Tracking lexical and semantic prediction error underlying the N400 using artificial neural network models of sentence processing

Alessandro Lopopolo
Milena Rabovsky

Abstract

Recent research has shown that the internal dynamics of an artificial neural network model of sentence comprehension displayed a similar pattern to the amplitude of the N400 in several conditions known to modulate this event-related potential. These results led Rabovsky, Hansen, and McClelland (2018) to suggest that the N400 might reflect change in an implicit predictive representation of meaning corresponding to semantic prediction error. This explanation stands as an alternative to the hypothesis that the N400 reflects lexical-prediction error as estimated by word Surprisal (Frank, Otten, Galli, & Vigliocco, 2015). In the present study, we directly model the amplitude of the N400 elicited during naturalistic sentence processing by using as predictor the update of the distributed representation of sentence meaning generated by a Sentence Gestalt model (McClelland, St. John, & Taraban, 1989) trained on a large-scale text corpus. This enables a quantitative prediction of N400 amplitudes based on a cognitively motivated model, as well as quantitative comparison of this model to alternative models of the N400. Specifically, we compare the update measure from the SG model to Surprisal estimated by a comparable language model trained on next-word prediction. The results reported in this paper corroborate the hypothesis that N400 amplitudes correspond to the change in an implicit predictive representation of meaning after every word presentation. Furthermore, we argue that a comparison of the Sentence Gestalt update and Surprisal might also uncover two distinct but probably closely related sub-processes that contribute to the processing of a sentence.

1 Introduction

Understanding a sentence requires extracting semantic information about the events or states that it describes. The way this is achieved by the human brain, and more specifically the nature of the processes involved is a central question for the study of the neurobiology of language. The identification of several language-related neural correlates – such as the N400 component of the event-related potential – has allowed scientists over the last decades to have a more direct view of the neurocognition of sentence processing. Nonetheless, despite the mounting number of studies and data collected, the nature of these processes and the computations signaled by their neural correlates is still widely debated.

In this paper, we advance and test the hypothesis that the N400 event-related potential (ERP) component signals the online update of the brain's implicit predictive representation of meaning during sentence processing, and compare it with the hypothesis that instead, it reflects lexical prediction error. We do so by mapping the internal dynamics of an artificial neural network trained on sentence comprehension onto electroencephalographic data collected during sentence reading and by comparing the results to the effect of lexical Surprisal on the same data. This quantitative model comparison is enabled by scaling up our cognitive model of the N400 (Rabovsky et al., 2018) to a large scale corpus of naturalistic texts.

The N400 ERP component is a well-established correlate of meaning processing in the brain (Kutas & Federmeier, 2011), which offers a window into the neural processes supporting sentence comprehension. Understanding the computational processes underlying it can help us understand language and meaning processing. The N400 is a negative deflection at centro-parietal electrode sites peaking around 400 ms after the onset of a word or another potentially meaningful stimulus. Its amplitude has been shown to be affected by a wide variety of linguistic variables. The N400 effect was first discovered as a larger negativity for incongruent sentence continuations (such as e.g., ”I take my coffee with cream and dog”) as compared to congruent continuations (Kutas & Hillyard, 1980). In addition, N400 amplitudes tend to decrease over the course of a sentence (van Petten & Kutas, 1990). Smaller amplitudes are observed for targets after semantically similar or related as compared to unrelated primes and for repeated words as compared to a first presentation (Bentin, McCarthy, & Wood, 1985). It has been shown that the amplitude of the N400 is sensitive to the stochastic properties of the word in isolation (e.g, its frequency; e.g., Rabovsky, Álvarez, Hohlfeld, and Sommer (2008)) and in relation to its context of utterance (Kutas & Hillyard, 1984; van Petten & Kutas, 1990; Parviz, Johnson, Johnson, & Brock, 2011; Frank et al., 2015). Various theories propose for example that the N400 reflects, among others, lexical-semantic access to individual words or semantic integration processes at the sentence level (Kutas & Federmeier, 2011). However, in most studies, it is difficult to unequivocally decide whether reduced N400 amplitudes reflect facilitated lexical access due to prediction/pre-activation of upcoming input (Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008), or whether reduced N400 amplitudes reflect facilitated bottom-up semantic integration processing because the incoming input better fits the preceding context (C. M. Brown & Hagoort, 1993). There is also a debate revolving around the involvement of this component in predictive processes. Even though the idea of predictive preactivation underlying N400 amplitude reductions has been traditionally linked to the idea of facilitated lexical access rather than facilitated sentence level integration (Kutas & Federmeier, 2011), the issue of whether the N400 reflects predictive processing and preactivation is actually independent of the question of at which level (word or sentence level) the processes operate (Rabovsky et al., 2018). In general, despite the large body of studies conducted on the N400 and the hypothesis that its amplitude modulations are related to meaning processing, an adequate neurally-mechanistic account of the relation between the N400 and language processing with its theorized functional sub-processes is still actively debated.

Recent developments in the field of artificial intelligence have given rise to the possibility of using computational models as explicit hypotheses regarding the computations and mechanisms involved in several cognitive processes. This holds true for language processing in general, and its neural correlates, including the N400 ERP component. In the last decade, several studies have tried to explain the computations signaled by this component by comparing it to the performance or mechanisms implemented by various types of computational language models. These models can be ascribed to two broadly defined categories, (neuro)cognitively motivated small-scale models linking the N400 to internal processes in the models (Brouwer, Delogu, Venhuizen, & Crocker, 2021; Fitz & Chang, 2019; Rabovsky et al., 2018), and large-scale stochastic language models trained on next word prediction in naturalistic language corpora (Merkx & Frank, 2020; Michaelov & Bergen, 2020).

1.1 Next word prediction language models and the N400

Using next word prediction language processing models, Aurnhammer and Frank (2019); Merkx and Frank (2020); Michaelov and Bergen (2020) have shown that the N400 amplitude is significantly influenced by word-level Surprisal – a stochastic measure that quantifies how unexpected a word is given its context. More recently Michaelov, Bergen, and Coulson (2022) used Surprisal estimated by state-of-the-art language models (e.g., Gpt-3, (T. B. Brown et al., 2020)) and showed that it explains the amplitude of the N400 better than cloze probability based on human judgement. Additionally, a set of neuroimaging experiments provide evidence of the relation between Surprisal and processing in cortical regions responsible for language processing (Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2016; Lopopolo, Frank, van den Bosch, & Willems, 2017). These findings are generally taken to indicate that language processing is largely supported by predictive processes. This is reflected by the amplitude of the N400 and, more generally, by the activity in a large portion of the perisylvian and temporal cortex.

Importantly, Surprisal can be estimated by a variety of language models, and thus can be implemented in different ways. These include n-gram models (also known as stochastic Markov models), recurrent neural networks (RNN's) as well as transformer models (Jurafsky & Martin, 2009). These models can be referred to as next-word prediction language models since they are trained on predicting the next word in a sequence. In general, language models are often used to test neuro-cognitive theories of predictive processing, which posit that the brain constantly updates its expectations regarding incoming stimuli (including, but not limited to lexical units) as a function of previous inputs and conditional probabilities learned from the environment (Bar, 2011; Bubic, von Cramon, & Schubotz, 2010; K. Friston & Kiebel, 2009).

An advantage of these large-scale natural language processing (NLP) models is that they are trained on large linguistic datasets, approximating human language exposure so that they can be linked directly to empirical N400 data. However, the measure used to predict N400 amplitudes, Surprisal, is an output measure of the models that can be computed in many different ways. Thus, it does not directly speak to the internal cognitive processes and neural activation dynamics underlying N400 amplitudes, which is our main interest here. Surprisal, as obtained by these language models, is a measure of the performance of the model on the task it is asked to perform – and not the internal mechanisms aimed at performing it – therefore offering only a "computational" explanation of the electrophysiological correlate.

1.2 Cognitively motivated models of the N400

On the other hand, cognitively motivated computational models of language comprehension that have been used to model the N400 (Laszlo & Plaut, 2012; Cheyette & Plaut, 2017; Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Brouwer et al., 2021; Fitz & Chang, 2019; Rabovsky & McRae, 2014; Rabovsky et al., 2018; Rabovsky & McClelland, 2020; Rabovsky, 2020) link N400 amplitudes to internal hidden layer activation processes and dynamics. This provides an explicit and mechanistic explanation of the (neuro)cognitive processes giving rise to the N400. In other words, the link established by (neuro)cognitively motivated small-scale models between their internal processes and the N400 might instead offer an "algorithmic" (and partly "implementational") explanation of the underlying processes.

In the first study that attempted to link neuro-cognitively motivated computational models with insights from electrophysiological studies of language processing, Laszlo and Plaut (2012) proposed that the N400 corresponds to the magnitude of the activation of a semantic representation arising from the dynamic interaction between excitatory and inhibitory connections between layers of a connectionist model of visual word recognition, and successfully simulated influences of orthographic neighborhood on N400 amplitudes. On the other hand, Rabovsky and McRae (2014) proposed that N400 amplitudes reflect an implicit prediction error at the level of meaning, which they simulated as the network error in a feature-based attractor model of word meaning. Using this model, they simulated seven empirical N400 effects, namely influences of word frequency, semantic richness, repetition, orthographic neighborhood, semantic priming, as well as interactions of repetition with both word frequency and semantic richness. Subsequently, Cheyette and Plaut (2017) addressed the same range of N400 effects as Rabovsky and McRae (2014) using a model building on Laszlo and Plaut (2012)' proposal. The different models are not necessarily incompatible though and could be seen as complementary as the models by Laszlo and Plaut (2012) and Cheyette and Plaut (2017) address primarily the implementational level while the model by Rabovsky and McRae (2014) targets the algorithmic level of analysis according to Marr's classification. Both models focused on N400 effects observed during the processing of single words and word pairs.

