Proposal for a Master's thesis

The predictive power of language model surprisal: Explaining N400 & P600 effects

Benedict Schneider

June 15, 2023

Abstract

In the research of online language comprehension, information-theoretic surprisal has been introduced as a high-level indicator of cognitive processing effort. In recent years, it has most commonly been operationalized through language models that compute the probabilities of upcoming words given the preceding context. This operationalization has also been used to predict the amplitude of the N400, a component of the event-related potential (ERP) that is sensitive to experimental manipulations of expectancy. However, the N400 has also been modulated by contextual semantic relatedness, and thus the question arises if and to what extent next-word prediction language models can capture this differentiation. Moreover, the P600 as another prominent ERP-component has recently been shown to be modulated by expectancy in a graded manner. Thus, it should provide a strong correlate to language model surprisal, yet to our knowledge no research has been conducted to investigate this hypothesis. Hence, the goal of this thesis will be to examine several EEG-studies in order to determine which distinctive N400 and P600 effects can be explained by surprisal estimates derived from different language model architectures (n-gram, RNN, transformer). The resulting implications for the explanatory power that language model surprisal can provide will be discussed.

1 Introduction

Within the field of computational psycholinguistics, increasingly capable language models have been deployed to predict the amplitude of the N400, an ERP-component

that has amongst other factors been linked to the expectancy of upcoming words (Kutas & Federmeier, 2011). While successful at this endeavour, expectancy itself is a very generalistic concept, whose constituting factors may be captured to a greater or lesser extent by its individual operationalizations.

Two important findings from recent studies indicate that the N400 is more sensitive to semantic association than it is to expectancy, while the P600, another prominent ERP-component, is sensitive to expectancy and plausibility but not to association (Aurnhammer et al., 2023; Aurnhammer et al., 2021; Delogu et al., 2019, 2021). While the first finding stands in an apparent contrast to recent work that operationalizes expectancy through GPT-3 surprisal (Michaelov et al., 2023), the P600 has not been well researched in the context of language models yet. Thus, the twofold task of this thesis will be to evaluate on the four above mentioned studies to which extent language model surprisal can predict N400 effects that have been differentially elicited by a manipulation of association and how strongly it correlates with the P600.

The proposal will have the following structure: §2 provides an overview of the related concepts and studies, while §3 lays out the research questions; the data and intended methods will be discussed in §4; the subtasks and a time frame for the thesis will be schemed in §5 whereas §6 concludes the proposal.

2 Background and related work

This section presents the theoretical concepts of expectancy and surprisal, experimental and model-derived operationalizations of these concepts as well as related work on neurophysiological research.

2.1 Surprisal Theory

Human sentence comprehension is driven by incremental expectations about upcoming words. In order to formalize this notion on a computational level (Marr, 2010), the concept of *surprisal* (1), originating from Information Theory (Shannon, 1948), has been introduced. Hereafter, the cognitive effort required to process a word *w* in a sentence at position *t+1* is proportional to its surprisal (Hale, 2001; Levy, 2008).

$$Surprisal(w_{t+1}) = -log_2 p(w_{t+1}|w_{1...t}) \tag{1}$$

Importantly, this formalization remains agnostic towards the exact origins and mechanisms that can lead to processing difficulty, providing only a generalistic measure of expectancy. Due to its nature as a linking theory, surprisal doesn't entail specific assumptions about the algorithmic or implementational level (Marr, 2010).

While the amount of information a specific word conveys in a specific context can be *computed* by probabilistic models, the amount of cognitive effort required to process this word may be *observed* through (neuro)behavioural methods (Frank et al., 2015). Specifically, a positive correlation between reading times and word surprisal has been established in the past (Fernandez Monsalve et al., 2012; Fossum & Levy, 2012; Mitchell et al., 2010; Roark et al., 2009; Smith & Levy, 2008)[1]. While reading times provide an overall index of word-by-word processing effort, neurophysiological methods aim to reveal the mechanisms that are underlying and driving this effort.

Event-related potentials (ERPs) provide a multidimensional window into language comprehension with an excellent temporal resolution and have been widely used to study the neural computations during sentence comprehension. Most saliently, the N400 and P600 components have been identified and shown to be differentially sensitive to a number of experimental manipulations related to the general concept of expectancy.

