <u>tourist</u>" to "the lady *welcomed* the <u>tourist</u>" the expectancy for the distractor word <u>"suitcase"</u> was eliminated in this context and the target word <u>"tourist"</u> should now have a higher expectancy, while being still less plausible compared to the baseline Condition A. Hence, the stimuli used in the current study differ only with respect to the main verb in Condition B from the stimuli created by Aurnhammer et al. (2023). Figure 2 shows an item

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

in the three conditions in the current study compared to an item in the study of Aurnhammer et al. (2023).

*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 2: 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.

As the three plausibility levels are maintained, the subsequent self-paced reading study should show, analoguous to the results of Aurnhammer et al. (2023), that reading times scale with plausibility for Conditions A < B < C, reflecting increased integration effort as plausibility decreases. In order to assess whether the items have been successfully manipulated in terms of plausibility and expectancy of the target and distractor words, and in particular whether the ambiguity in Condition B has been removed, two norming studies are conducted prior to the self-paced reading studies.

---

*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 3: Three example items, transliterated from German. Target words are underlined, distractor words are highlighted in boldface.

This may not be relevant for the following self-paced reading study as the assumptions of multi-stream models and RI theory regarding the ERP components cannot be tested in behavioural studies. However, a future EEG study could further test the predictions of RI theory and multi-stream models more extensively. Based on the design and stimuli employed in the current study the predictions of multi-stream models and RI theory diverge not only for Condition C but also for Condition B. Since the semantically attractive alternative in Condition B has been removed, multi-stream models no longer predict a P600 effect, but also an N400 effect [2]. In contrast, RI theory still predicts a graded P600 effect and no N400 effect, since the presence of a semantically attractive alternative has no influence on the predictions of RI theory. Finally, this could also shed on whether the early negativity observed in Aurnhammer et al. (2023) was due to the unfulfilled expectations in Condition B. If this is the case, the negativity should disappear when conducting an EEG study based on the current data, as the expectancy for the distractor word in Condition B has been is lowered in this study.

## 4.2 Language Model Architecture

The LMs used for computing surprisal in a pre-study of this thesis belong to the Transformer model family. As discussed in Chapter 2, 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). Transformers differ from RNNs primarily in their ability to process sequential information in parallel rather than token by token, allowing them to capture long-range dependencies more effectively.

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 on the context vector generated by the encoder. However, the LMs used to calculate surprisal for the current study, GPT-2 (Radford et al., 2019) and LeoLM (Plüster, 2023), are both decoder-only architectures, which are based solely on the decoder component of the Transformer architecture.

Figure 2 shows the decoder-only architecture. The Transformer model takes a sequence of words as input, which is split into subword units by the tokeniser, before mapping each token to its respective embedding. Since Transformer models don't have an inherent notion of word order, positional encoding vectors are added to the input embeddings for each of the token positions in the input sequence.

This is followed by a decoder block, which consists of a masked multi-head attention layer, a normalisation layer, a feed-forward neural network and a

[2] The model proposed by Rabovsky et al. (2018); Rabovsky and McClelland (2019) as described in Chapter 2 is an exception to this.

normalisation layer. Consequently, the input embeddings are first passed through a layer of masked multi-head attention, in which self-attention is applied. Self-attention is a mechanism that allows each token to attend to all other tokens in the input sequence (including itself), enabling the model to capture the relationships between them. To achieve this, each token is associated with three vectors: *query (q)*, *key (k)* and *value (v)* that are derived from the input embeddings. To calculate attention scores for all queries in parallel $q$, $k$ and $v$ are stored in the respective matrices $Q$, $K$ and $V$ by multiplying their weight matrices learned during training with the input embedding matrix. Attention scores are calculated for each *query* by taking the dot product between the *query* vector of the current token and the *key* vectors of the current and all preceding tokens. A softmax function is then applied to normalise the attention scores into probabilities. Each *value* vector is multiplied by the softmax score and the resulting products are subsequently summed. Thus, the output of the self-attention layer for the current token is the weighted sum of the *value* vectors of all tokens in the input sequence. Tokens with higher attention scores, reflecting higher association with the current token, contribute more to the weighted sum compared to tokens with lower scores. Multi-head attention extends this capability by applying the attention mechanism multiple times in parallel, allowing the model to focus on different aspects of the relationships between the input tokens.

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, enabling the model to generate text in a left-to-right manner without access to future tokens.

In the next step, the output of the multi-head attention layer is added to the original positional input embedding. This so-called residual connection prevents the vanishing gradient problem during backpropagation. Then, the output of the residual connection is normalised to stabilize and speed up training. The normalised residual output is passed through a Feedforward Neural Network (FFNN), the output of which is again added to the input of the FFNN and further normalised.