Among the small-scale models addressing N400 effects during sentence processing, Brouwer et al. (2017, 2021) tested the hypothesis that the N400 component reflects the retrieval of word meaning from semantic memory, and the P600 component indexes the integration of this meaning into the unfolding utterance interpretation. They did so by implementing the two processes – retrieval and integration – in two separate modules of a recurrent neural network-based model trained on sentence comprehension. Patterns of ERP amplitudes were compared to the internal activity of the retrieval and integration modules. Specifically, they monitored the dynamics of the model's modules during the processing of reversal anomalies. Reversal anomalies are sentences such as ”Every morning at breakfast, the eggs would only eat...”, where N400 amplitudes are small despite the semantic incongruity, while amplitudes of the subsequent P600 component are increased. They observed that, in this situation, the amplitude of the N400 displays similar behavior with the activity of the retrieval module (i.e., both were small), therefore leading to their proposal that the N400 signal lexical-semantic retrieval mechanisms. On the other hand, the P600 was instead observed behaving similarly to the activity of the integration module in the sense that both were large (see Section 4 for further discussion).

Rabovsky et al. (2018) proposed an explanation of the N400 ERP component in terms of update of an implicit predictive representation of meaning as captured by the change of the inner states of the Sentence Gestalt (SG) model, a connectionist model of predictive language processing that maps a sentence to its corresponding event (McClelland et al., 1989). At every given moment during sentence processing, this representation not only contains information provided by the words presented so far, but also an approximation of all features of the sentence meaning based on the statistical regularities in the model's environment internalized in its connection weights. Rabovsky et al. (2018) showed that the SG model update (referred to as Semantic Update, or SU for short) simulates a number of N400 effects obtained in empirical research. These included the influences of semantic congruity, cloze probability, word position in the sentence, reversal anomalies, semantic and associative priming, categorically related incongruities, lexical frequency, repetition, and interactions between repetition and semantic congruity. These results foster the idea that N400 amplitudes reflect surprise at the level of meaning, defined as the change in the probability distribution over semantic features in an integrated representation of meaning occasioned by the arrival of each successive constituent of a sentence. This surprise at the level of meaning has also been shown to correspond to a learning signal driving adaptation and learning in the SG model. Specifically, because at any given point in sentence processing, the model attempts to predict all aspects of meaning of the described event, the change in SG activation induced by each new incoming word corresponds to the prediction error contained in the previous SG representation. Based on the idea that prediction errors drive learning, Rabovsky et al. (2018) used the difference in SG activation between the current and the next word as a learning signal to adjust connection weights in the model (simulation 16). This was taken to suggest that N400 amplitudes reflect an implicit error based learning signal during language comprehension (Rabovsky et al., 2018). Thus, the model by Rabovsky et al. (2018) refined the notion of an implicit prediction error at the level meaning proposed by Rabovsky and McRae (2014) by extending it to the sentence level, directly linking it to neural activation, and implementing it dynamically as a change in an implicit predictive representation of meaning.

Fitz and Chang (2019) extended the perspective that N400 amplitudes reflect a learning signal also to the P600. Specifically, they proposed that both the N400 and P600 arise as side effects of an error-based learning mechanism that explains linguistic adaptation and language learning (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Holroyd & Coles, 2002). Their model is based on Chang's Dual-path model – a connectionist model of language acquisition and sentence production (Chang, 2002). The model decouples sequence and meaning processing in two distinct pathways, which – during training – learn syntax and semantic regularities. Their theory is instantiated by observing how the model's processing error simulates data from several studies on the N400 (amplitude modulation by expectancy, contextual constraint, and sentence position), and on the P600 (agreement, tense, word category, sub-categorization and garden-path sentences).

These cognitively motivated models proposed to account for N400 amplitudes have as of yet been only trained on small artificial language corpora. This has some advantages with respect to the models' transparency but also entails important limitations. Specifically, the models' N400 correlate – its internal dynamics – can be related to empirical N400 data only in a qualitative way and only concerning specific experimental manipulations.

For instance, even though Rabovsky et al. (2018) showed that the SG model covers a wide range of distinct N400 effects, including ones that could not be accounted for by lexical surprisal, quantitative assessment of the relation between its activity and the amplitude of the N400 itself as well as quantitative model comparison was impossible. This is because it was trained only on a small synthetic language and therefore it could not be presented with the same stimuli presented in empirical experiments. For this reason, the relation between the model's N400 correlate and empirical N400 data remained somewhat abstract.

1.3 The present study

In the present study we overcome this limitation and quantitatively investigate the hypothesis that N400 amplitudes might reflect predictive processes at the level of sentence meaning based on our cognitively motivated SG model trained on a large scale corpus. Moreover, we quantitatively compare this hypothesis with the alternative hypothesis that they instead signal prediction error at the word level. These hypotheses are tested by quantitatively comparing two alternative artificial neural network-based models of language processing: the SG model and a next word prediction LM. From the SG model we estimate the change over successive words of its internal representations of the meaning of the sentence, otherwise known as Semantic Update (SU). Whereas, the LM, trained on next word prediction, is employed to estimate word-level

Surprisal, a measure that has already been shown to correlate with this ERP component (e.g., Frank et al., 2015). Using SU based on the internal dynamics of the SG model implies that N400 amplitudes reflect predictive processes at the level of sentence meaning. Whereas, using Surprisal derived from a next word prediction language model (LM) implements the idea that the N400 reflects lexical prediction and prediction error.

In the analyses conducted in the present study, we use these measures to directly predict the amplitude of the N400 generated during sentence processing. Moreover, in order to make the quantitative comparison between the effects of SU and Surprisal meaningful and fair, we compute this latter measure from a language model having an architecture as similar as possible to the SG model and that has been trained on the same linguistic material. This ensures, in our view, that the models' different performances with regard to the EEG data are not due to differences in training data or (as far as possible) basic architectural features, but that these differences are ascribable to differences in training task (next word prediction versus predictive meaning comprehension) and to the comparison between a performance-oriented measure (Surprisal) and an internal-processing one (SU).

The results reported in this paper show that SU significantly predicts the amplitude of the N400 and the time course of EEG activity between 300 and 500 ms post-word onset. This is true even when controlling for lexical Surprisal (and vice versa), hinting at the fact that these 2 measures might tackle different aspects or processes within the wider spectrum of activity elicited by sentence comprehension. As compared to lexical Surprisal, our measure of semantic update shows a more sustained temporal similarity to brain activity which also expands beyond the typical N400 time-window. The significant results of both SU and Surprisal, we argue, can be potentially explained as suggesting two distinct – yet presumably closely intertwined – processes.

1.4 The Sentence Gestalt model

The **Sentence Gestalt (SG) model** is a model of language comprehension which maps sentences to a description of the meaning of a utterance approximated by a list of arguments encompassing the action, the various participants (e.g., agent and patient) as well as information concerning, for instance, the time, location, and the manner of the situations or events conveyed by the speaker (McClelland et al., 1989). The model implements a theory of sentence processing centered around the computation of an internal representation of meanings informed and constrained by the sentence uttered by the speaker.

For instance, when processing the sentence "the boy opened the door slowly", the **task** of the SG model is to recognize that *opened* is the action, and that *the boy* and *door* are its agent and patient respectively and that *slowly* is a modifier specifying the way in which the event takes place. In order to process such a sentence, the model implements a mechanism consisting in building internal representations (here referred to as Sentence Gestalts) that can be used as a basis to respond to probes regarding the meaning of the sentence approximated by the argument structure of the utterance in terms of role filler pairs.

The model is probabilistic in the sense that it can be conceptualized as approximating a function computing a conditional probability distribution between role-filler pairs, representing the arguments of the sentence's event, and string of words composing the sentence itself: $P(< r, f > |w_{1:n})$, where r refers to the role (e.g., *agent*) and f to the filler (e.g., *boy*) of an argument, and $w_{1:n}$ is the sequence of the first n words of a sentence. For instance, given the sequence *the, boy, opened*, the model learns the conditional probability of the role-filler pair $< agent, boy >$. This is done on the basis of the statistical properties of the training environment and by using probes. Given a sentence and a probe, the training environment specifies a probability distribution over the possible answers to the probe and the goal of processing is to form a representation that allows the model to match this probability distribution. Moreover, the SG model is predictive in light of its training regime, which forces it to estimate role-filler pairs representing arguments of the whole sentence event structure word-by-word, even before the actual lexical items that correspond to the probed argument fillers are available. The model is therefore expected to attempt to use the representation to anticipate the expected responses to these probes. Section 2.2 below contains the details regarding the training procedure adopted in this study. Finally, the predictive and probabilistic processing at the level of meaning is internally represented in the model by the hidden Sentence Gestalt representations. These representations are implicit in that they represent sentence meaning using features that do not correspond to symbolic features, either semantic or grammatical. That makes them formally comparable to semantic representations obtained from vector-space models such as the ones produced by distributional semantics and deep neural network word embeddings.

The SG model differs from **next word prediction language models (LMs)** with regard to aspects of the theory of language processing that it implements. Despite the fact that both models are predictive, LMs implement a theory of processing at the level of lexical items alone. In contrast, as seen above, SG models are trained to predict sentence meanings from sequences of words. In line with this, both models approximate functions estimating a conditional probability distribution. If the SG model estimates conditional probabilities between lexical cues and role-filler pairs of event arguments, an LM instead estimates such probabilities between sequences of words and their following word. Given a sequence of words of length n, an LM assigns a probability $P(w_{n+1}|w_{1:n})$ to the next word w_{n+1}.