2.2 Cloze probability, association & plausibility

Cloze probability (oftentimes abbreviated as *cloze*) has been originally introduced by Taylor (1953) as result of a normative sentence-completion study. Hereafter, the cloze probability of a word equals the percentage of people who continued a sentence fragment with it. As such, cloze can be viewed as a more informal measure of word probability gathered from human judgements rather than from probabilistic models (Frank et al., 2015). While providing a good estimate about the more likely range of completions that humans may expect in a given context, one potential weakness is a lack of differentiation in the lower probability spectrum of possible continuations. That is, reasonable but less likely continuations may result in zero cloze probability.

Importantly, while being correlated, the *semantic association* of a target word with the prior context is distinct from but often confounded with expectancy (as pointed out by Aurnhammer et al., 2021). In norming studies, this metric may be collected by participants rating the strength of relatedness between a target and usually a single content word from the preceding context (Aurnhammer et al., 2021; Delogu et al., 2019, 2021) or by participants producing related words under time constraints (Battig & Montague, 1969; Keppel & Postman, 1970; Kutas, 1993).

Extending beyond linguistic information, *plausibility* reflects to which extent an utterance aligns with our knowledge of the world (Brouwer et al., 2021). Usually, experimental stimuli are assessed in their entirety (respectively until including the target word) in a rating task during a pre-study (Aurnhammer et al., 2023; Delogu et al., 2019, 2021; Michaelov et al., 2023).

[1] These references stem from Frank et al. (2015, p.2)

Grammar-based	Frank et al. (2015) Parviz et al. (2012) Mitchell et al. (2010) Hale (2001)	
n-gram	Degaetano-Ortlieb and Teich (2022) Goodkind and Bicknell (2018) Frank (2017) Frank and Willems (2017) Frank et al. (2015)	Mitchell et al. (2010)
RNN	Slaats and Martin (2023) Michaelov et al. (2022) Merkx and Frank (2021) Michaelov et al. (2021) Michaelov and Bergen (2020)	Aurnhammer and Frank (2019a) Frank and Hoeks (2019) Aurnhammer and Frank (2019b) Goodkind and Bicknell (2018) Frank et al. (2015)
Transformer	Michaelov et al. (2023) Michaelov et al. (2022) Oh and Schuler (2022) Merkx and Frank (2021) Michaelov et al. (2021)	

Table 1: A selection of studies that collected surprisal estimates from different language model architectures (sorted by model type and publication date).

A challenging problem arises from the correlation between these metrics and their relation to the more generalistic notion of expectancy. While cloze probability can be viewed as a human-based operationalization of expectancy (that stands conceptually near to surprisal), it seems intuitively sensible that these probabilities are influenced by the distinct concepts of both semantic association and plausibility. That is, words that are strongly associated with each other are also more likely to appear in the same contexts (Ettinger et al., 2016). Moreover, under rational theories of communication (Grice, 1967) we expect plausible continuations to be more likely than implausible ones.

2.3 Language model operationalizations

Since surprisal is formalized in terms of logarithmic probabilities, statistical language models have been extensively deployed to derive these word-by-word probabilites in the past (see Table 1). Note, that the unit in question doesn't need to be restricted to words. For example, surprisal can and has also been computed for POS-tags (Frank et al., 2015) and CCG-tags (Arehalli et al., 2022) to operationalize a notion of syntactic versus lexical surprisal.

Importantly, the surprisal estimates from these studies stem from a range of dif-

ferent model-architectures. Originally, Hale (2001) used a *probabilistic Early parser*, relying on a *phrase structure grammar (PSG)*. But, while having the advantage of capturing hierarchical structure in language, the limitations of grammar-based approaches in terms of scalability have led to research turning to architectures that rely solely on distributional information.

As such, *n-gram* models are contextually restricted to the *n-1* preceding words when estimating word surprisal. While these estimates rely straightforwardly on word frequencies gathered from corpora, due to data sparseness the order of *n* is usually ceiling at 5, even after including smoothing techniques, meaning that only the 4 preceding words are taken into account for the estimate. Though this renders n-gram models cognitively implausible, they have shown to be remarkably accurate (Frank et al., 2015).

In terms of cognitive plausibility, the *Recurrent Neural Network* (RNN; Elman, 1990) offers a considerable improvement. As a subtype of this model class, the *Simple Recurrent Network* (SRN; Rumelhart et al., 1986) has been succesfully used within psycholinguistic research to model aspects of incremental language comprehension (see Brouwer et al., 2017 as an example). These models take into account a wider context and are (in contrast to simpler neural architectures) capable of capturing the sequential nature of language. Still limited though when it comes to long sequences due to the vanishing gradient problem (Hochreiter, 1998), the *Gated Recurrent Unit* (GRU; Bahdanau et al., 2014) and the *Long-Short-Term-Memory* network (LSTM; Hochreiter and Schmidhuber, 1997) have been proposed to address this shortcoming.