Since Transformers typically consist of multiple decoder blocks stacked on top of each other (12 in the case of GPT-2 and 32 in the case of LeoLM), the output of the normalised output of the FFNN can either serve as input to the next decoder block or it can be passed to a final linear layer, which projects the output to a layer with the same dimensionality as the vocabulary of the model. The output of the final linear layer represents the logits, which are then converted to a probability distribution over entire the vocabulary by applying the softmax function. The token predicted at the current position corresponds to the index with the highest probability score. Surprisal is calculated as the negative logarithm of the probability assigned to the observed token in the probability distribution.

The LMs used to compute surprisal for the current study both belong to the family of transformer-based autoregressive causal language models, but differ in terms of their size and training data. The first model is a pre-trained German-GPT-2 model (Schweter, 2020), which belongs to the smallest GPT-2 version trained with 124 million parameters. The authors used the same training data as for a German BERT

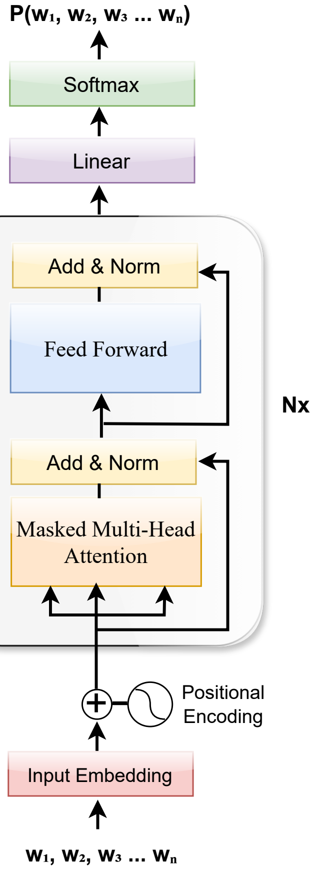

FIGURE 4: The Decoder-only Transformer Architecture on which GPT-2 and LeoLM are based.

model [3], which consists of Wikipedia articles [4], 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 [5]. 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 (Touvron et al., 2023) and exceeds GPT-2 in terms of the number of parameters and the amount of data it was trained on. LLama-2 is a family of LLMs ranging from 7 billion to 70 billion parameters that are pre-trained on approximately 2 trillion tokens of predominantly English texts. For the current study, the medium-sized 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).

[3] https://huggingface.co/dbmdz/bert-base-german-cased. [Accessed: 2024-04-16].
[4] https://dumps.wikimedia.org. [Accessed: 2024-04-16].
[5] https://commoncrawl.org. [Accessed: 2024-04-16].

LeoLM was also trained on two smaller datasets, comprised of Wikipedia [6] and news (Tagesschau) [7] articles, resulting in a training dataset of approximately 600 gigabytes.

As the overview of the parameters of both models in Table 1 indicates, LeoLM is a more advanced model in terms of its size and the resources required for training and inference. Since the models were not evaluated using metrics such as perplexity in the present study, their performance cannot be compared in this context. However, given the sheer difference in size, LeoLM would very likely outperform (i.e. achieve lower perplexity) than GPT-2 in a performance test.

| | **GPT-2** | **LeoLM** |
|---|---|---|
| **Parameters** | 124M | 13B |
| **Vocabulary size** | 50,257 | 32000 |
| **Context size** | 1024 | 4096 |
| **Embedding dimension** | 768 | 5120 |
| **Decoder/Hidden layers** | 12 | 32 |
| **Attention Heads** | 12 | 32 |

TABLE 1: Parameter overview for the two language model architectures used to compute surprisal values in a pre-study: GPT-2 and LeoLM.

## 4.3 Linear Mixed Effects Regression Re-estimation

lalal

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

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

# Chapter 5

# Pre-studies

Prior to the main experiment, two norming studies were carried out to test whether changing the main verb of the final sentence had the desired effects. 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 confirm whether the expectancy for the distractor word in Condition B was lowered, indicating that the ambiguity was removed. The plausibility norming study, as well as the subsequent self-paced reading studies were conducted in a web-based experiment 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 had 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, 7 indicating "plausible" and 1 indicating "implausible". 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. Plausibility ratings were collected for both target words and distractor words in all three conditions to assess if the main verbs were chosen in a way that results in high plausibility for Condition A, intermediate plausibility for Condition B, and low plausibility for Condition C on average and if plausibility is higher for the target words (especially for Condition B) than for the distractor words across conditions on average. The main goal was to assess whether the plausibility of the items in the target condition is graded, as reading times are collected for these items in the subsequent self-paced reading study. Regarding the distractor condition, it is likely that the plausibility in condition B has been decreased compared to Aurnhammer et al. (2023), if the expectancy in Condition B has been successfully lowered given that less plausible items are often also less expected. Ideally, this would result in the plausibility of the distractor condition being graded as well. 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 addition to the 60 critical items, each participant was presented with 12 out of 45 filler items, which are plausible, intermediately plausible or implausible, analogous to the critical items. The purpose of the filler items was to make sure that participants read the texts carefully. Therefore, filler items contained instructions in the middle of the paragraph asking participants to rate the trial either 1 or 7, regardless of the actual plausibility of the sentence. If more than two of these 12 attention checks were failed, the participant's data were excluded from further analyses. Before starting the actual study, participants were also presented with three practice items. Consequently, each participant encountered a total of 75 items, comprising critical items, filler items, and practice items.