2 Materials and methods

2.1 The implementation

The **SG model** is constituted of two components: an update network (encoder) and a query network (decoder), as described in Fig. 1. The update network sequentially processes each

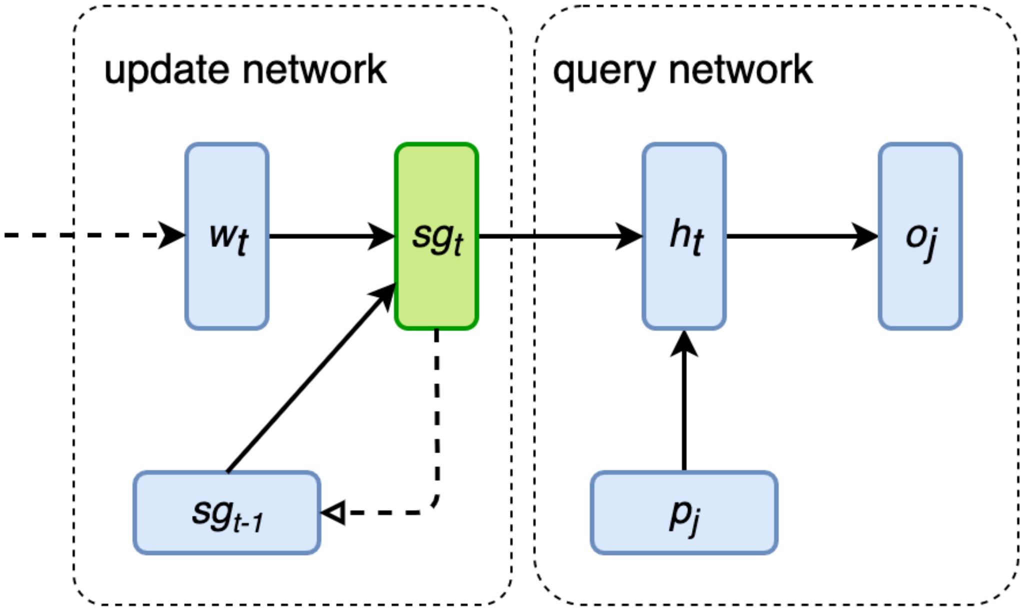

Fig. 1: The architecture of the Sentence Gestalt Model, with the **update network** on the left hand-side and the **query network** on the right hand-side.

incoming word to update activation of the Sentence Gestalt layer, which represents the meaning of the sentence after the presentation of each word as a function of its previous activation and the activation induced by the new incoming word. The query network, instead, extracts information concerning the event described by the sentence from the activation of the Sentence Gestalt layer. The sentence comprehension mechanism is implemented in the update network. The query network is primarily used for training.

The **update network** of the SG model is composed of an input layer, which generates a vectorial representation $\vec{w}_t$ for each input word of the incoming sentence, and a recurrent layer implemented as a long short-term memory (LSTM) unit generating a Sentence Gestalt representation $\vec{sg}_t$ as a function of the current input word $\vec{w}_t$ and its previous Gestalt representation $\vec{sg}_{t-1}$ (Hochreiter & Schmidhuber, 1997). LSTM have the advantage of being better at processing long and complex sentences compared to traditional recurrent layers, and being still simpler in structure and number of parameters compared to even more performative types of deep learning components (e.g. Transformers). The update network is essentially a recurrent neural network encoding a string of words as a function of the current presented word and the words preceding it.

The **query network** is composed by an hidden layer $\vec{h}_t$ which combines the Sentence Gestalt vector $\vec{sg}_t$ and a probe vector $\vec{p}_i$. The output $\vec{o}_i$ of the query network is generated from the hidden state $\vec{h}_t$.

The **task** the SG model is asked to perform is to map a sentence to its corresponding situation or event, defined as a list of role-filler pairs representing an action or state, its participants (e.g. agent, patient, recipient), and eventual modifiers. A sentence is defined as a sequence of words, each represented as an integer i_t, defined on a lexicon associating a unique index to every word. Figure 2 exemplifies the mapping from words to event performed, word-by-word, by the SG model. In this example, given the sentence *the boy opened the door*, the model tries to identify the filler of the *patient* role. The sentence is presented word-by-word. The model is first presented only with the words *the boy*, leading to a list of wrong predicted patient fillers. As soon as the model is presented with word *opened*, its prediction is adjusted to produce a list of potentially correct fillers. When the model is presented with the whole sentence, it converges on the correct patient of the described event: *door*.

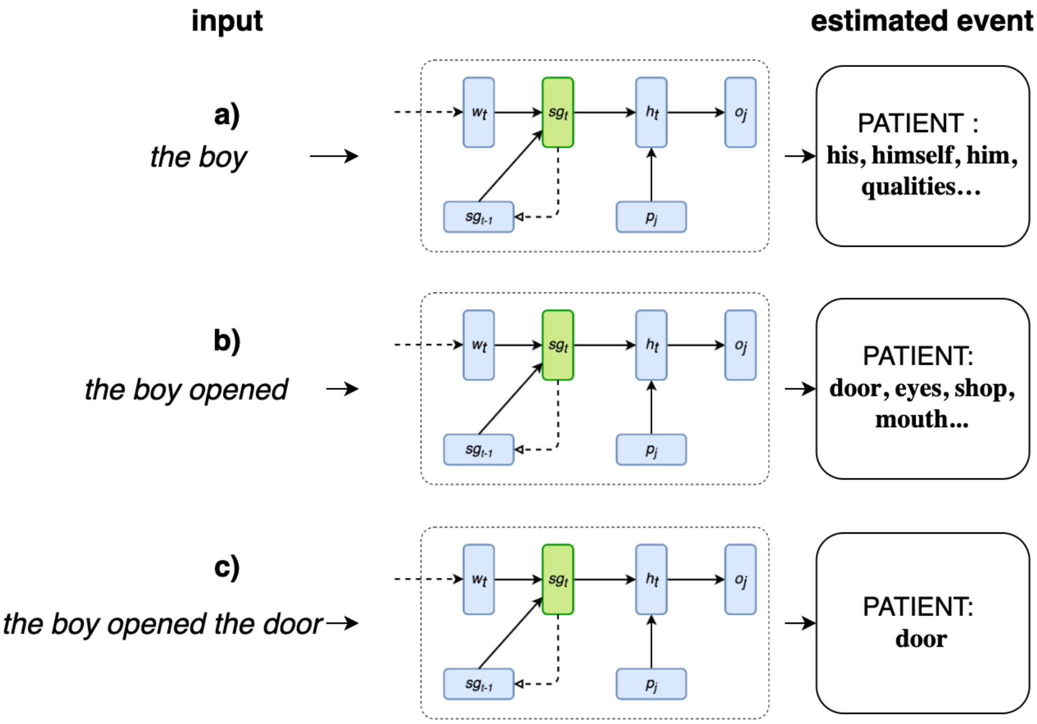

Fig. 2: Example of how the SG model processes the sentence *the boy opened the door*. The model is trained on identifying the roles and fillers constituting the event described in the sentence. Here, for reasons of space, we focus only on mapping filler to role *patient*, but the model also estimates the agent and action of the event. Sentences are presented word-by-word. Initially (a), the model is first presented only with the words *the boy*, leading to a list of wrong predicted patient fillers. Subsequently (b), the model is presented with word *opened*, causing the prediction of potentially correct fillers. Finally (c), when the model is presented with the whole sentence, it converges on the correct patient of the described event: *door*.

As shown in Fig. 3.a, the event consists of a set of role-filler vectors $\vec{o}_i$, each of which consists of the concatenation of the feature representation of a word and a one-hot vector of the role of that word in the context of the event described by the sentence. Therefore, the sentence “the boy opened the door slowly” will consist of a sequence of 6 one-hot word representation vectors. Its event contains 4 role-filler combinations representing each role of the event (agent, action, patient, manner) with its corresponding concept filler (boy, open, door, slowly).

During **training**, the model is presented with sentences (fed word by word to the input

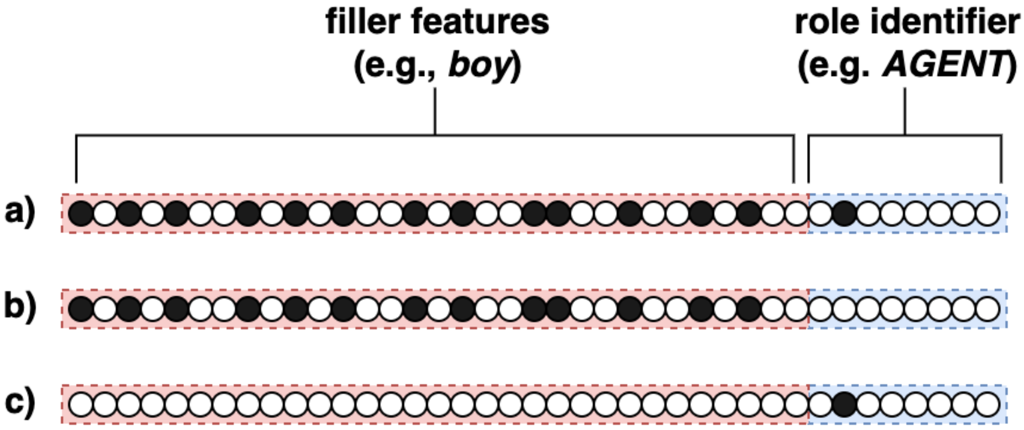

Fig. 3: The role-filler vector $\vec{o}_i$ (a), and its corresponding two types of probes $\vec{p}_i$ (b) and (c). The left hand-side of the vectors correspond to the embedding representation of the filler concept, whereas the right hand-side to the one-hot representation of the thematic role played by the filler. When probing for the thematic role, probe (b) is presented. When probing for the filler instead, probe (c) is presented. In both cases the SG model is expected to produce the full role-filler vector (a).

layer) such as "he opened the door slowly". Every time a word is presented, the model is probed concerning the event described by the sentence. The model is probed concerning the complete event after each presented word, even if the relevant information has not yet been presented at the input layer, allowing the model to learn to predict sentence meaning from partial input based on the statistical regularities of its environment. A probe consists of a vector $\vec{p}_i$ of the same size of a corresponding role-filler vector $\vec{p}_i$, but with either the thematic role identifier zeroed (Fig. 3.b) – if probing for roles –, or filler features zeroed (Fig. 3.c) – if instead probing for fillers. Responding to a probe consists therefore of completing the role-filler vector. When probed with either a thematic role (e.g., agent, action, patient, location, or situation; each represented by an individual unit at the probe and output layer) or a filler, the model is expected to output the complete role-filler vector. Fillers are represented using word embeddings obtained by binarizing *Fasttext*, a computational semantic model representing 1 million words and trained on both the English Wikipedia and the Gigaword 5 corpora (Bojanowski, Grave, Joulin, & Mikolov, 2017). The discrepancies between the observed role-filler vector $\vec{o}_i$ and generated output $\vec{\hat{o}}_i$ is computed using cross-entropy and is back-propagated through the entire network to adjust its parameters in order to minimize the difference between model-generated and correct output. Binarization of the filler semantic feature representations was performed in order to allow for a probabilistic interpretation of the model generated activation of semantic feature units afforded by the cross-entropy error used during training.