In recent years, technical progress in the field of NLP has led to the emergence of *transformer-based* models (Vaswani et al., 2017). Increasingly popular and successful when it comes to NLP tasks, their cognitive plausibility is more or less limited depending on their specific architecture. In contrast to all types of RNNs, which process words in a strictly incremental manner, transformer models exhibit self-attention layers that allow them to selectively attend to previous (or all) parts of the input. While this mechanism seems cognitively implausible at first, Merkx and Frank (2021) discuss how it could relate to cue-based retrieval theories rather than recurrent ones.

Though at least some of the model architectures clearly lack cognitive plausibility, an important observation is that there seems to be a tendency that more sophisticated models perform closer to human data (Goodkind & Bicknell, 2018; Merkx & Frank, 2021; Michaelov et al., 2021). Nevertheless, this conclusion does not universally hold (Frank et al., 2015) and specifically for reading times, raising model complexity within transformer-based models appears to have a reverse effect (Oh & Schuler, 2022).

While the operationalization of surprisal through language models can be viewed

to be conceptually related the metric of cloze probability[2], there have also been a number of different methods for a model-derived quantification of semantic association in psycholinguistically motivated studies (see Table 2).

The most popular existing methods can be divided into two distinctive approaches, based on either *counts* or *predictions* (Mandera et al., 2017). For the former, two prominent examples are the *Latent Semantic Analysis* (LSA; Landauer and Dumais, 1997) and *Global Vectors for Word Representation* (GloVe; Pennington et al., 2014), the latter are most prominently represented by *Word2Vec* (Mikolov et al., 2013) and its more recent instantiation of *fastText* (Bojanowski et al., 2017). In favor of prediction based methods, arguments have been made for their stronger psychological plausibility (Mandera et al., 2017) and overall better performance (Baroni et al., 2014; Nieuwland et al., 2020).

Regardless of the underlying approach, all methods derive high-dimensional vector representations of words, based on their co-occurence patterns in large text corpora. This form of representation brings along the useful property of comparability on a mathematical level. That is, the semantic relatedness between two words is quantified by the distance between their vector representation, usually measured by cosine similarity. When comparing the relatedness between a single word and its preceding context, the context is usually represented by the sum of the vector representations that it is spanning over. Usually, this sum representation is averaged, though Michaelov et al. (2023) point out that with respect to cosine similarity the magnitude is irrelevant and hence including this step should yield the same result as taking into account only the sum (Frank & Willems, 2017). In some cases, only content words are taken into account as context (Frank, 2017; Frank & Willems, 2017).

Since plausibility builds upon world knowledge outside of the linguistic domain, models that are solely trained on language do not naturally capture this notion. While surprisal in human listeners is characterized by expectations from both world knowledge and linguistic experience, surprisal in language models originates only from the latter source, that is, the linguistic input they have been trained on. This distinction has lead to the definition of *comprehension-centric surprisal* (*cc-surprisal*; Venhuizen et al., 2019). In their model of comprehension, world knowledge is introduced into the model within the framework of *Distributional Formal Semantics* (*DFS*; Venhuizen et al., 2022).

[2]Crucially, with the important difference that language model surprisal solely reflects distributional information of language, while human judgements may also be modulated by numerous factors such as world knowledge.

LSA	Smolka and Eulitz (2018) Frank (2017) Van Petten (2014) Parviz et al. (2012) Mitchell et al. (2010)	Pynte et al. (2008) Chwilla and Kolk (2002)
Word2Vec/fastText	Michaelov et al. (2023) Nieuwland et al. (2020) Frank and Willems (2017) Frank (2017) Ettinger et al. (2016)	
GloVe	Michaelov et al. (2023)	
Other	Van Petten (2014)	

Table 2: A selection of studies that collected semantic association estimates from different distributional methods (sorted by method and publication date).

2.4 ERP-components

While behavioural measures such as reading times provide a solid estimate of overall processing effort and have therefore been succesfully linked to surprisal, neurophysiological research seeks to reveal the sub-processes underlying comprehension and to answer how those processes interactively unfold in real-time. In this line of research, two components in the *event-related-potential* (*ERP*) signal have taken a prominent role: the *N400* and the *P600*.