A total of 66 participants were recruited through Prolific Academic Ltd. and each was paid £4.95. Six participants were excluded due to exceptionally fast completion of the study (more than 3 standard deviations below the mean) or more than 2 out of 12 failed attention checks. A Latin square design was employed to ensure that each participant read each item in only one condition, and that all conditions received an equal number of ratings from each participant. Therefore, the remaining 60 participants were assigned to six distinct lists, each comprising ten participants who read identical lists of items. These lists varied in terms of conditions from those presented in the other groups. Thus, each participant rated 60 items in total, 10 in each of the six conditions, and conversely, a total of 10 plausibility ratings were collected for each of the 360 different final sentences, resulting in 3600 trials in total.

Participants who met the criteria of being a native German speaker, aged between 18 and 32 years, and without any language-related disorders 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, intermediately 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 actual study began. Throughout the study, the context paragraph was presented along with the final sentence and the 1 to 7 scale to ensure that participants rated the final sentence in the context of the preceding paragraph, thus avoiding any instances of forgetting or skipping the context. In contrast to Aurnhammer et al. (2023) the continuation of the final sentence ("and after that he left the store") was not excluded for consistency reasons, because plausibility ratings were also collected during the subsequent self-paced reading study, in which the continuation is crucial to capture spillover effects. After completing the study, participants were asked to provide demographic information.

### 5.1.2 Results

On average, participants completed the study in 30.57 minutes and 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 2 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 the target word plausibility is graded ($A > B > C$) across conditions, indicating that the main verbs in the final sentences were successfully selected such that participants considered Condition A as plausible, Condition B as intermediately 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 very high (7) or low (1) ratings to items in Condition B. However, this is to be expected given the challenge of rating items of intermediate plausibility on a scale of 1 to 7, compared to rating items that are either clearly plausible or clearly implausible.

Regarding the distractor word, the average plausibility ratings are also graded ($A > B > C$). However, the average ratings for all three conditions show only slight differences and fall within the lower (implausible) range overall. 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 rather implausible ("Then **dismissed** (A)/**welcomed** (B) the lady the tourist (target)/suitcase (distractor)"). Even though the expectancy of the target and distractor word is assessed in a separate norming study, this study can also offer insights on their expectancy, as more plausible continuations generally tend to be more expected than implausible ones. Since the average target word plausibility in Condition B is higher than the distractor word plausibility, this indicates that the expectancy of the distractor word has been successfully lowered in Condition B, while keeping the plausibility levels across conditions unchanged[1]. Finally, it is worth noting that while the plausibility of the target word is higher than that of the distractor word in Conditions A and B, the opposite is found for Condition C. This could be attributed to the combination of the implausible main verb in condition C and the, compared to the target word, usually less expected distractor word, which, at least in some cases, renders the final sentence slightly more plausible than when the target word is used (see also the example items in Table 3). 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 than signing people. However, this should not affect the subsequent analyses, as both target and distractor words were rated low in plausibility on average and the main objective, i.e. to obtain a graded effect for target plausibility, was achieved.

Figure 5 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, averaged per item across participants, for the target word. The vertical dashed lines show, (similarly to Table 2), the average plausibility ratings per condition and demonstrate that the average plausibility decreases from A to C. Conditions A and C show a unimodal distribution, peaking at plausibility levels 1.5 and 6.5 respectively. The distribution of Condition C is slightly skewed to the right and the distribution of Condition A is slightly skewed to the left, with both their tails extending towards the centre of the plausibility range (4). 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 high ratings to

[1] For comparison, see also Aurnhammer et al. (2023), who observed higher average plausibility and expectancy values for the distractor word in Condition B compared to Condition A due to the active semantically attractive alternative.

the items in Condition C and low 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 average ratings in the range of intermediate plausibility. However, the average ratings are more evenly distributed 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. In addition, the slightly more prominent peak at 3 indicates that most items in Condition B fall at the lower end of the (still) intermediate part of the scale. As mentioned previously, this result is not surprising, since assessing intermediate nuances is inherently more challenging than assessing clearly plausible or implausible ones.