Besides being very powerful and widespread tools in natural language processing (NLP), **next word prediction LMs** have been used, as mentioned in the introduction, in numerous studies of language processing in humans, in psycholinguistics and neurobiology of language (Frank et al., 2015). RNN implementations of the LMs share significant similarities in the way they treat the linguistic input and in internal architecture with the SG model. In both cases language – for instance, sentences – are treated as sequences of words which are feed to an input layer which generates a per-word vectorial representation which is then integrated with a contextual representation by the internal recurrent layer. The LM is composed of an input layer, which generates a vectorial representation $\vec{w}_t$ for each input word of the incoming sentence, and a recurrent layer implemented as a long short-term memory (LSTM) computing a recurrent representation $\vec{r}_t$ as a function of the current input word $\vec{w}_t$ and its previous activation $\vec{r}_{t-1}$. The hidden layer $\vec{h}_t$ takes the recurrent representation $\vec{r}_t$ and feeds it to the output $\vec{o}_t$ which consist of a probability distribution over the model's lexicon representing the likelihood of the next word w_{t+1}.

2.2 Training the models

Both LM and SG model are trained on the same data, the Rollenwechsel-English (RW-eng) corpus (Sayeed, Shkadzko, & Demberg, 2018). The only crucial difference is – as mentioned above – the task the two models are called to perform: predict the next lexical item (i.e. mapping sequences to words), or understanding and predicting the meaning of the sentence (i.e. mapping sentences to role-filler pairs).

The SG model was trained on the whole Rollenwechsel-English (RW-eng) corpus (Sayeed et al., 2018). The RW-eng corpus is annotated with semantic role information based on PropBank roles (Palmer, Gildea, & Kingsbury, 2005) and obtained from the output of the SENNA semantic role labeller (Collobert et al., 2011; Collobert, 2011) and the MALT syntactic parser (Nivre, 2003). Each sentence annotation consists of the list of event frames it describes. An event frame is defined by its main predicate (usually a finite verb) and its arguments. Following PropBank, RW-eng frames can contain arguments of 26 types spanning from agent, patient, benefactive, starting and end point and a series of modifiers describing the temporal, locational, causal, final and modal circumstances of the event. Therefore, the SG model in this study is trained on mapping each RW-eng sentence to its PropBank-style event structure as provided in the RW-eng corpus. For more detail on the argument structure proposed by PropBank we refer to Palmer et al. (2005). A sentence can contain multiple event frames. For instance, the following sentence (taken from the RW-eng corpus):

	she	took	great	heaving	breaths	as	though	she	had	come	up	a	hill
frame 1	*agent*	*action*	*patient*			*manner*							
frame 2				*action*	*agent*								
frame 3								*agent*	*action*		*direction*		

The sentence is composed by 13 lexical items and describes 3 nested events represented by 3 separate PropBank-style frames. The first, color-coded in red, contains a patient ("great heaving breaths") and a manner modifier ("as though she had come up a hill") which in turn are frames on their own right (color-coded in green and blue respectively). It is worth noting that the SG model is a model of language comprehension in general, and that it is trained not only on events. The training dataset includes also other types of utterances such as stative sentences, questions and commands among others. This example illustrates how the SG model is trained on naturalistic, complex, English sentences, by presenting it with one word at a time, following the sentential order, and probed each time for the role-fillers of the argument of potentially several events at the same time.

Instead, the LM was trained only on the raw sequences of words (sentences) contained in the RW-eng corpus, disregarding the semantic annotation based on PropBank. When trained on the example sentence above, the model is asked to predict the simple linear sequence of lexical items. It is presented, for instance with the input sequence "...she took great heaving" and asked to match it with the next lexical item "breaths", as per the example sentence above.

The parameters of both models were optimized using Adamax (Kingma & Ba, 2015) with learning rate equal to 0.01 for the SG model and 0.002 for the LM (after hyper-parameters optimization on 5% of the data). The whole dataset was split in mini-batches of 32 sentences each. Training was conducted for a maximum of 150 epochs on 90% of the batches, the remaining 10% was kept for validation. Only sentences having between 6 and 15 words and having a maximum of 8 frames were used for training. Sentence length and number of frames were constrained in order to limit the number of complex subordinate events and to facilitate the mini-batch training.

The size of the input layers of both model was equal to 600 and their recurrent layers size equal to 1200 (the Sentence Gestalt layer for the SG model), implemented as a 1-layer LSTM. The SG model's probe and output layers had size 328 due to the concatenation of the 300-size binarized embedding vector, the frame number and the argument type. Instead the output layer of the LM was equal to the extension of its lexicon: 300000.

2.3 Model-based predictors

The goal of these two modelling approach is to investigate the relation between brain activity (more specifically N400 amplitudes and the related EEG time-course post word onset) and the update of an internal predictive representation of sentence meaning, on one hand, and lexical prediction error, on the other. For this reason we compute two main measures from our models: semantic update (SU) and Surprisal.

Semantic update (SU) is the update of the SG model's recurrent layer's internal representations after the presentation of each word composing a sentence. It is computed as the mean absolute error between the activation of the Sentence Gestalt layer after the presentation of a word and its activation before the presentation of that word. It represents the amount of change driven by the new incoming word to the implicit predictive representation of sentence meaning computed by the SG model.

Surprisal instead is the negative log-probability of the new incoming word given the previous words: $surp(w_t) = -log(P(w_t|w_{1:t-1}))$. It is estimated from the probability distribution generated as output of the LM. Surprisal operates on lexical items, and not on the semantic representations, and is taken to represent the error in predicting such items in a sequential input.

2.4 EEG dataset

The elecrophysiological recordings of the N400 were obtained from an EEG dataset provided by Frank et al. (2015). The dataset consists of data collected from twenty-four participants (10 female, mean age 28.0 years, all right handed and native speakers of English) while they were reading sentences extracted from English narrative texts.

The stimuli consisted of 205 sentences (1931 word tokens) from the UCL corpus of reading times (Frank, Monsalve, Thompson, & Vigliocco, 2013), and originally from three little known novels. The sentences were presented in random order, word by word. About half of the sentences were paired with a yes/no comprehension question to ensure that participants read attentively. The sentence stimuli were not manipulated to control or elicit any particular linguistic phenomenon.

The sentences were presented in random order, word-by-word. Each word was presented in the center of the screen for a variable amount of time, depending on their length as number of characters. Word presentation duration equalled 190 ms plus 20 for each characters in the word. Each word was followed by a 390 ms interval before appearance of the next word.

The EEG signal was recorded continuously at a rate of 500 Hz from 32 scalp sites (Easycap montage M10) and the two mastoids. Signal was band-pass filtered online between 0.01 and 35 Hz. Offline, signals were filtered between 0.05 and 25 Hz (zero phase shift, 96 dB roll-off). The N400 amplitude for each subject and word token was defined as the average scalp potential over a 300-500 ms time window after word onset at electrode sites in a centro-parietal region of interest (ROI).

For further details regarding the stimuli see Frank et al. (2013). More detailed information regarding the EEG dataset, its stimulation paradigm, preprocessing, and the electrode positions contained in the ROI can instead be found in Frank et al. (2015).

3 Analyses

In this study we want to test the hypotheses that language-elicited EEG activity, specifically the amplitude of the N400 component, reflects prediction errors at the level of sentence meaning or lexical-items.

In Section 3.1 we fit a linear mixed effect model with the aim of predicting the amplitude of the N400 (average activity over a 300-500 ms time window after word onset at centro-

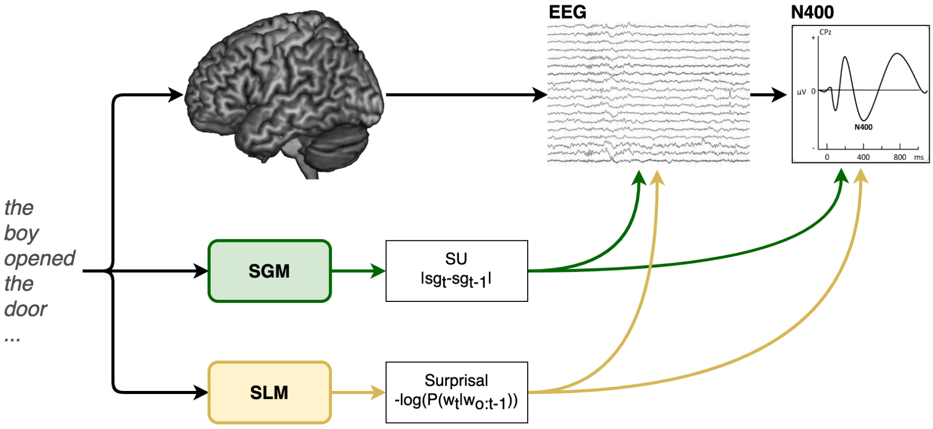

Fig. 4: This study aims to model the amplitude of the N400 and the electrophysiological activity locked to the presentation of each word during sentence comprehension as a function of the update of an implicit predictive semantic representation (SU) generated by a SG model trained on a large scale corpus of naturalistic texts. The effect of SU is compared to the one of Surprisal as computed by an LM.

parietal electrodes) as a function of the update of the semantic representation generated by the SG model during language processing (the Semantic Update or SU). The SU is computed as the mean absolute error between $\vec{sg}_t$ and $\vec{sg}_{t-1}$ generated by the Sentence Gestalt layer of the update network. Fig. 4 presents a graphical summary of our approach. Since the N400 is a negative deflection of the electrophysiological signal, the SU is multiplied by -1.