The N400 is a negative voltage deflection peaking around 400 ms after stimulus onset and has for many years been taken to reflect the process of semantic integration, while the P600, a positive deflection emerging from around 500-600 ms, has been linked to syntactic integration. However, several findings have challenged this view, leading to more complex multi-stream accounts, which in turn have struggled to account for all of the data (see Brouwer et al., 2012 for a review). Utilizing their neurocomputational model, Brouwer et al. (2017) demonstrate a single-stream account of *retrieval* and *integration*, that is able to account for all of the previous findings. Hereafter, the N400 indexes the retrieval of word meaning from semantic memory, while the P600 reflects the integration of this meaning into an interpretation of the unfolding utterance.

In a series of following studies (which will from now on be referred to as **RI-studies**), experimental manipulations of expectancy, association and plausibility have led to results that provide further evidence for the *Retrieval-Integration* (*RI*) account (see Table 3). It has been observed that manipulations of expectancy may result in a biphasic *N400* and *P600*, although these effects may sometimes be obscured by component overlap (Van Petten and Luka, 2012; see Brouwer et al., 2017 for a discussion). Consistent with this finding, Aurnhammer et al. (2021), henceforth **ADSBC21**, have shown both the N400 and P600 to be sensitive to manipulations

	DBC19[1]		DBC21[2]		ADSBC21[3]		ADBC23[4]	
	N4	P6	N4	P6	N4	P6	N4	P6
Association	✓	X	✓	X	✓	X	–	–
Plausibility	X	?	X	✓	–	–	X	✓
Cloze	X	?	X	✓	✓	✓	X	✓

[1] Delogu et al. (2019)
[2] Delogu et al. (2021)
[3] Aurnhammer et al. (2021)
[4] Aurnhammer et al. (2023)
– not manipulated
✓ effect found
X no effect found
? conflicting results

Table 3: Overview of the N400 & P600 modulation pattern found in the four RI-studies.

of cloze probability. In their context manipulation design, an intervening adverbial clause either maintained or dissolved association to a preceding context, while the target that followed the intervening clause was either expected or unexpected[3] given the selectional restrictions of the main clause. While expectancy modulated both the N400 and P600, association only influenced the N400. As pointed out by the authors, the influence of expectancy on both components appears valid under the retrieval-integration account, since a higher expectancy of a word may facilitate its retrieval from long-term memory (N400) as well as ease the effort of updating the utterance meaning representation (P600), whereas differences in association should only affect retrieval effort but not integration effort.

Also finding an N400-effect of association, Delogu et al. (2019), henceforth **DBC19**, additionally manipulated plausibility, expecting (in line with the retrieval-integration account) to elicit a P600 effect for implausible versus plausible target words, since the former would be more difficult to integrate into the unfolding utterance representation. While this was indeed the case for the *event-related & implausible*[4] condition, no P600 effect was observed in the *event-unrelated & implausible* condition (both relative to the plausible baseline). As hypothesized and confirmed by the authors in a follow-up study, the apparent absence of a P600 effect was due to spatiotemporal-overlap, i.e. the N400 and P600 overlapping both in space in time due to the additive nature of the waveform-based component structure (see

[3] operationalized by cloze probability
[4] Event-related being equivalent to associated here

Brouwer and Crocker, 2017 for a discussion and Brouwer et al., 2021; Delogu et al., 2021 for empirical evidence). Interestingly, cloze probability didn't pattern with the N400 in this study. Rather, it only patterned with the P600, in that average cloze in the two implausible conditions (event-related and event-unrelated) was significantly lower than in the baseline condition.

Delogu et al. (2021), henceforth **DBC21**, replicated these results, providing further evidence for a strong link between semantic association and the N400. Crucially, applying the technique of *regression-based ERP* (rERP) estimation (Brouwer et al., 2021; Smith & Kutas, 2015), the authors revealed an increase in P600 amplitude in their event-unrelated implausible condition, that had been attenuated by a preceding increase in N400 amplitude. Moreover, the rERP analysis showed, that using *association & plausibility* versus *cloze & plausibility* as continuous predictors resulted in a closer fit to the EEG-data, rendering association a stronger predictor of N400 amplitude than cloze.

Further, Aurnhammer et al. (2023), henceforth **ADBC23**, established that the P600 continuously indexes integration effort: compared to a plausible baseline, increasingly implausible conditions led to increasingly positive P600 amplitudes. Since lexical association was high across conditions, achieved by lexical repetition of the target word, the difference in plausibility did not elicit any N400 modulations. While the mean target cloze probability didn't explicitly enter the analyses, it was considerably lower in the implausible conditions while being highly correlated with plausibility at the same time.

In sum, these studies provide evidence that both N400 and P600 are modulated by expectancy since it respectively facilitates retrieval and integration. For the N400 it has been found that this influence of expectancy can be overridden by association, while this is not the case for the P600.