The plot on the top right shows the distribution of plausibility ratings, averaged per item across participants, for the distractor word. The plausibility ratings per item, particularly in Conditions A and B, are more similar to each other than those of the target word. This is also reflected in the average ratings per condition in Table 2. All three distributions show a unimodal distribution and are skewed to the right. Condition C has a higher peak (at 1.5) and a shorter tail compared to Conditions A and B, indicating that most items were perceived as clearly implausible on average. Conversely, Conditions A and B also have a (smaller) peak in the implausible range at around 1.8 and their tails extend across the entire plausibility scale. This indicates that on average the majority of items in Conditions A and B were also considered to be implausible. However, there are more instances where items received on average higher plausibility ratings compared to Condition C. The distribution of the target and distractor plausibility may appear similar in Condition C, as both are implausible. However, Condition B and, to an even greater extent, Condition A are rendered mostly implausible only when they contain the distractor word.

In this study, the average plausibility for the target word is slightly higher across all conditions compared to the average target plausibility reported in Aurnhammer et al. (2023), who collected plausibility ratings using the same stimuli, except of Condition B. Importantly, the observed pattern for the target words is consistent with the findings of Aurnhammer et al. (2023). However, in the study by Aurnhammer et al. (2023) the average plausibility for the distractor word in Condition B is higher than the average plausibility for the distractor word in Condition A. This is due to the higher expectancy of the distractor word within their Condition B, which renders the distractor word in Condition B more plausible compared to the distractor word inpreceding the subsequent pre-study where the expectancy of both target and distractor words will be explicitly evaluated Condition A. As one of the goals of this study was to reduce the expectancy of the distractor word in Condition B, this resulted also in lower average plausibility of the distractor word in Condition B compared to the plausibility of the distractor word in Condition A. At the same time, this observation suggests that the expectancy of the distractor word has been effectively reduced in Condition B, prior to conducting the second pre-study in which the expectancy of the target and distractor word is explicitly evaluated.

It remains to be seen whether the same pattern will be observed for the per-trial plausibility ratings collected (online) during the first self-paced reading study. Based on the results of this plausibility study and those of Aurnhammer et al. (2023), the same pattern $(A > B > C)$ can be expected. In addition, the study will show whether the reading times for the target word will be graded for plausibility, with implausible items being read more slowly compared to plausible items, $(C > B > A)$, indexing 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, indicating the absence of the distractor word as a semantically attractive alternative. 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), the expectancy of the target words should again 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. The latter is considered to be more objective and cheaper, and has the advantage of generating reliable probability distributions over the entire vocabulary. However, it is unclear whether and if so, which of these metrics performs better on which type of data (see also Chapter 2).

Since the average plausibility was shown to be graded across conditions $(A > B > C)$ in the plausibility norming study, it is likely that the average expectancy of the target and distractor words will also be graded across conditions, given that sentences with higher plausibility are usually also more expected. Since surprisal is inversely proportional to expectancy, the average surprisal values per condition are expected to follow the reversed pattern $(C > B > A)$ compared to the average plausibility values per condition. However, more crucially, the expectancy of the target words should be higher than the expectancy of the distractor words, reflected in higher surprisal values for the distractor words compared to the target words across conditions.

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) [2] 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 in the plausibility rating study [3]. They consist of target and distractor words in conditions A, B and C for 60 items, resulting in 360 item variations in total. First, the stimuli were preprocessed using a regular expression, which inserted a whitespace

[2] In fact, the surprisal values were calculated based on a re-trained GPT-2 version that the author had shared due to a bug in the tokeniser (https://huggingface.co/stefan-it/secret-gpt2. [Accessed: 2024-4-15]).

[3] Except of the filler items, for which obviously no surprisal values were calculated

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. BPE is a data compression technique that was adapted by Sennrich et al. (2016) for word segmentation. It works by iteratively merging the most frequent pairs of consecutive characters in the input text. The input text is segmented into subword units, usually individual characters, that are treated as separate tokens. The most frequent pair of consecutive tokens, such as "*A*" and "*B*", is merged into a new single token "*AB*", with which the vocabulary is subsequently updated. This procedure is repeated for a fixed number of iterations or until a stopping criterion (e.g. a predefined vocabulary size) is reached. Finally, each word in the input text is tokenised into a sequence of subword units. BPE is a commonly used tokeniser in transformer models because it reduces the model's vocabulary size while maintaining its expressive power. Simultaneously, this enables the model to generate accurate predictions for rare or Out-of-Vocabulary words by using the subword representations it has learned during training.