Since the information reflected by the SU might go beyond the time window of the N400 most typically defined as the average per-trial activity between 300 and 500 ms post-word onset, and to demonstrate the specificity of the observed relationship between SU and N400 amplitudes, in Section 3.3 we predict the complete time-course of the EEG signal in order to explore whether earlier or later latencies also correlate with the SG model's internal dynamics. These analyses are compared and contrasted to the effect of lexical Surprisal on the same electrophysiological measurements. In Section 3.2, we assess the contribution of the SU above and beyond purely lexical predictive processing, and directly compare how well SU versus surprisal predict N400 amplitudes.

Finally, in Section 3.4, we investigate the effect of model training, i.e. the amount of exposure to the training data, of the fit between SU and the amplitudes of the N400.

3.1 Predicting the N400

Tab. 1 contains the results of a linear mixed effect model predicting the N400 ERP component amplitude obtained from Frank et al. (2015) as a function of the SU over the stimulus words. SU is included together with the ERP baseline (the activity of the 100 ms leading to the onset of each word). In order to avoid potential artefacts, the baseline is not subtracted directly from the dependent variable, but instead included as a variable of no interest in the model. The model is fit with per-subject and per-word random intercepts. The N400 was computed as the average activity between 300 and 500 milliseconds after word onset in central, posterior and lateral electrodes.

	β	t	p
ERPbase	-0.09	-21.00	< 0.0001
SU	7.15	8.27	< 0.0001

Tab. 1: Linear mixed effect model fitted with the SU and aimed at predicting the amplitude of the N400 component.

The results in Tab. 1 clearly indicate that SU significantly predicts the amplitude of the N400 ($\beta = 7.21$, $t = 8.27$, $p < 0.0001$ FDR corrected). This indicates that larger word-wise updates of the Sentence Gestalt layer representation correspond with stronger negative deviation of the ERP signal in the N400 time segment.

3.2 Comparing Surprisal and SU as predictors of the N400

In order to assess the contribution of the SU on the amplitude of the N400 above and beyond the effect of Surprisal, we fitted two nested linear mixed effects models, one containing as predictors only Surprisal, the other containing also SU. Both models were fit with per-subject and per-word random intercepts. The results of the log-likelihood test between the two models show the difference in model fit to be significant ($\chi^2 = 71.08$, $p < 0.001$), the ΔAIC amounts to 70.6.

	β	t	p
ERPbase	-0.09	-21.00	< 0.001
Surprisal	0.05	4.52	< 0.001
SU	7.28	8.45	< 0.001

Tab. 2: Results of a *Full* model fitted with both Surprisal and SU and aimed at predicting the amplitude of the N400 component.

Table 2 contains the β estimates for the *Full* model predictors. Even with the presence of Surprisal ($\beta = 0.05$, $t = 4.52$, $p < 0.001$ FDR corrected), SU makes a significant contribution to the amplitude of the N400 ($\beta = 7.28$, $t = 8.45$, $p < 0.001$ FDR corrected).

We additionally compared two linear-mixed effect models predicting the N400 either as a function of Surprisal or SU, both fit with per-subject and per-word random intercepts. The results of the log-likelihood test between the two models yield the following results: The results of the log-likelihood test between the two models show the difference in model fit to be significant ($\chi^2 = 50.89$, $p < 0.001$). The ΔAIC between the SU and Surprisal models amounts to 59.5 and indicates a general better fit for SU as compared to Surprisal.

3.3 Predicting the EEG time-course and topographical distribution

In order to delineate the relation between SU and Surprisal with brain activity besides the N400, we fit a series of linear mixed effect models on each time-point of the EEG signal averaged across a region of interest defined over posterior-lateral electrodes (the same as used for the N400 analyses reported in the previous sections). The linear models followed the same structure as in Section 3.1, controlling for the baseline defined as the pre-word onset activity and with either SU or Surprisal as predictor of interest. We report the t-values over time (from 0 to 650 ms after word onset) relative to the estimates of SU or Surprisal.

Figure 5 shows the time course of the fit of SU and Surprisal to the EEG activity averaged over lateral-posterior electrodes, corresponding to the electrodes employed by Frank et al. (2015) to define the N400. The solid red line indicates the effect of SU, and the solid blue line the one of Surprisal.

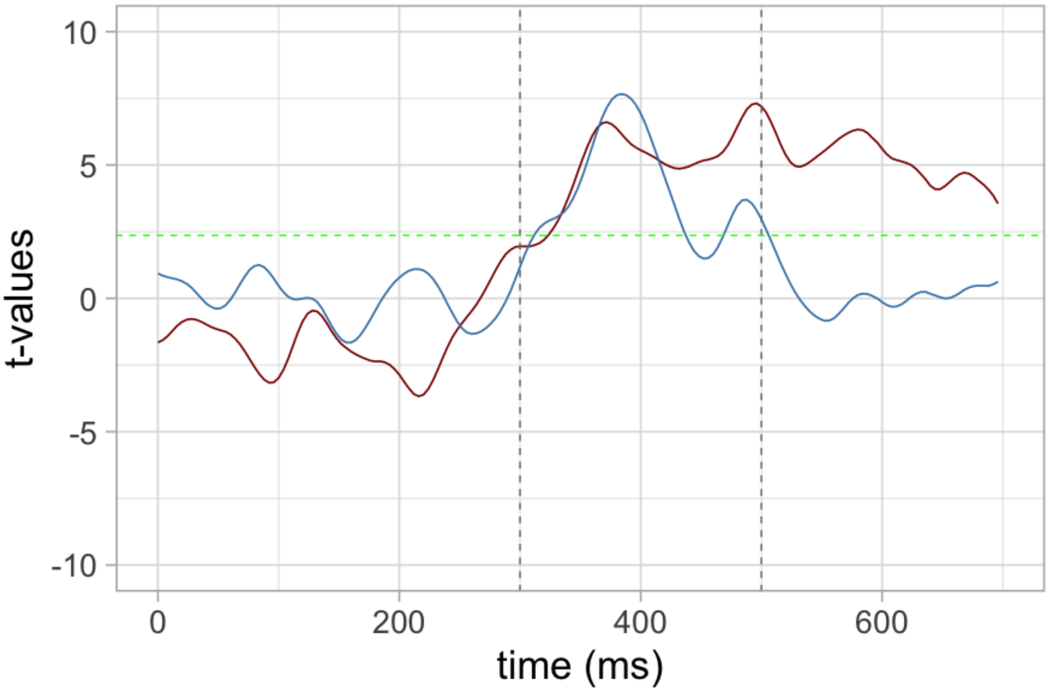

Fig. 5: The time-wise results of a series of independent linear mixed effect models predicting the EEG signal in a ROI defined over N400-sensitive electrodes as a function of the SU (red) and Surprisal (blue) from the LM.

SU significantly explains EEG activity from 300 ms onward after stimulus onset. Two separate peak appear at 375 ms ($t = 6.51$, $p < 0.0001$ FDR corrected) and at 500 ms after stimulus onset ($t = 7.21$, $p < 0.0001$ FDR corrected). Interestingly, and in line with the previous literature, Surprisal shows a strong effect peaking at 388 ms after stimulus onset ($t = 7.51$, $p < 0.0001$ FDR corrected). The effect of Surprisal, compared to SU, decreases rapidly after this time point.

Furthermore, we replicated the same time-wise analyses on each separate electrode in the dataset and plotted the results in a series of topographical maps representing the distribution of the fit between SU or Surprisal to electrophysiological activity using 50 ms wide non-overlapping time-windows. Figures 6 and 7 display the topographical distribution of the fit between EEG activity and SU (Figure 6) and Surprisal (Figure 7). Here it is evident again how the effect of SU extends the whole time-span of the N400 at centro-parietal electrode positions, in line with the N400 spatio-temporal profile. Surprisal shows a similar distribution, although limited to the early temporal phases of the latencies of the N400 ERP component. Moreover, both regressors display an inverse effects in frontal regions at later latencies. For SU this is mainly limited to the 650 to 700 ms time-window and only slightly left-lateralized, whereas for Surprisal this phenomenon starts to appear right after the N400 time-window, around 500 ms post stimulus onset and is more strongly left-lateralized.

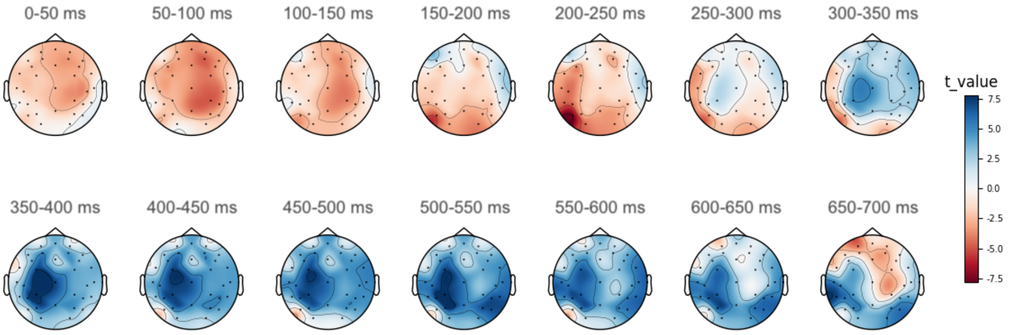

Fig. 6: Topographical plot of the goodness of fit of a linear mixed effect model predicting the EEG signal over time – from 0 to 600 ms post stimulus onset – as a function of SU (50 ms wide non-overlapping time-windows).