3 Research questions

The overarching goal of the thesis will be to gain further insights into the power of language model surprisal in explaining the ERP profile of human language comprehension. A number of studies have found language model surprisal to be a good predictor of the N400 amplitude (Frank et al., 2015; Frank & Willems, 2017; Merkx & Frank, 2021; Michaelov et al., 2023; Michaelov & Bergen, 2020; Michaelov et al., 2022; Parviz et al., 2012). In contrast, the P600 has not been studied in this context yet.

One strength of the four RI-studies is that they dissociate the influence of association, plausibility and expectancy and show how these manipulations lead to either isolated or bi-phasic N400 and P600 effects. More precisely, an isolated N400 effect of association was observed in DBC21 and ADSBC21, whereas an isolated P600 ef-

Stimulus	A	E	Con
Yesterday sharpened the lumberjack, before he the wood stacked, the *axe*...	+	+	A
Yesterday sharpened the lumberjack, before he the movie watched, the *axe*...	-	+	B
Yesterday ate the lumberjack, before he the wood stacked, the *axe*...	+	-	C
Yesterday ate the lumberjack, before he the movie watched, the *axe*...	-	-	D

Table 4: **RQ1 (a)**: showing the manipulations of Association (A) and Expectancy (E) in the conditions (Con) of ADSBC21.

fect of plausibility was elicited in DBC19, DBC21 and ADBC23. Moreover, DBC19 and DBC21 demonstrated how a simultaneous manipulation of association and plausibility can lead to a bi-phasic ERP profile, where the association-driven N400 effect is spatiotemporally concealing the plausibility-driven P600. The influence of expectancy was inconsistent in the N400 but consistent in the P600 window.

The results of the RI-studies indicate that it is not expectancy alone that modulates the N400. In fact, semantic association may be able to overwrite its influence. This is an apparent contradiction to the results of Michaelov et al. (2023), who found GPT3-surprisal to be a stronger predictor than two semantic distance metrics (GoVe and fastText). This leads to the first research question of the thesis:

1. Can language model surprisal capture N400 effects that have been modulated distinctively by expectancy or association?

As cloze probability and association are usually correlated, an important task is to assess language model surprisal on pairs of experimental conditions that fall under one of two cases:

(a) Keeping expectancy constant, a difference in association elicited an N400 effect.

(b) Keeping association constant, a difference in expectancy did *not* elicit an N400 effect.

For (a), table 4 shows the relevant conditions from ADSBC21. For (b), table 5 shows the relevant conditions from DBC19 and DBC21. Assuming that language model surprisal does not capture semantic relatedness, the expected outcome will be that language model surprisal (henceforth **LM** surprisal) falsely predicts no N400 effect in (a) and that it falsely predicts an N400 effect in (b).

Another finding from ADSBC21 is that the P600 is sensitive to expectancy but crucially not to association. Thus, this leads to the second research question:

	Stimulus	A	E	Con
DBC19	John entered the restaurant. Before long, he opened the *menu*...	+	+	A
	John left the restaurant. Before long, he opened the *menu*...	+	-	B
DBC21	John left the restaurant. Before long, he opened the *umbrella*...	-	+	B
	John entered the restaurant. Before long, he opened the *umbrella*...	-	-	C

Table 5: **RQ1 (b)**: showing the manipulations of Association (A) and Expectancy (E) in the conditions (Con) of DBC19 and DBC21.

2. Does language model surprisal provide a strong predictor of the P600 amplitude?

It has to be noted that in all four RI-studies the target word that elicited the P600 was a less plausible or implausible continuation given the context. While it is possible, that a rather unexpected continuation may be a plausible one (e.g. "She likes to eat her pizza with *artichokes*."), the reverse case seems unlikely from a communicative point of view. That is, implausible continuations should always simultaneously be unexpected. Considering that the language model which is generating the surprisal estimates was trained on a naturalistic corpus (rather than carefully manipulated experimental stimuli), the number of implausible continuations it was trained on is most likely small. Therefore, the expected outcome with respect to the second research question is an overall strong correlation. That being said, it would be conceivable that LM surprisal might overestimate the plausibility-driven P600 effect and not capture its gradedness.

As raised by Michaelov et al. (2023), another point to consider is how different language model architectures may yield a better or worse fit to the N400 and P600 window respectively. The authors cited a related study (Michaelov & Bergen, 2020), which found a better correlation of RNN-based estimates with post-N400 positivites than with the N400. Therefore, surprisal estimates of different architectures will be considered for both research questions.