In this context it should be briefly noted that 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 of predicting reading times using BPE tokenisation compared to morphological segmentation, the former should be used with more caution, as it is less psychologically plausible. To avoid this problem, Michaelov et al. (2023) only included items that were not split by the tokeniser. 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.

As only the surprisal values for the target and distractor words are relevant, in a second preprocessing step, the final sentence of each item was truncated after the target/distractor words, excluding the continuation of the final sentence. For example, the original stimulus "*[...] Dann verabschiedete die Dame den Touristen und danach ging er zum Gate*" was presented as "*[...] Dann verabschiedete die Dame den* ***Touristen***", i.e. without the final sentence continuation ("und danach ging er zum Gate").

Subsequently, the tokeniser segmented each item into BPE tokens and converted them into a sequence of ids which was used as in input to the LM. The considerable size of the context (paragraph) preceding the target/distractor words was easily handled by GPT-2's and LeoLM's content window size of 1024 tokens. In the next step, the models processed the input sequence and generated logits, which are raw scores associated with each token in the model's vocabulary. These logits 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, given the context provided by the input sequence. Then surprisal values for the tokens were the computed by applying the $-log_2$ function to the probability estimates. This measures how unexpected or surprising the token is, given the context provided by the preceding sequence. Finally, the tokens, i.e. the subword units created during tokenisation, are recombined to form the original stimuli based on the different word encodings for words that are preceded by a whitespace compared to words that are not preceded by a whitespace. Similarly, the surprisal values of the combined tokens are summed to obtain a single surprisal value for each word.

### 5.2.2 Results

Table 2 shows the descriptive statistics of the surprisal values computed by GPT-2 and LeoLM. Crucially, the average surprisal values computed with GPT-2 are higher for the distractor word across (almost all) conditions compared to the target word. 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. Firstly, this provides evidence that the main verb in Condition B was effectively manipulated in a way that resulted in higher average expectancy of the target word compared to the distractor word, while maintaining an intermediate level of plausibility. This result implies that the semantically attractive alternative in Condition B has been removed, such that the distractor word does not present a semantically attractive alternative to the target word in any condition. Secondly, the average surprisal values of Condition A and B are consistent with the cloze probabilities in the study of Aurnhammer et al. (2023), both indicating higher average expectancy for the target word compared to the distractor word. The surprisal values calculated by LeoLM follow the same pattern, except for Condition C. The expectancy of the distractor word in condition C is slightly higher than the expectancy of the target word for unknown reasons. On the one hand, this result seems reasonable, considering the higher average plausibility for the target word compared to the distractor word in Condition C, alongside the relatively strong negative correlation between the human-based plausibility judgements and the LeoLM-based surprisal values (see Table 3). One potential explanation, that was previously discussed in Chapter (add ref. 5.1), is that the rather unexpected distractor word in the context of the implausible Condition C renders some items more plausible compared to items in which the target word appears in the context of Condition C (see also Figure 3). On the other hand, these differences might not be crucial, as Conditions A and C were not changed for the current study and the main goal of reducing the expectancy of the distractor word in Condition B was achieved. Furthermore, the surprisal values computed with GPT-2 match the expectancy levels of the cloze probabilities in Conditions A and C, as reported by Aurnhammer et al. (2023).

As assumed, the average surprisal values of both models GPT-2 and LeoLM show a graded pattern across conditions for the target word. The pattern is reversed

$(C > B > A)$ compared to plausibility, as higher surprisal indicates lower expectancy, and vice versa. However, the average surprisal values calculated by GPT-2 for the distractor word do not align with 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: $A > B > C$. This result is consistent with the target and distractor word pattern for plausibility, but rather unexpected in terms of surprisal. Intuitively, one would also expect a lower expectancy (i.e. a higher average surprisal value) for the distractor word in Condition C compared to Conditions A and B. Again, there might be several reasons for these discrepancies, one of them being that in some cases the distractor word in combination with Condition C is slightly more plausible than Conditions A and B. Furthermore, the relatively high standard deviations for the distractor word indicate greater variability in the data. This implies that the distractor word is generally less predictable and therefore assessed less consistently by the different LMs across conditions. Partially, this may be due to the fact that LMs estimate only one surprisal value per item, while per-item plausibility represents the average plausibility based on the ratings from 10 participants, which makes it less susceptible to outliers and provides a more stable estimate of overall plausibility compared to the surprisal estimates. For example, although one (or more) participant may have rated Condition C with a number such as 6 or 7, indicating high plausibility, this single rating does not heavily influence average the plausibility of the item since it is likely that at least some of the other nine participants rate the same item as implausible.

Another difference between plausibility and surprisal is in the way they are measured. For this plausibility rating study final sentence continuations were included for consistency reasons since plausibility ratings were also collected during the self-paced reading study in which the continuation is crucial to capture spillover effects. Participants assessed the plausibility of an item in terms of the context paragraph and the final sentence, including the final sentence continuation, while LMs only process the context paragraph and the final sentence up to the target or distractor word, without taking the final sentence continuation into account. The collected plausibility ratings thus reflect the plausibility of the entire final sentence in the context of the preceding paragraph, while surprisal only reflects the expectancy of the target or distractor word in the given context. Whether the final sentence continuation is included or not, however, seems to influence especially the expectancy of the distractor word in the context of the entire final sentence. For example, the distractor word "*Gaul*" seems to have a different, possibly higher, expectancy in the context of the truncated final sentence (of Condition C) ("Dann füllte der Tierliebhaber den *Gaul*"/"Then filled the animal lover the *horse*") than in the context of the entire final sentence ("Dann füllte der Tierliebhaber den *Gaul* und darüber hinaus forderte er ihn auf, den Gaul in Ruhe zu lassen."/"Then filled the animal lover the *horse* and asked him to leave the horse alone."), due to the repetition of the distractor word combined with another, sometimes contradictory, action in the final sentence.

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. This may confirm the findings of Oh and Schuler (2023) that larger LMs tend to underpredict open-class words, to which the target and distractor words in this study belong. However, LeoLM's stronger (negative) correlation with human plausibility ratings (see Table 3) suggests that it may better capture the expectancy of the target and distractor words compared to GPT-2.

| | | **Plausibility** | | | **Surprisal (GPT-2)** | | | **Surprisal (LeoLM)** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cond. | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| | 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 |
| **Target** | 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 |
| | 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 |
| **Distractor** | 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 2: 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 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, they are assigned very low probabilities, and therefore very high surprisal values, which skew the distributions to the right. 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. Both GPT-2 and LeoLM have the highest peak, corresponding to the highest density of surprisal values, in Condition A. Most of the surprisal values computed by GPT-2 fall in the lower range of approximately 0 to 5 and gradually decrease afterwards, whereas virtually all surprisal values computed by LeoLM fall in this range. The generally high surprisal values observed for Condition A are consistent with the summary statistics in Table 2, confirming that both LMs, particularly LeoLM, predict low surprisal values, reflecting high expectancy, for the target words in Condition A. In contrast, the density curves for Condition B and especially for Condition C have less pronounced peaks and follow a more even distribution across a wider range. This shows that a part of the target words in Condition B and especially in Condition C were assigned higher surprisal values, reflecting lower expectancy levels, which increases the surprisal of the target words in these conditions on average. This implies that the target words in Conditions B and C are generally less predictable compared to the target words in Condition A. Consequently, the variability in the surprisal values is greater, as shown by the relatively flat distributions in Conditions B and C.

In contrast, the distributions of the surprisal values computed for the distractor word are more smooth and spread over a wider range compared to the distributions of the target word. 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 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. This is consistent with the previously discussed descriptive statistics, which show that the average expectancy of the distractor word is lower than the expectancy of the target word in all conditions except of Condition C in the case of LeoLM. Furthermore, the variability of the surprisal values appears to be relatively high across all conditions, due to the additionally rather low expectancy of the distractor word compared to the target word.

Table 3 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, yet rather moderate, 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 word GPT-2 surprisal ($r$ = -0.36). The negative sign indicates that as target word plausibility increases, target word surprisal decreases proportionally, reflecting higher expectancy due to the inverse relationship between surprisal and expectancy. This demonstrates that plausibility and expectancy, operationalised as surprisal, are to some extent proxies for each other: Events or statements that are plausible in a given context tend to be more expected and events or statements that are less plausible tend to be less expected, although it is possible for an event to be both unexpected and plausible. Furthermore, the correlations indicate that LeoLM surprisal aligns more closely with human plausibility judgments, suggesting that larger LMs capture human plausibility judgments (and probably reading times) better compared to smaller variants. As only distractor word surprisal is used as a predictor of reading times in the current study, which is unlikely to modulate reading times, no conclusions can be drawn as to which LM, GPT-2 or LeoLM, is more predictive of reading times. However, the higher correlation of target word surprisal with human target word plausibility suggests that LeoLM target word surprisal would be more predictive of RTs than GPT-2 target word surprisal. This would contradict the findings of Oh and Schuler (2022) and be instead in line with the findings of Goodkind and Bicknell (2018) and Wilcox et al. (2020) who found larger LMs with more parameters and better next-word prediction performance (i.e. lower perplexity) to be more predictive of human reading times.

There is a weak positive correlation between target word plausibility and LeoLM distractor surprisal, as both follow the pattern $A > B > C$. In contrast, there is virtually no linear relationship between distractor word plausibility and target 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 ($r$ = 0.28), which are also used as predictors for reading times, should ideally be closer to zero. However, the relationship is rather weak and should not be problematic in terms of multicollinearity.

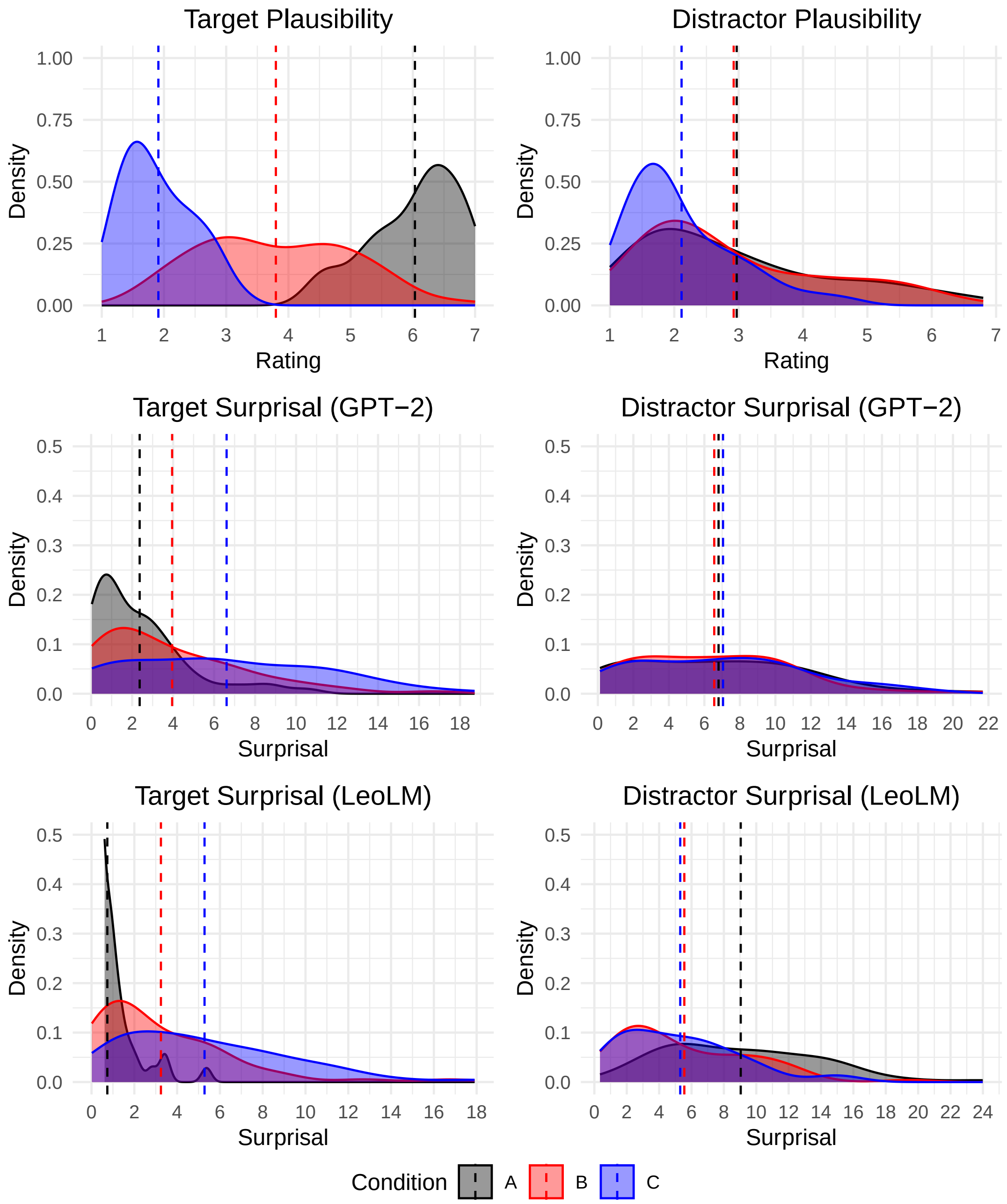

FIGURE 5: Densities for the results of the plausibility rating study that collected seven-point scale plausibility ratings and the surprisal values calculated by GPT-2 and LeoLM for the target and distractor words.

| | | **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 3: Correlations between plausibility ratings and (GPT-2 and LeoLM) surprisal of the target and distractor words.

In summary, the average surprisal values per condition have been shown to be lower for the target word than for the distractor word. Thus, the stimuli have been manipulated successfully in terms of plausibility and expectancy such that the distractor word expectancy is lower than the target word expectancy, while maintaining graded plausibility across conditions. The LeoLM surprisal values for Condition C are an exception, as the surprisal values for the distractor word are higher, although they are almost identical to the surprisal values for the target word. The average target word surprisal $(C > B > A)$ pattern is reversed compared to the average target word plausibility pattern, indicating that both target word plausibility and target word expectancy are high in Condition A, intermediate in Condition B and low in Condition C. In contrast, the patterns for the distractor word differ from the target word and with respect to the LM by which they were computed, as the lower plausibility and expectancy of the distractor word lead to greater variability in the predictions.

LeoLM's surprisal demonstrates a stronger correlation with human plausibility judgements compared to GPT-2's surprisal, which can be attributed to LeoLM's larger amount of training data and number of parameters. Consequently, it would not be surprising if LeoLM surprisal also proved to be a better predictor of reading times than GPT-2 surprisal. However, the correlation between plausibility and LeoLM surprisal is lower compared to the correlation between plausibility and expectancy operationalised as cloze probability that was observed in Aurnhammer et al. (2023). These differences may be caused by various factors, such as the relative stability of the average plausibility ratings and cloze probabilities per condition, which rely on per-item averages, contrasting with single surprisal estimates computed for each item. Additionally, the calculation of the surprisal values itself is not unproblematic as it relies on subword units created during tokenisation (Nair and Resnik, 2023).

Ultimately, these issues may not impact the results of the current study. In line with previous research (Rich and Harris, 2021), Aurnhammer et al. (2023) found no significant reading time modulations due to distractor word cloze probability despite the high distractor word expectancy in Condition B. This suggests that behavioural measures such as reading times might not be sensitive to unfulfilled expectations at all. Since the pre-studies have shown that the expectancy of the distractor word is low across conditions, including Condition B, distractor word surprisal should certainly not modulate reading times in the current study. In other words, given that the semantically attractive alternative in Condition B has been removed, no significant reading times modulations due to distractor word surprisal should be observed, even if reading times were to be sensitive to unfulfilled expectations. However, Aurnhammer et al. (2023) observed an early negativity (~250-400 ms) in the ERP signal for the unexpected target words in Condition B, which could be due to the unfulfilled expectation of the distractor word on a lexical level. To verify this, a future EEG study based on the current stimuli could test whether the early negativity disappears after the ambiguity in Condition B has been removed.

# Chapter 6

# Self-Paced Reading Study I

As Michaelov 2022 and 2023 showed surprisal can be a better predictor of EEG (not RTs?) in Aurnhammer et al RTs were not sensitive to expectancy, so here shouldnt be at all bc expectancy is lower in B, no diff almost if I use GPT-2 or LeoLM to calculate surprisal because distractor surprisal is not significant predictor anyways (and coefficients?)

## 6.1 Participants

## 6.2 Procedure

## 6.3 Analysis

## 6.4 Results

### 6.4.1 Single-trial Plausibility Ratings

### 6.4.2 Comprehension Questions

### 6.4.3 Reading Times

### 6.4.4 Model Comparison

## 6.5 Discussion

# Chapter 7

# Self-Paced Reading Study II

## 7.1 Participants

## 7.2 Procedure

## 7.3 Analysis

## 7.4 Results

### 7.4.1 Comprehension Questions

### 7.4.2 Reading Times

## 7.5 Discussion

# Chapter 8

# General Discussion

# Chapter 9

# Conclusion

# *Acknowledgements*

lalal

# Bibliography

Arehalli, S., Dillon, B., and Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processng difficulty from syntactic ambiguities. In Fokkens, A. and Srikumar, V., editors, *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 301–313.

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. 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.

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., 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.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

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.

Ferreira, F. and Tanenhaus, M. K. (2007). Introduction to the special issue on language-vision interactions. *Journal of Memory and Language*, 57(3):455–459.

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 33(4):465–495.

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.

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.

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.

Hahn, M., Futrell, R., Levy, R., and Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 119, pages 1–8.

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.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diplomarbeit, Technische Universität München*.

Keller, F. (2010). Cognitively Plausible Models of Human Language Processing. In Hajič, J., Carberry, S., Clark, S., and Nivre, J., editors, *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67.

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. (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).

Merkx, D. and Frank, S. L. (2021). Comparing Transformers and RNNs on predicting human sentence processing data. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

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.

Misra, K., Ettinger, A., and Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP*, pages 4625–4635.

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.

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.

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*, volume 1, 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.

Parker, D., Shvartsman, M., and Van Dyke, J. A. (2017). *Language processing and disorders*, chapter The cue-based retrieval theory of sentence comprehension: New findings and new challenges, pages 121–144. Cambridge Scholars Publishing.

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

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

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

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

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 Kevy, 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.

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.

Van Petten and B. J. Luka, C. (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.

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

# Post Hoc Analysis