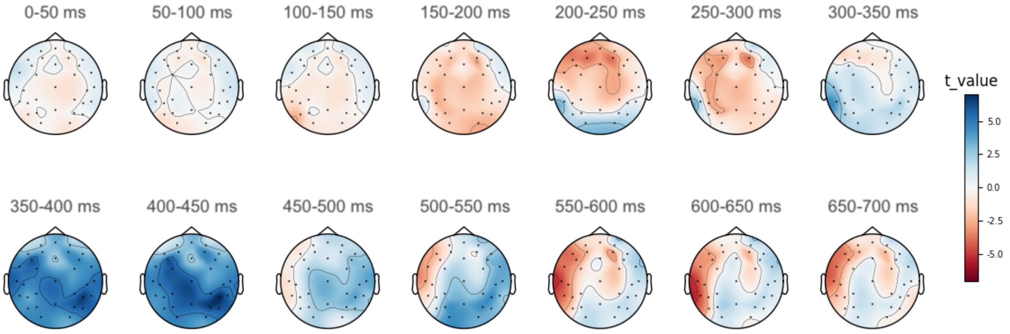

Fig. 7: Topographical plot of the goodness of fit of a linear mixed effect model predicting the EEG signal over time – from 0 to 600 ms post stimulus onset – as a function of of RNN language model estimated Surprisal (50 ms wide non-overlapping time-windows).

3.4 The effect of training

As humans acquire semantic knowledge about the world by experiencing their environment, the model learns knowledge about event semantics by being exposed to the statistic regularities in its environment – in this case the sentences contained in the RW-EN corpus – during

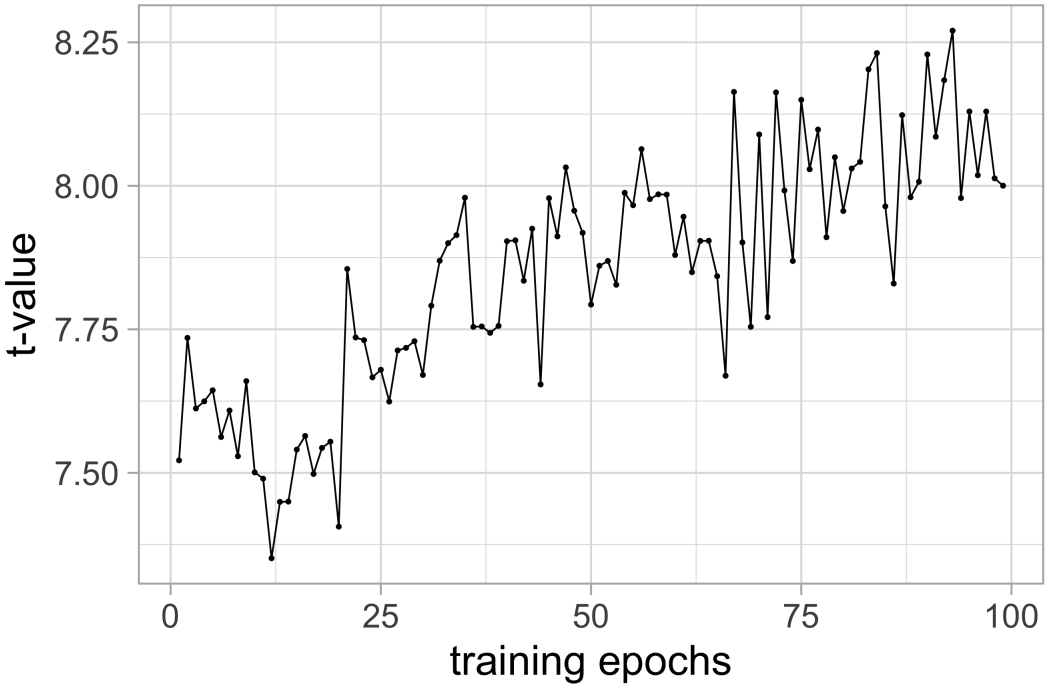

Fig. 8: The fit of SU on the amplitude of the N400 quantified as t-value for the SG model during training. Each point reports the t-value relative to the SU as predictor of the N400 in a linear mixed effect model; SU's values are obtained from the same SG model at different stages (epochs) of training, from after the first epoch of training to epoch 100. Please note that epoch 0 is not displayed in this Figure and is instead reported in Table 3.

training. In order to investigate whether the amount of training affects the similarity between the model's internal dynamics and the N400 amplitudes, we conducted similar analyses as in Section 3.1 using SU measures obtained from the same SG model over different training stages.

In Figure 8 we plot the effects of the Semantic Update of the SG model as predictor of the amplitude of the N400 from after the first training epoch to epoch 100. During one training epoch the model is exposed to the whole training data and its parameters are updated in order to minimize its output error. The analyses, including the definition of the N400 and the linear mixed effect model, were conducted as in Section 3.1. The observed trend indicates that the amount of training undergone by the model affects the similarity between its internal dynamics and the amplitude of the N400.

model	chisq	p	ΔAIC
untrained SGM	0.23	0.6326	4.3
trained SGM	67.92	< 0.001	67.5

Tab. 3: Results of an analysis of variance and AIC reduction between a linear model of the N400 without SU and one with SU from an untrained SG model (top) and between the same base model and one containing the SU values from a trained SG model (bottom).

Next we compared mixed linear regressions with the previously-discussed random effects structure to those also including a main effect of SU from either a completely untrained SG model and a trained one. Table 3 reports the difference in AIC and the analysis of variance. Only the Semantic Update obtained from the trained model appears to improve the fit to the amplitude of the N400 . The effect of the SU of the untrained SG model is not significant and negative ($\beta = -3.91$, $t = -0.48$, $p = 0.63$ FDR corrected).

4 Discussion

Although there is a substantial amount of evidence linking the N400 to the processing of language, there is still a great deal of debate as to the nature of the processes it represents. It is especially true in the case of the mechanisms that give rise to it. In this study, we advanced and tested the hypothesis that the N400 might signal the update of the implicit predictive representation of the meaning of a sentence during comprehension. This was accomplished by implementing the process in the form of a cognitively motivated artificial neural network model of predictive sentence comprehension trained on a large scale corpus, and by quantitatively mapping its internal dynamics (Semantic Update) to the N400 amplitude and to the EEG measured during sentence comprehension.

We quantitatively compared this account of the N400 to the alternative hypothesis that this signal instead reflects lexical Surprisal and is hence driven by predictive processing at the level of lexical units. We did so by using naturalistic stimuli and models trained on large-scale naturalistic corpora, yet based on cognitively-informed architectures. The usage of naturalistic training material allowed the models to learn from a linguistic environment as complex and close as possible to the one humans experience in the course of their lives. On a more technical level, it also allowed the models to be presented with the same stimuli used during human electrophysiological data collection, allowing for direct and quantitative comparison between model-derived variables and brain activity.

4.1 The time-course of semantic update

Our results show a significant relationship between Semantic Update and the amplitude of the N400 ERP component in a time window between 300 and 500 ms post stimulus (see Section 3.1). This is confirmed by time-resolved analyses (see Section 3.3), which show that SU has a significant effect on the EEG activity in line with the latency (Fig. 5) and the topographical distribution (Fig. 6) of the N400. More precisely, the effect of SU peaks both at 350 ms and 500 ms post-word onset, and extends beyond 500 ms post-word onset in line with studies reporting N400 effects extending beyond the typical N400 time window (Rabovsky et al., 2008). These observations are in line with the results reported by Rabovsky, Hansen, and McClelland (2016) and Rabovsky et al. (2018) which observed how the magnitude of the SU estimated by a SG model trained on a small-scale world behaves strikingly similar to the amplitudes of the N400 under experimental conditions manipulating semantic, stochastic and positional variables, among others. Overall, the results support the hypothesis that the N400 component of the event-related brain potential reflects the change induced by an incoming stimulus in an implicit predictive representation of meaning during sentence comprehension.

4.2 Lexical and sentence meaning prediction

The same time-wise analyses (Section 3.3, Fig. 5) show that also lexical Surprisal significantly explains electrophysiological activity in the N400 time frame. Its fit partially overlaps with the one of the SU, especially at early stages of activity. Nonetheless, the profiles of these two predictors with regard to the data differ in two interesting aspects. Even though they both show a sharp increase in fit around 300 ms post-word onset, the fit of Surprisal displays only one strong peak around 350 ms and then a drop after 400 ms post-word onset, contrary to the effect of SU which remains strong during the entire N400 time window and beyond. Our results demonstrate that when looking at the complete N400 time window, SU shows a better fit to N400 amplitudes than Surprisal (see Section 3.2). Moreover, the results reported in this Section show a significant effect of SU on the amplitude of the N400 component after controlling for Surprisal (and vice versa).

One possible explanation of these results is that the activity after 300 ms from the presentation of a word and the amplitude of the N400, in particular, do not reflect a single monolithic process, but rather the combined activity of multiple sub-components which are associated with related yet distinguishable cognitive processes. It is possible that language processing is supported by predictive processes across several levels of representation, such as at the level of individual words (lexical level) and sentence meaning. The existence of at least two distinct predictive sub-components at the level of lexical and semantic processing might be also compatible with a hierarchy of predictive processes (K. J. Friston, 2005; K. Friston & Kiebel, 2009). This is the hypothesis that predictions and prediction errors are passed up the hierarchy from lower to higher processing stages. That might be reflected by initial parts of the N400 potentially signaling primarily lexical prediction errors (explaining the high peak for Surprisal in the initial part of the N400), while later parts of the N400 may reflect primarily sentence meaning related prediction errors (explaining the second high peak for SU). This is potentially in line with Nieuwland et al. (2020)'s analysis which revealed temporally distinct (though overlapping) effects of predictability and plausibility on the N400. In particular, lexical predictability negatively correlated with widespread activity peaking around 350 ms after word onset. By contrast, plausibility was associated with a smaller, right-lateralized effect that started after the peak of the predictability and continued after the end of the N400 time frame. In light of these results, the authors propose that semantic facilitation of predictable words is produced by a cascade of processes that activate and integrate contextual word meaning into semantic representations at the level of the sentence.

This is in line with predictive coding and parallel processing theories, which have found support also from several brain-imaging studies. Using stochastic language models trained on different levels of linguistic description, Lopopolo et al. (2017) showed that predictive processes may be decomposed in separate domain-specific sub-processes mapped onto distinct cortical areas. This hypothesis is supported also by Heilbron, Armeni, Schoffelen, Hagoort, and de Lange (2020) who observed dissociable MEG neural signatures of lexical, syntactic, phonemic, and semantic predictions.