Finally, an important observation would be to see how model-derived metrics of association (as described in section 2) correlate with the human-derived association ratings from the RI-studies and if the model-derived metrics are predicting the respective N400 and crucially not predicting the P600 effects. But, depending on the progress of the project (as laid out in section 5) this may need to be evaluated in future work.

	DBC19	DBC21	ADSBC21	ADBC23
Association	✓	✓	✓	–
Plausibility	✓	✓	–	✓
Cloze	✓	✓	✓	✓
#Stimuli	90	90	120	60
#Conditions	3	3	4	3

Table 6: Overview of the stimuli and ratings of the four RI-studies.

4 Data and methodology

With the research questions laid out, this section offers a closer look into the foundational data as well as the (computational) methods that will be used to answer them.

4.1 Experimental stimuli and ratings

In order to address the questions, the thesis seeks to specifically evaluate the four RI-studies, that have been conducted in German. Depending on which factors were manipulated, a combination of cloze, association and plausibility ratings for the target words have been collected during pre-studies (see Table 6). Having these ratings available alongside the stimuli, i.e. on a by-stimulus basis, the goal will be to evaluate how they correlate with the surprisal estimates gathered from different language models. In particular, a correlation between lm surprisal and plausibility would be a first indicator that lm surprisal might be a good predictor of the P600 amplitude (RQ2). Alongside assessing overall correlations, analyses will be conducted for specific sets of conditions. From RQ1, it is expected to observe the following:

1. There is no systematic difference for lm surprisal estimates comparing conditions A versus B and C versus D in ADSBC21.
2. Lm surprisal estimates will be systematically higher in condition B versus A in DBC19 and C versus B in DBC21.

4.2 Language model architectures

Based on the literature as reviewed in section 2, different language model architectures should be considered for the estimation of surprisal. Therefore, at least one instantiation of the three most prominent architectures used for modeling the ERP

profile will be taken into account respectively: n-gram, RNN and transformer-based models.

The SRILM software (Stolcke, 2004) will be used to train a 5-gram model. If it is implementable, modified Kneser-Ney smoothing will be applied (Chen & Goodman, 1996). Furthermore, the PyTorch library (Paszke et al., 2019) in Python (Van Rossum & Drake, 2009) will be applied to train a standard RNN. While it is also conceivable to train a more complex GRU or LSTM here, the results of Aurnhammer and Frank (2019b) indicate that this wouldn't necessarily add value to the ERP modeling task. Finally, as a representative of transformer-based models, the German version of GPT-2 (Schweter, 2020) will be deployed via the HuggingFace interface (Wolf et al., 2020).

While training set size has varied in previous studies, an important point with respect to model comparability is that ideally the same training set should be used. Since transformer-based models are usually trained on enormous data sets that may not easily be recreated,[5] a challenge will be to aim for a training set for the n-gram model and the RNN , that is comparable to the training set of the available German GPT-2 version.[6] Another challenge arises from the internal representation of words in GPT-2 in terms of Byte-Pair-Encoding (BPE; Sennrich et al., 2016). That is, with the goal of a more parsimonious vocabulary, words are often splitted into subwords according to an algorithm during training. Resulting from this process, the surprisal estimates for some target words may need to be composed from the surprisal of the respective subwords. The usual workaround here is to simply add the subword surprisals, but this leads to an overestimation. To circumvent this problem, Michaelov et al. (2023) use only items containing target words that were not splitted, i.e. present as a vocabulary unit within the model. It needs to be verified how severely this problem affects the stimuli of the four RI-studies. Consequences with respect to the analyses need to be discussed.

4.3 The rERP framework

An essential part of the analyses of most of the RI-studies is the application of the rERP framework (Brouwer et al., 2021; Smith & Kutas, 2015). Not only were Brouwer et al. (2021) able to reveal a hidden P600 effect due to this method, but it also allows to isolate the individual contributions of experimentally manipulated factors to the observed ERP profile. In its essence, "the core idea is to replace each individual voltage measurement in the ERP data—each observed voltage scalar—with a voltage estimate from a linear regression model that optimally combines the manipulated variables to explain the variance in the signal." (Brouwer et al., 2021, p. 976). Moreover, in the re-estimated signal the contributions of single variables can become visible by keeping

[5] Note, that Merkx and Frank (2021) trained their own transformer in PyTorch.

[6] see https://github.com/stefan-it/german-gpt2

the other variables constant. With the EEG-data available, the goal for this thesis is to re-estimate the amplitudes from the RI-studies, crucially using the surprisal estimates from the previously trained language models as predictor. Methodologically, the aim is to follow ADBC23 who used an implementation of this approach in Julia (Bezanson et al., 2017).

5 Work plan

The flowchart presented in Figure 1 provides an overview of the relevant steps that will be taken to answer the research questions. The steps are divided into three phases: model training, data collection and evaluation.

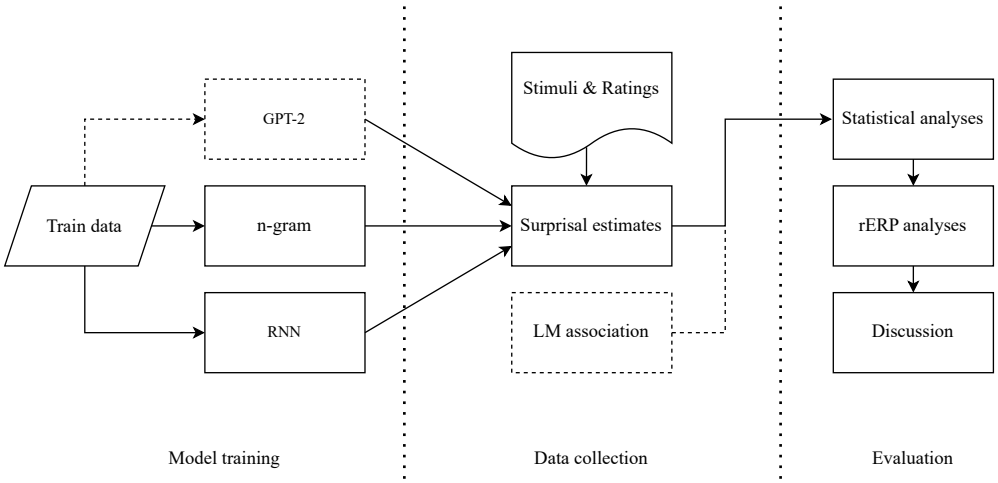

Figure 1: Flowchart presenting the three phases and their respective sub-processes of the thesis. Dashed lines index optional steps.

The project needs to be completed within three months, that is until the 30th of September 2023. Depending on the progress, it is conceivable to extend the project with analyses regarding model-derived association estimates and a transformer model that is trained from scratch, instead of using the pre-trained GPT-2 model (Schweter, 2020). These two optional steps are indexed by dashed boxes and lines in the flowchart.

6 Conclusion

Utilizing language model surprisal to predict the ERP-signal, the N400 has received much attention in the past years. As a very recent study, Michaelov et al. (2023) have found lm surprisal to be a stronger predictor for this component than model-derived semantic association metrics. This stands in contrast to the results of studies showing association as the driving factor (Aurnhammer et al., 2021; Delogu et al., 2019, 2021). Since these studies offer the advantage of modulating either association or expectancy while respectively keeping the other factor constant, this thesis investigates the predictions of lm surprisal derived from different model architectures in these cases. Moreover, Aurnhammer et al. (2021) established that the P600 is modulated by expectancy alone and Aurnhammer et al. (2023) elicited an isolated P600 effect graded for plausibility. Thus, the P600 is expected to strongly correlate with lm surprisal, and this constitutes the second hypothesis that the thesis addresses. Overall, considering surprisal as a "causal bottleneck" in the comprehension process (Levy, 2008), the thesis seeks to shed light on the capabilities and limitiations of language models in predicting the ERP-profile.

Bibliography

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 301–313. https://aclanthology.org/2022.conll-1.20

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*. https://doi.org/10.1111/psyp.14302

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), 1–31. https://doi.org/10.1371/journal.pone.0257430

Aurnhammer, C., & Frank, S. (2019a). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 112–118). Cognitive Science Society. https://scholar.google.de/citations?view_op=view_citation&hl=de&user=SPeiq88AAAAJ&citation_for_view=SPeiq88AAAAJ:9yKSN-GCB0IC

Aurnhammer, C., & Frank, S. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv*, *1409*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, *80*, 1–46.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information.