SU and Surprisal are measures that represent two different processes, and they tackle these processes from two different levels of analysis framed along the lines of Marr's distinction between the *computational* and *algorithmic* (Marr, 1982) levels. Generally speaking, the computational level refers to the goal of the process carried out by a system (artificial or biological). The algorithmic, on the other hand, refers to the manner this process is implemented, and, in particular, the type of representations and operations involved. Semantic Update measures the change over time of the SG model's activity during sentence processing. Therefore it is concerned with the representations and operations (for instance the recursive integration over time implemented by the SG layer) used by the system to perform the task of sentence comprehension. It thus offers a description of the algorithmic level of the process. For this reason, the observed effect of SU on the amplitude of the N400 might reflect the internal dynamics of a meaning construction mechanism signaled by the component. Surprisal, on the other hand, is obtained from the probability assigned by a language processing system to an incoming word given its context of utterance. As a stochastic measure, after the presentation of a novel word, Surprisal quantifies how much the system is off with regard to the estimation of the probability of that word appearing. For this reason, it can also be interpreted as a lexical prediction error. Having this in mind, it is therefore a measure of the performance of a system in solving the task of predicting the next word, therefore making it a descriptor of the computational level of the process, in the sense in which Marr defines it.

4.3 Relation to anterior post N400 positivity

Besides predicting N400 amplitudes, i.e., more negative amplitudes between 300 and 500 ms post stimulus onset, SU and Surprisal appear to be have an inverse effect on activity recorded in anterior regions following the N400 time-window (Figures /reffig7 and Figure /reffig8), indicating a relation between the processes of lexical and sentence meaning prediction as instantiated by our predictors and a frontal post N400 positivity. This seems in line with previous evidence showing positive components in anterior areas after the N400 (Thornhill & Van Petten, 2012; Petten & Luka, 2012a; DeLong, Quante, & Kutas, 2014; Brothers, Wlotko, Warnke, & Kuperberg, 2020; DeLong & Kutas, 2020; Kuperberg, Brothers, & Wlotko, 2020). The exact functional interpretation of these late anterior positivities in language processing is still actively debated. Kuperberg et al. (2020) argue that the post N400 frontal positivity might reflect a successful update of the meaning representation as a function of novel unpredicted input, whereas Thornhill and Van Petten (2012) interpret it as a function of lexical predictability independent from semantic similarity. One picture that seems to be emerging is that different from the late posterior positivity (P600; see next section), the late anterior positivity is increased for sentence continuations that are unexpected yet not implausible.

4.4 Comparison to alternative cognitive models of the N400

Our results seem to indicate that the N400 is indexing the change induced by an incoming stimulus to the representation of sentence meaning as recorded from the recurrent layer of an SG model, and that larger sentence meaning update continues to predict more negative ERP amplitudes over centro-parietal regions beyond the typical N400 segment. These conclusions are seemingly at odds with Brouwer et al. (2017) and Brouwer et al. (2021) that claim that the N400 instead indexes lexical retrieval, i.e. the activation from semantic memory of the meaning of each separate word in a sentence during its processing, and that the P600, a subsequent positivity over centro-parietal regions, instead, correlates with integrative processes aimed at constructing sentence level semantic representations. It is important to note that our results seem compatible with both word and sentence level contributions to the amplitude of the N400, which seem to partially overlap in time, a finding that seems at least partly consistent with Brouwer's view on the N400 (lexical Surprisal does not exactly correspond to Brouwer et al's measure of the N400 because it is based on word form rather than word meaning, but our current study cannot disentangle both measures). Importantly, however, our results provide strong evidence against the view that the P600 reflects sentence meaning update. If the P600 would reflect sentence meaning update during naturalistic reading as suggested by Brouwer et al. (2017, 2021), one would expect that larger SU (which reflects sentence meaning update) would predict more positive amplitudes at centro-parietal electrode sites in the P600 time segment starting at about 500 ms and extending beyond the end of Frank et al. (2015)'s recording segment. However, we find the opposite: Larger SU continues to predict more negative amplitudes at centro-parietal electrodes beyond the typical N400 window and into the P600 time window.[1]

How can we explain this discrepancy that Brouwer et al. (2017, 2021) reported that larger P600 amplitudes could be explained by larger sentence meaning update in their small-scale model, while we find in our quantitative analysis during naturalistic reading that larger sentence meaning update predicts more negative centroparietal ERPs during the P600 segment? It is important to note that both of Brouwer's simulations concerned very specific experimental conditions. The 2017 model focused on reversal anomalies, i.e. sentences such as "Every morning at breakfast, the eggs would eat..." (Kuperberg, Sitnikova, Caplan, & Holcomb, 2003). In the model, the small N400 observed in these situations is taken to reflect facilitated lexical access due to semantic priming (from breakfast and eggs to eat) and the large P600 is taken to reflect more effortful sentence meaning update, due to the incongruency that the eggs in the

[1] In the present study, we predict the EEG data from the negative of SU, in order to accommodate the fact that the N400 is a negative component, and the fact that we expect (and find) larger SU to predict more negative N400 amplitudes.

sentence are agents of an eating action even though they are not usually able to do so. An alternative explanation of the same pattern of results is that the small N400 reflects an initial *semantic illusion*, in which the readers temporarily take the eggs to be the patient rather than the agent of the eating action, based on semantic priors (Kim & Osterhout, 2005; Rabovsky et al., 2018) and in line with good enough approaches to language comprehension (Ferreira, Bailey, & Ferraro, 2002; Ferreira, 2003). In this view, the large P600 might reflect an internal revision, i.e., the process that after the initial illusion participants detect that there is an anomaly in the sentence and revise their internal sentence representation so that it matches the syntactic structure. Thus, even though the large P600 can be interpreted and modeled as large sentence meaning update (as done by Brouwer et al. (2017)), it is arguably a very specific update process potentially including an internal revision process. The interpretation of the P600 as reflecting an internal revision process is compatible with long held views on the P600 (even though this interpretation has initially focused on the P600 as reflecting syntactic revision processes (Kaan & Swaab, 2003; Friederici, Mecklinger, Spencer, Steinhauer, & Donchin, 2001), it has been expanded to revisions of internal representations more generally (Petten & Luka, 2012b; Rabovsky & McClelland, 2020)). The second simulation of an N400 and P600 pattern (reported in Brouwer et al. (2021)) follows the same lines of the one above in that the observed increased P600 might again reflect an internal revision process rather than a simple sentence meaning update. Specifically, the experimental conditions were as follows (based on a study conducted by Delogu, Brouwer, and Crocker (2019)):

Baseline:	John entered the restaurant. Before long, he opened the menu.
Event-related violation:	John left the restaurant. Before long, he opened the menu.
Event-unrelated violation:	John entered the apartment. Before long, he opened the menu.

The event-related violation was the only condition that elicited a large P600. Since aspects of the context (e.g., “the restaurant”) and event (”opened the menu”) appear to be related, it is possible that, upon first reading, participants may initially believe that the sentence makes sense, resulting in a reduced N400, before realizing that something is off and revising their initial interpretation (resulting in a larger P600). Alternatively, as proposed by Brouwer et al. (2021), the small N400 can be explained as reflecting facilitated lexical access due to priming, while the P600 can, instead, reflect sentence meaning update. According to this latter explanation, one would expect an equally large P600 for the event-unrelated violation in which the sentence meaning update (in the absence of illusions and revisions) would be equally large. Crucially, the data did not show this: The P600 in the event-unrelated violation was equally small as in the baseline condition (Delogu et al., 2019). This pattern of results (of a large P600 only in the event related violation condition) makes sense from the perspective that the P600 reflects revision processes as these revision processes would be expected in the event-related violation but not in the event unrelated violation.[2]

[2] However, Brouwer et al. (2021) claimed that there was no possible explanation for this specific pattern of

To sum up, in both experimental settings in which Brouwer et al. (2017, 2021) simulated larger P600 amplitudes as larger sentence meaning update, the increased P600 might be alternatively explained as reflecting internal revision processes. Against this backdrop, our finding that larger sentence meaning update as measured by a large scale SG model during naturalistic reading does definitely not predict larger P600 amplitudes (instead predicting more negative amplitudes at centro-parietal electrode sites beyond the N400 time segment) might be seen as further evidence strengthening the view that the increased P600 observed in the specific experimental conditions targeted by Brouwer et al. (2017)'s models might rather reflect internal revision processes and that under natural conditions larger sentence meaning update is reflected in larger N400 amplitudes.

Our finding that N400 amplitudes are overall better predicted by SU than by surprisal (see Section 3.2) seems to provide evidence against the model by Fitz and Chang (2019) who link the N400 to purely lexical prediction error. However, our results seem consistent with lexical contributions to the early part of the N400 (see Sections 3.2 and 3.3), which thus seems partially consistent with their claims (please note that in the current study we do not differentiate between lexical contributions at the level of word form versus word meaning). Our results do not seem to speak to Fitz and Chang (2019)'s account of the P600 as a sequencing prediction error. However, the results of reversal anomalies with two animate event participants, such as "The fox hunted the poacher", which do not include sequencing prediction errors but produce large P600 amplitudes (van Herten, Kolk, & Chwilla, 2005), seem inconsistent with their implementation.

4.5 The role of training

In Section 3.4, we showed that the model's ability to predict the N400 amplitude is strongly affected by the amount of exposure to its training environment quantified in terms of the number of training epochs. It is evident from these results that the SU obtained from a model better approximates the N400 the more the model is exposed to the statistical regularities of the training corpus. Conversely, our results indicate that a random model, i.e. an SG model with randomly initialized parameters fits poorly our target electrophysiological variable.

Recently, Schrimpf et al. (2021) analyzed the fit between fMRI and ECoG data in the frontal-temporal cortex and vectorial representations generated by 43 deep learning models. Similarly to our study, besides using fully trained models, they also evaluate the same models before training. Interestingly, they observed that untrained networks yield representations that still significantly predict fMRI data, although training significantly improves fit. These results led to the proposal that the architecture of the networks can work as brain models of

results and subtracted the complete activity in the N400 time segment from the activity in the P600 segment to correct for "component overlap". Even though the partial component overlap between the N400 and P600 might sometimes be an issue, in this specific dataset the wave-forms from the three conditions had already rejoined after the N400 and before the P600 time segment (see Fig. 1 in Delogu et al. (2019)). Because there was a large N400 in the event-unrelated violation condition, subtracting this negativity from the P600 time segment indeed resulted in a large P600. Only after this major change of the data, the data fit the model's predictions.

language even without extensive training because the hierarchical structures implementing the deep learning networks might resemble similar neural mechanisms implemented by the cortical regions under analysis. We cannot conclude, based on our observations, that the architecture of the SG model alone – i.e. without training on a cognitively plausible task – is enough to approximate the electrophysiological processes under scrutiny. This partially diverges from Schrimpf et al. (2021)'s position. Nonetheless, we think it is important to stress that the difference between their and our conclusions might simply be due to the fact that the most successful models in their study were implemented using architectures – such as the multi-head attention mechanism (Vaswani et al., 2017) – which were not used for the present iteration of the SG model. Moreover, instead of analyzing fMRI and ECoG data, we focused on EEG activity and in particular on the amplitude of the N400. Therefore, it could also be the case that our results simply reiterate the fact that the N400 ERP component's behavior evolves during an individual's experience of the statistical regularities of their environment, paralleled by the activity of the Sentence Gestalt layer during training epochs. Importantly, our results also emphasize that the implicit semantic prediction error reflected in N400 amplitudes inherently depends on the statistics of the environment as the predictions formed by the model (and presumably by human comprehenders) are generated based on the experience of these statistics.

Previously, Rabovsky et al. (2018) also investigated the effect of training on the SG model's ability to simulate N400 amplitudes, specifically during the processing of sentences containing semantically incongruent nouns. They reported that the SU shows at first an increase and later a decrease with additional training. These results are in line with the variation of the N400 during human language acquisition, which also first increases and then decreases across development (Friedrich & Friederici, 2004; Atchley et al., 2006; Kutas & Iragui, 1998). Moreover, Rabovsky et al. (2018) observed that the output layer activation approximates more and more the probability distributions embodied in the training corpus. This second point is in line with our results in confirming the role of training in improving the fit between a computational model and human data, mediated by the fit to the statistics of the world.

5 Conclusions

In the present study, we quantitatively investigated the hypothesis that mechanisms underpinning the N400 ERP component might be related to the update of an implicit predictive representation of sentence meaning during language comprehension, and quantitatively compared it to the alternative hypothesis that the component is instead signaling lexical prediction error. These quantitative investigations were afforded by training an artificial neural network model of sentence comprehension, called the Sentence Gestalt (SG) model (McClelland et al., 1989) on a large scale naturalistic corpus, and directly comparing the update of the distributed predictive representation of sentence meaning after each incoming word (Semantic Update, or SU) with the EEG recordings obtained during sentence reading. Our work corroborates and crucially extends previous studies which have shown that the Semantic Update of an SG model trained on a small synthetic environment responds similarly to the N400 amplitude to a series of lexical semantic manipulations, including semantic congruity, cloze probability, semantic and associative priming, and repetition, among others (Rabovsky et al., 2018). The small scale training prevented a direct assessment of the similarity between electrophysiological data and the model's dynamics as presented in the current study. In addition, our approach allowed to quantitatively compare the effects of the SG model's update on the electrophysiological data with the effect of lexical Surprisal estimated by a language model trained on the same data.

The analyses reported in this paper demonstrated that there is a significant relationship between the amplitude of the N400 component and SU recorded from the SG model (Tab. 1) and that SU predicted N400 amplitudes overall better than Surprisal (see Section 3.2). In addition, we observed that the effect of SU on N400 amplitudes remains significant after controlling for Surprisal, and vice versa. This seems to indicate that the activity corresponding to the N400 component might not correspond to a single monolithic process but might rather reflect at least two distinct sub-processes at the level of lexical and sentence meaning prediction error.

Acknowledgments

The research presented in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG) (B09 318763901 - SFB1294) awarded to Sebastian Reich and Milena Rabovsky, and the Emmy Noether grant (RA 2715/2-1) awarded to Milena Rabovsky.

References

Atchley, R. A., Rice, M. L., Betz, S. K., Kwasny, K. M., Sereno, J. A., & Jongman, A. (2006). A comparison of semantic and syntactic event related potentials generated by children and adults. *Brain and Language*, *99*, 236-246.

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), (pp. 112–118). Cognitive Science Society: Austin, TX.

Bar, M. (2011). *Predictions in the brain: using our past to generate a future*. Oxford University Press.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343 – 355. doi: https://doi.org/10.1016/0013-4694(85)90008-2

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi: 10.1162/tacl_a_00051

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020, 03). Going the Extra Mile: Effects of Discourse Context on Two Late Positivities During Language Comprehension. *Neurobiology of Language*, *1*(1), 135-160. Retrieved from `https://doi.org/10.1162/nol_a_00006` doi: 10.1162/nol$_{a0}$0006

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

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

Brown, C. M., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*, 34-44.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*. doi: 10.3389/fnhum.2010.00025

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, *26*(5), 609-651. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0364021302000794`

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, *162*, 153-166.

Collobert, R. (2011). Deep learning for efficient discriminative parsing. In *International conference on artificial intelligence and statistics (AISTATS)*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011).

Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, *12*, 2493-2537.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*.

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-n400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *35*, 1044 - 1063.

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

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164-203.

Ferreira, F., Bailey, K., & Ferraro, V. (2002, 11). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11(1)*, 11–15.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15-52.

Frank, S. L., Monsalve, I., Thompson, R., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, *45*, 1182—1190. doi: 10.3758/s13428-012-0313-y

Frank, S. L., 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. doi: 10.1016/j.bandl.2014.10.006

Friederici, A. D., Mecklinger, A., Spencer, K. M., Steinhauer, K., & Donchin, E. (2001). Syntactic parsing preferences and their on-line revisions: a spatio-temporal analysis of event-related brain potentials. *Brain research. Cognitive brain research*, *11 2*, 305-23.

Friedrich, M., & Friederici, A. D. (2004). N400-like semantic incongruity effect in 19-month-olds: Processing known words in picture contexts. *Journal of Cognitive Neuroscience*, *16*, 1465-1477.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 815 - 836.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385-390. doi: 10.1111/j.1467-9280.1993.tb00586.x

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2020/12/03/2020.12.03.410399` doi: 10.1101/2020.12.03.410399

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. Retrieved from https://doi.org/10.1162/neco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, *109 4*, 679-709.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall.

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*, 98-110.

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

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Kuperberg, G. R., Brothers, T., & 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*, 12-35.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117-129. Retrieved from https://www.sciencedirect.com/science/article/pii/S0926641003000867 doi: https://doi.org/10.1016/S0926-6410(03)00086-7

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

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. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading between the lines: Event-related brain potentials during natural sentence processing. *Brain and Language*, *11*(2), 354 – 373. doi: https://doi.org/10.1016/0093-934X(80)90133-9

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

Kutas, M., & Iragui, V. J. (1998). The N400 in a semantic categorization task across 6 decades. *Electroencephalography and clinical neurophysiology*, *108 5*, 456-71.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*, 271-281.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*, 920-933.

Lopopolo, A., Frank, S., van den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PLoS ONE*, *12*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc.

McClelland, J. L., St. John, M. F., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, *4*, 287–335.

Merkx, D., & Frank, S. (2020). Comparing transformers and rnns on predicting human sentence processing data. *ArXiv*, *abs/2005.09471*.

Michaelov, J., & Bergen, B. (2020, November). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th conference on computational natural language learning* (pp. 652–663). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.conll-1.53` doi: 10.18653/v1/2020.conll-1.53

Michaelov, J., Bergen, B., & Coulson, S. (2022, 05). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*. doi: 10.1109/TCDS.2022.3176783

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., ... 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. Retrieved from `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2018.0522` doi: 10.1098/rstb.2018.0522

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th international workshop on parsing technologies (IWPT 03)* (pp. 149–160).

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106. doi: 10.1162/0891201053630264

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400 neural response. In *Proceedings of the australasian language technology association workshop 2011* (pp. 38 – 46). Canberra, Australia.

Petten, C. V., & Luka, B. J. (2012a). 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.

Petten, C. V., & Luka, B. J. (2012b). 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.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, *143*, 107466. doi: https://doi.org/10.1016/j.neuropsychologia.2020.107466

Rabovsky, M., Hansen, S., & McClelland, J. L. (2016, 08). N400 amplitudes reflect change in a probabilistic representation of meaning: Evidence from a connectionist model. In (pp. 2045–2050). Cognitive Science Society: Austin, TX.

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

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190313. doi: 10.1098/rstb.2019.0313

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*, 68-89.

Rabovsky, M., Álvarez, C. J., Hohlfeld, A., & Sommer, W. (2008). Is lexical access autonomous? evidence from combining overlapping tasks with recording event-related brain potentials. *Brain Research*, *1222*, 156 - 165. doi: https://doi.org/10.1016/j.brainres.2008.05.066

Sayeed, A., Shkadzko, P., & Demberg, V. (2018). *Rollenwechsel-English: a large-scale semantic role corpus.* European Language Resources Association. doi: http://dx.doi.org/10.22028/D291-30972

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45). doi: 10.1073/pnas.2105646118

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and n400 erp components. *International Journal of Psychophysiology*, *83*(3), 382-392. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167876011003862` doi: https://doi.org/10.1016/j.ijpsycho.2011.12.007

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An erp study of p600 effects elicited by semantic anomalies. *Brain research. Cognitive brain research*, *22 2*, 241-55.

van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related potentials. *Memory & cognition*, *18*, 380-93. doi: 10.3758/BF03197127

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Willems, R. M., Frank, S., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral cortex*, *26 6*, 2506-2516.