Brouwer, H., Crocker, M., Venhuizen, N., & Hoeks, J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/https://doi.org/10.1111/cogs.12461

Brouwer, H., & Crocker, M. W. (2017). On the Proper Treatment of the N400 and P600 in Language Comprehension. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01327

Brouwer, H., Delogu, F., & Crocker, M. (2021). Splitting Event-Related Potentials: Modeling Latent Components using Regression-based Waveform Estimation. *European Journal of Neuroscience*, *53*, 974–995. https://doi.org/10.1111/ejn.14961

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/https://doi.org/10.1016/j.brainres.2012.01.055

Chen, S. F., & Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. *34th Annual Meeting of the Association for Computational Linguistics*, 310–318. https://doi.org/10.3115/981863.981904

Chwilla, D., & Kolk, H. (2002). Three-step priming in lexical decision. *Memory & cognition*, *30*, 217–25. https://doi.org/10.3758/BF03195282

Degaetano-Ortlieb, S., & Teich, E. (2022). *Corpus Linguistics and Linguistic Theory*, *18*(1), 175–207. https://doi.org/doi:10.1515/cllt-2018-0088

Delogu, F., Brouwer, H., & Crocker, M. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569. https://doi.org/https://doi.org/10.1016/j.bandc.2019.05.007

Delogu, F., Brouwer, H., & Crocker, M. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, *1766*, 147514. https://doi.org/10.1016/j.brainres.2021.147514

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/https://doi.org/10.1016/0364-0213(90)90002-E

Ettinger, A., Feldman, N. H., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. *Cognitive Science*.

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

Fossum, V., & Levy, R. (2012). Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, 61–69. https://aclanthology.org/W12-1706

Frank, S., & Hoeks, J. C. J. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. *Annual Meeting of the Cognitive Science Society*.

Frank, S. (2017). Word Embedding Distance Does not Predict Word Reading Time.

Frank, S., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S., & Willems, R. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18. https://doi.org/10.18653/v1/W18-0102

Grice, H. P. (1967). Logic and Conversation. In P. Grice (Ed.), *Studies in the way of words* (pp. 41–58). Harvard University Press.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings of NAACL 2001*, *2*. https://doi.org/10.3115/1073336.1073357

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*, 107–116. https://doi.org/10.1142/S0218488598000094

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, *9*, 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

Keppel, G., & Postman, L. (1970). *Norms of Word Association, Edited by Leo Postman and Geoffrey Keppel*. https://books.google.de/books?id=1RUkcgAACAAJ

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., & 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*, 621–47.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*, 211–240.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/https://doi.org/10.1016/j.cognition.2007.05.006

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/https://doi.org/10.1016/j.jml.2016.04.001

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Merkx, D., & Frank, S. (2021). Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Michaelov, J., Bardolph, M., Coulson, S., & Bergen, B. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude?

Michaelov, J., Bardolph, M., Van Petten, C., Bergen, B., & Coulson, S. (2023). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 1–71. https://doi.org/10.1162/nol_a_00105

Michaelov, J., & Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? *Proceedings of the 24th Conference on Computational Natural Language Learning*, 652–663. https://doi.org/10.18653/v1/2020.conll-1.53

Michaelov, J., Bergen, B., & Coulson, S. (2022). So Cloze yet so far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. *IEEE Transactions on Cognitive and Developmental Systems*, *PP*. https://doi.org/10.1109/TCDS.2022.3176783

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th annual meeting of the association for computational linguistics*, 196–206.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Oh, B.-D., & Schuler, W. (2022). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2012). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision research*, *48*, 2172–83. https://doi.org/10.1016/j.visres.2008.02.004

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving Lexical and Syntactic Expectation-Based Measures for Psycholinguistic Modeling via Incremental Top-down Parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, 324–333.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation.

Schweter, S. (2020). *German GPT-2 model* (Version 1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4275046

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slaats, S., & Martin, A. E. (2023). What's surprising about surprisal. https://doi.org/10.31234/osf.io/7pvau

Smith, N., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*. https://doi.org/10.1111/psyp.12317

Smith, N., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*(30).

Smolka, E., & Eulitz, C. (2018). Psycholinguistic measures for German verb pairs: Semantic transparency, semantic relatedness, verb family size, and age of reading acquisition. *Behavior Research Methods*, *50*, 1540–1562. https://doi.org/10.3758/s13428-018-1052-5

Stolcke, A. (2004). Srilm — An Extensible Language Modeling Toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, *2*.

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

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International*

Journal of Psychophysiology, *94*, 407–419. https://doi.org/10.1016/j.ijpsycho.2014.10.012

Van Petten, C., & Luka, B. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *83 2*, 176–90.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need.

Venhuizen, N., Crocker, M., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*, 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Venhuizen, N., Hendriks, P., Crocker, M., & Brouwer, H. (2022). Distributional formal semantics [Special Issue: Selected Papers from WoLLIC 2019, the 26th Workshop on Logic, Language, Information and Computation]. *Information and Computation*, *287*, 104763. https://doi.org/https://doi.org/10.1016/j.ic.2021.104763

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing.