UNIVERSITÄT ZU LÜBECK

Aus dem Institut für Psychologie I
der Universität zu Lübeck
Direktor: Prof. Dr. Nico Bunzeck

Artificial and biological neural networks of semantic prediction in speech comprehension

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

Aus der Sektion Naturwissenschaften

vorgelegt von
Lea-Maria Schmitt
aus Preetz

Lübeck, 2021

1. Berichterstatter: Prof. Dr. Jonas Obleser

2. Berichterstatter: Prof. Dr. Stefan Kiebel

Tag der mündlichen Prüfung: 11. Juni 2021

Zum Druck genehmigt. Lübeck, den 14. Juni 2021

Artificial and biological neural networks of semantic prediction in speech comprehension

Acknowledgements

I thank my supervisor JONAS OBLESER for giving me the freedom to follow my scientific interests and for guiding me when going astray, for challenging me to go beyond my limits and encouraging me to let go; my Dissertation Committee: STEFAN KIEBEL and MATTIAS HEINRICH; my colleagues in the Auditory Cognition group for making science fun with their dedication, ingenuity and camaraderie: MOHSEN ALAVASH, FELIX DEILMANN, JULIA ERB, LORENZ FIEDLER, ANNE HERRMANN, MARKUS KEMPER, JENNIFER KLOTKE, JENS KREITEWOLF, TROBY LUI, HONG-VIET VICTOR NGO, MARTIN ORF, MICHAEL PLÖCHL, SEBASTIAN PUSCHMANN, FRANZISKA SCHARATA, SARAH TUNE, LEONHARD WASCHKE, and MALTE WÖSTMANN, as well as all STUDENT RESEARCH ASSISTANTS and administrators TANJA KRUSE and CHRISTA MARX; my collaborators from the Max Planck Institute for Human Cognitive and Brain Sciences: GESA HARTWIGSEN for putting things into perspective with her pragmatism and rigour, but also ANNA RYSOP for taking the first steps in science together with me and opening her home to me during my visits in Leipzig; URI HASSON and his lab for welcoming me to Princeton as a visiting graduate student and introducing me to the magic of artificial neural networks; my friends and family for having my back: This is a shout out to my parents BARBARA and THOMAS, my sister RUTH, my LENNARD, as well as SIMON, ANNALINA, KAY, MAX, and ANDREAS.

Lea-Maria Schmitt, January 2021

Contents

1 General introduction

There are hundreds of thousands of words at my fingertips to write an opening line never read before. Yet, were you surprised to read any of the words eventually strung into this sentence? If not, one explanation might be that upcoming words met the expectations raised by preceding sentence context. The formation of such expectations, or predictions, is thought to be a fundamental principle in the nervous system. An ongoing effort in the field of neuroscience is directed towards the identification of the neural computations that enable predictive processing. In the present thesis, I leverage the human faculty of speech comprehension to investigate the neural underpinnings and behavioural consequences of making predictions. With speech evolving over time, the constraints posed by context are subject to constant alteration and call for the dynamic adaptation of biological neural networks in the brain to changing task demands. The aim of this thesis is to advance the toolbox available to study speech prediction and to foster our understanding of the functional interplay within and between biological neural networks of speech prediction.

1.1 Prediction in language comprehension

What is a prediction?—a weather forecast, a sports betting, a horoscope? In everyday language, we usually refer to a prediction as a guess on what will happen in the future. Let us stay with the example of a meteorologist for a moment to formalize the basic steps of making a prediction. The first step towards making an accurate weather forecast is to gather as much relevant information on the current weather situation as possible. For example, today is a rainy and windy day in early December. In a next step, the meteorologist will evaluate this context based on a set of rules she has acquired over the years. For example, one rule might be that most days at this time of a year were snowy since records began, another one might be that a rainy day is rarely followed by a snowy day. This internal model of weather development can be used to link the past (context) to the future (prediction). Finally, the meteorologist will announce in the evening news that it is not very likely that there will be snow the next day. The same logic applies to the prediction of speech, where we can use the previous context and our knowledge about the regularities and dependencies of language to inform our predictions on upcoming speech.

Psycholinguists have debated for decades whether humans make predictions during language comprehension or not. What have we learned already about these predictive processes and which questions remain unanswered so far?

Are humans exploiting the contextual constraints of language to make predictions?

One of the earliest experimental works on language prediction date back to the origins of information theory in the middle of the 20th century. While primarily concerned with the efficient and reliable transmission of information in telecommunication systems, Shannon (1948) laid, rather incidentally, the foundation for quantifying the information content and with that the predictability of (written or spoken) language.

In a series of experiments, Shannon (1951) instructed participants to guess the next letter in a text excerpt (e.g., “The room was not very”) and moved on to the next letter as soon as their guess was correct (letter “l” of word “light”). These experiments demonstrated that participants needed less guesses to name the correct letter when more context, here up to 100 letters, was available to inform the prediction. The guessing paradigm was widely adopted in linguistics and longer sentence context was shown to also help participants to guess the correct continuation on the level of words (Aborn et al., 1959). These studies were the first to show systematically that humans are capable of making predictions on upcoming language and that these predictions are not exclusively informed by local but also more global dependencies.

The predictions of humans on upcoming language are not always perfect. One limiting factor is the constraint posed by preceding context: Tulving and Gold (1963) showed that the time a target word has to be visually presented before participants are able to correctly report it decreased when preceding context was longer, but only when context was actually relevant to the target word. Consistently, participants were faster to classify words in a lexical decision task when presenting them in sentences with high as opposed to low constraint (Schuberth et al., 1981). For example, “She made the bed with new” provides rich context for the prediction of the final word “sheets”, whereas the context “We are very pleased with the new” does not (see Erb et al., 2012).

Are humans naturally making predictions in language comprehension?

The traditional experimental paradigms used to study language prediction either directly ask participants to make a prediction or implicitly push participants to make use of contextual constraints to give faster responses. A line of research that offers a more natural approach to language prediction uses eye tracking while participants read a text, a task that we encounter regularly in everyday life.

When a word is highly predictable from its context, readers not only spend less time fixating that word but are also more likely to completely skip the fixation of that word (Ehrlich & Rayner, 1981). On the one hand, this reading behaviour suggests that participants make natural use of contextual constraints. On the other hand, it leaves open whether the constraint is actually used to make a prediction. The skipping effect is thought to arise from parafoveal preview, that is, word information is available to readers already during the saccade into the word (Schotter et al., 2012). The critical (and open) question is whether this parafoveal word information is used to compare a prediction to the actual next word or to simply integrate more predictable word information into current context more easily (see e.g., Balota et al., 1985).

What are the costs of making a wrong prediction on upcoming language?

A frequent objection to predictive processing in language comprehension is that even the most probable continuation is oftentimes not the correct one and should therefore slow down response times (for review, see Van Petten & Luka, 2012). This concern has not proven to be true.

Indeed, words more congruent with constraining sentence context are associated with faster responses in lexical decision tasks (Schuberth & Eimas, 1977; Forster, 1981), also when related only loosely (Kleiman, 1980). Additionally, reading times for words violating contextual constraints were just as long as for words in an unconstrained, neutral context (Frisson et al., 2017). This absence of negative behavioural effects for incorrect predictions is also reflected in skipping rates (Luke & Christianson, 2016) and later eye movements (Frisson et al., 2017).

On the neural level, an eminent marker of speech predictability in electroencephalography (EEG) is the N400 component, which is sensitive to the disconfirmation of predictions. In line with behavioural results, the N400 amplitude gradually increases when a word is less related to the most probable prediction (Kutas & Hillyard, 1984; Federmeier & Kutas, 1999; DeLong et al., 2005). This suggests that language prediction is probabilistic (e.g., Jurafsky, 1996; Norris & McQueen, 2008), with all (or most) words in someone's vocabulary activated according to their probability of being the next word instead of a potentially costly all-or-nothing activation for only the most probable word (Brothers & Kuperberg, 2021).

What are the benefits of making predictions on upcoming speech outside the laboratory?

The prediction of upcoming speech is thought to be a fundamental principle in human communication. What is the functional role of language prediction?

A prominent example, directly relevant to the present thesis, is the comprehension of speech under adverse conditions. While laboratory conditions allow presenting clearly enunciated speech in a quiet environment, intelligibility of natural everyday speech is often severely reduced (Mattys & Liss, 2008). According to Mattys and colleagues (2012), there are three sources of disrupted speech comprehension: the speaker (e.g., foreign accent), the speech signal (e.g., competing background noise), and the listener (e.g., hearing impairment). A compensatory factor under such adverse conditions is the availability of context in the form of acoustic, linguistic, pragmatic or multisensory information (Obleser, 2014). In particular, the presentation of sentences with a final keyword predictable from its preceding semantic context facilitates comprehension when speech is acoustically degraded, both in terms of accuracy (Kalikow et al., 1977) and response speed (Golestani et al., 2013). This semantic predictability gain has also been found when disruption was not limited to the speech signal but extended to the listener, that is, in participants of older age (Pichora-Fuller et al., 1995) or with hearing impairment (Bilger et al., 1984; Winn, 2016; Holmes et al., 2018).

Is timing everything when studying language prediction?

Together, previous studies on language prediction showed that (1) contextual constraint in general and timescale-specific constraint in particular facilitate comprehension, (2) contextual constraint is naturally exploited in comprehension, (3) language prediction is probabilistic, (4) violated predictions do not deteriorate comprehension, and (5) contextual constraint benefits comprehension under challenging listening conditions. This highlights the ubiquity of predictive processing in language comprehension.

The ultimate evidence for predictive processing is to demonstrate the presence of a prediction before the actual *on*set of a word. First evidence in this direction came from studies showing that the N400 component occurs before the *off*set of a word (McCallum et al., 1984; Holcomb & Neville, 1991), and before the isolation point allowing to distinguish a word from other candidates (Van Petten et al., 1999). More direct evidence came from a series of studies showing that articles (Wicha et al., 2004) or gender-marked adjectives (van Berkum et al., 2005) evoke an enhanced positive deflection in EEG when they are incongruent with the ensuing noun, thereby indicating that participants had already predicted the noun before onset of the article. However, Nieuwland and colleagues (2018) were not able to replicate a similar study on the N400 (DeLong et al., 2005) in a large multi-lab endeavor (but for a response to the critique, see DeLong et al., 2017).

Beyond evidence from event-related potentials, a spatial representational similarity analysis in magnetoencephalography showed more similar responses already before the onset of a word when different contexts constrained the same in comparison to a different prediction (Wang et al., 2018). A limitation of this study is that different contexts constraining the same prediction nevertheless appeared to be semantically more similar to one another than sentences constraining a different prediction. Therefore, we cannot rule out that not word prediction but semantic similarity between contexts drives the observed effect.

Another recent study showed that semantic features of an upcoming word can be decoded from high gamma activity in electrocorticography before word onset (Goldstein et al., 2020). Surprisingly, the effect was not tested as a function of word predictability but across all words presented in a story. This suggest that also other, confounding effects might account for the early semantic response, such as temporal smearing of responses due to filtering, misalignment of defined word onsets or similarity of words to their preceding context.

In sum, even though most studies have some limitations that make it hard to interpret results in terms of predictive processing, the consistency in finding results that are compatible with predictive processing is remarkable. The studies gathered in this thesis are no different in this respect. Here, I will report two functional magnetic resonance imaging (fMRI) studies, which track a notoriously sluggish signal at a low temporal resolution. While this makes it impossible to trace changes in activation back to time intervals before vs. after word onset, this approach has merit for two reasons. First, fMRI allows to investigate the dynamics within and between biological neural networks during language prediction in great spatial detail. Second, not being able to recover the exact timing of effects does not preclude that these effects arise from predictive processes. For example, prediction error responses are directly tied to the prediction but become evident only after the prediction has proven correct or incorrect (Blank & Davis, 2016; Kandylaki et al., 2016).

Theoretical accounts of language prediction

The prediction of language takes effect on multiple representational levels. On the one hand, context in many forms facilitates language comprehension. For example, comprehension is facilitated by constraining context on the level of phonology (Farmer et al., 2006), prosody (Kurumada et al., 2014), syntax (Altmann & Kamide, 2007) and discourse (Xiang & Kuperberg, 2015). On the other hand, context facilitates language comprehension on many levels. For example, constraining context facilitates comprehension on the level of phonology (Allopenna et al., 1998), syntax (Rohde et al., 2011), and semantics (Metusalem et al., 2012). This suggests that the brain makes use of context represented at different timescales and, in turn, makes predictions at different timescales.

One general account that combines multiple timescales of predictive processing is the family of predictive coding frameworks (Spratling, 2017). In its most basic sense, predictive coding describes how a generative model that maps causes to consequences is inverted to infer causes from consequences (Ramstead et al., 2020). For example, as a speaker we produce an auditory signal (i.e., consequence) that reflects the story we would like to tell (i.e., cause), while the listener infers that story from the auditory signal (Friston et al., 2020).

In this scenario, perception is an inference problem that is often cast in terms of Bayesian inference. In detail, the posterior probability of observing causes given the consequences is calculated by multiplying the prior probability of causes with the likelihood of consequences given the causes (Berger, 1985). Put differently, beliefs are updated based on sensory input and prior beliefs, with prior beliefs representing hypotheses or expectations about the world. Critically, these expectations must not necessarily be interpreted as temporal forecasts but rather describe the statistical extrapolation to any unseen data (de Lange et al., 2018).

While Bayesian inference explains behaviour, predictive coding explains its underlying neural responses (Aitchison & Lengyel, 2017). Accounts of predictive coding propose that inference is hierarchically organized in the brain (Rao & Ballard, 1999; Friston, 2005). Generally speaking, prior beliefs (or predictions) are thought to be fed back from higher-order areas, while the residual error in prediction is fed forward from lower-order areas (Bastos et al., 2012).

By extension, the predictive coding hierarchy has been suggested to engage also in temporal predictions (Kiebel et al., 2008). In line with such an approach of temporal prediction, predictive coding has been laid forward as a viable framework for the multi-scale prediction processes suggested to constitute speech comprehension (Bornkessel-Schlesewsky et al., 2015).

1.2 Artificial and biological neural networks of language prediction

While previous studies established that predictive processes are of behavioural and neural relevance, they did not characterize the neural network dynamics that give rise to these predictions. In the present thesis, I aim to approach these network dynamics from two angles. First, I will make use of artificial neural networks to model the computations underlying the formation of predictions in the brain. Second, I will investigate the dynamic interplay within and between large-scale brain networks in light of changing task demands.

Artificial neural networks of language prediction

When talking about the predictability of speech, most studies are talking about the predictability of speech in controlled experimental settings using isolated sentences. However, such stimuli with short context drastically underestimate the challenges listeners face in everyday communication. Imagine yourself sitting in on a lecture and context piles up word by word until someone in the audience asks a question about a topic discussed a while ago. To make an accurate prediction about the end of the question, a representation of long-term context is needed.

Until recently, it has been technically difficult to study effects of long-term context. In traditional cloze probability procedures (Taylor, 1953), participants fill in the blanks of a text and the proportion of correct word predictions is calculated across participants. As deriving word-by-word cloze probabilities for long texts takes a lot of time, this procedure is rarely used to investigate effects of long-term context (but see Goldstein et al., 2020). Beyond the cloze procedure, other language models assigning probabilities to words suffer from the problem that they are limited to short context per se. For example, relative frequency counts (or n-grams) are based on counting the occurrence of specific word combinations in a large corpus of text. The limitation of such an approach becomes apparent when considering that we can easily create new combinations of words so rare that they do not occur in text corpora but are still predictable from their context.

With the advent of artificial neural networks in machine learning (Schmidhuber, 2015), it has become more straightforward to derive word probabilities based on complex context (for a detailed description, see Jurafsky & Martin, in preparation). Three criteria make an artificial neural network a powerful language model. First, the network is trained on the objective to make a prediction on the next word, so that we can derive word probabilities. Second, the network is trained on a large corpus of text, so that it can learn the dependencies of speech. Third, the network operates on multiple layers, so that it can represent context at different timescales.

The inputs to most artificial neural networks in natural language processing are multidimensional vector representations of words (e.g., Mikolov et al., 2013). A unique vector defines each word in our vocabulary, with single dimensions coding for how strongly the word loads on a linguistic feature. For example, one dimension might represent animate vs. inanimate objects. The vectors (or embeddings) are either trained within the current artificial neural network or derived from a pre-trained network. In a previous study, Huth and colleagues (2016) demonstrated that the semantic features of word vectors map onto a "semantic cortical atlas", with specific semantic features represented in dedicated cortical regions.

An artificial neural network in its most shallow form has three layers: an input layer, a hidden layer, and an output layer. When stacking multiple hidden layers, the artificial neural network becomes a deep network, allowing for computations that are more complex.

Put simply, a hidden layer is composed of many units (or neurons). In a fully connected network, each unit takes as input the weighted output from all units of the previous layer and linearly combines these inputs. The combined input is passed through a non-linear activation function and becomes the input to the next layer. To provide the artificial neural network with a context memory, recurrent connections are implemented at hidden layers.

The outputs of most artificial neural networks in natural language processing are probability distributions assigning the probability of being the next word to each word in a large vocabulary of candidate words. These probability distributions can be used to derive information theoretic measures, which provide a proxy of predictability on the word level. For example, Frank and Willems (2017) derived the surprisal evoked by a word from such probability distributions and found left inferior temporal sulcus and bilateral posterior superior temporal gyrus to be sensitive to machine-derived surprisal.

Together with cognitive science and computational neuroscience, artificial neural networks have been proposed to constitute the discipline of cognitive computational neuroscience (Kriegeskorte & Douglas, 2018). Artificial neural networks come in the form of many different architectures, optimization functions and objective functions. This opens the possibility to test the explanatory power of different artificial neural networks for biological neural networks, ultimately pursuing the goal to understand the computational principles underlying processing in the brain (Richards et al., 2019). In this thesis, I will focus on the potency of artificial neural networks to elucidate the underpinnings of predictive language processing.

Biological neural networks of language prediction

No task that humans face is solved by only one brain region. Instead, multiple regions distributed across the entire brain interact within large-scale networks of cognitive processing (Bressler & Menon, 2010). There are a handful of such functional networks in the brain (Yeo et al., 2011), like the default mode network and the fronto-parietal control network (Power et al., 2011). Which of these large-scale networks are involved in language prediction?

An obvious candidate for predictive language processing is the language network, which includes superior and middle temporal gyrus, inferior frontal gyrus, premotor and inferior parietal cortex (Friederici & Gierhan, 2013) but also regions in medial prefrontal and ventral temporal cortex (Binder et al., 2009). Previous studies have shown that this network is sensitive to manipulations of speech predictability (Willems et al., 2016), especially when speech is acoustically degraded (Obleser et al., 2007; Obleser & Kotz, 2010; Golestani et al., 2013). The language network is a domain-specific network highly specialized for language-related tasks.

A subnetwork (or module) of the language network is the auditory dorsal pathway (Sporns & Betzel, 2016). This region has been shown to be sensitive to the different timescales of speech. In Study 2 and Study 4, we aimed to elucidate the computations along this pathway and its implication in speech prediction.

In contrast, domain-general networks are implicated in different tasks (Geranmayeh et al., 2014), representing core brain networks combining regions with different cognitive functions (Bressler & Menon, 2010). Domain-general and domain-specific networks are conceptualized to dynamically interact depending on task demand (Hartwigsen, 2018). When task demands are high, the allocation of additional resources from domain general networks may facilitate processing. When task demands are low, domain-specific networks are thought to be equipped with the computations most efficiently solving the task. In Study 1, we investigated interactions between domain-specific and domain-general networks depending on the predictability and intelligibility of speech.

1.3 Research questions

The overarching goal of this thesis was to shed light on the biological neural network dynamics that underlie the prediction of speech when comprehension is challenged by poor acoustics.

More specifically, the present thesis set out to answer four questions: (1) How do contextual constraints modulate the interplay between domain-general and domain-specific biological neural networks under challenging listening conditions to facilitate speech comprehension? (2) How can we derive a measure that represents the predictability of words at multiple timescales of context in natural speech? (3) Does the human brain predict speech along a temporo-parietal processing hierarchy at multiple timescales? (4) Does the temporo-parietal processing hierarchy draw on event-based context representations to make predictions?

In Study 1, we asked how the domain-specific language network and the domain-general cingulo-opercular network dynamically interact to adapt to the task demands posed by semantic predictability and acoustic intelligibility of speech. A special focus of this study was on the network dynamics that enable successful comprehension under challenging listening conditions by exploiting the contextual constraints of speech. We hypothesized that the recruitment of the language network facilitates comprehension at intermediate levels of intelligibility when semantic constraints are available to make predictions. In contrast, we hypothesized that the cingulo-opercular network is up-regulated when no such semantic constraints are available. During functional MRI, participants repeated isolated sentences, which varied at six levels of intelligibility (signal-to-noise ratio [SNR] of speech-shaped noise) and two levels of predictability (low vs. high).

In Study 2, we asked how the more complex contextual constraints of natural speech are exploited to facilitate speech comprehension in challenging listening scenarios. We recorded functional MRI while participants listened to a story embedded in resynthesized natural sounds (SNR of 0 dB). Our central hypothesis was that context is resolved into its timescales to inform predictions on upcoming words at multiple levels of abstraction, which we expected to be represented along a temporo-parietal processing hierarchy. We employed a measure of semantic similarity between each word in the story and its preceding words at different context lengths. To this aim, we represented words in a multi-dimensional vector space derived from pre-trained artificial neural networks and calculated the similarity between vectors of single words and average context vectors.

In Study 3, we developed a new measure to determine how informative context at different timescales is for the prediction of the next word. It was particularly important to us that this measure captured the complex temporal dependencies constituent of language. We trained two artificial neural networks to predict the next word in the story based on a context window of 500 preceding words. Both artificial neural networks operated on five layers (or timescales) and critically differed in one specific computational feature. One artificial neural network continuously integrated new word input into context representations at all timescales, whereas the other network updated context representations only at the boundary of events in the narrative. We read out the surprisal evoked by a word at specific timescales, yielding a measure of multi-scale word surprisal. This allowed us to narrow down differential effects of surprisal on activations recorded in Study 2 to the continuous vs. event-based updating rule of artificial neural networks.

In Study 4, we applied the surprisal measure developed in Study 3 to fMRI data acquired in Study 2. Here, we asked whether the organization of context along the auditory dorsal pathway enables successful prediction on upcoming words. We expected that surprisal evolves along a hierarchy, with longer timescales represented in more parietal regions of the auditory dorsal pathway. Additionally, we expected this hierarchy to be sensitive to event-based context representations.

2 General methods

This chapter gives an overview of the central methods used in the studies presented in the current thesis. More detailed information can be found in the method sections of respective studies.

2.1 Controlled vs. natural speech stimuli

All studies in the present thesis used language as a stimulus. Along the lines of classical language experiments, we performed a sentence-repetition task in Study 1. The sentences used in this experiment were constructed to vary along one specific dimension, that is, the predictability of the sentence-final word (Kalikow et al., 1977; Erb et al., 2012). The use of such isolated, low-dimensional sentences has the advantage that, in theory, effects resulting from contrasting sentences with low vs. high predictability can be attributed to the manipulation of predictability but not any other (confounding) variables.

However, there is also a number of concerns with such controlled stimuli. An obvious criticism is that results from controlled stimuli often do not generalize to other and especially more realistic stimuli. Two important considerations in this respect are that the human brain is seldom exposed to such low-dimensional speech and that by controlling for some dimensions we might introduce spurious relationships between other dimensions (Nastase et al., 2020).

In another, more natural language experiment (Study 2), participants listened to Herta Müller, a Nobel laureate in Literature, reminiscing about her childhood as part of the German-speaking minority in the Romanian Banat ("Die Nacht ist aus Tinte gemacht", 2009). To determine the *naturalness* of language perception in an experiment, Hamilton and Huth (2018) proposed to ask three questions:

First, is this a stimulus that a person might reasonably be exposed to outside of an experimental setting?

The recording of the story is available for purchase, so it is intended to appeal to at least a certain target audience. In addition, audiobooks are popular with many people who regularly listen to recorded stories. However, the speaker has a striking dialect that we are rarely exposed to in everyday life and most of the time we are engaged in a conversation when listening to speech rather than a one-hour monologue.

Second, does the stimulus appear in the same context as it would in real life?

It is certainly rare to listen to an audiobook while lying in an MRI scanner. Apart from this obvious contrast to everyday contexts, it is unusual to listen to an audiobook with such prominent and extremely variable background noise as realized in Study 2 (i.e., five-second snippets of resynthesized natural sounds at an SNR of 0 dB). Conforming to contexts of real life communication, the semantic context of the story is extremely rich and the speaker does not read the story to us but spontaneously narrates it.

Third, is the subject's motivation for perceiving and understanding the stimulus particular to the experimental setting, or is it a motivation that the subject would feel in real life?

While some participants were enthusiastic about the audiobook, others found the story difficult to understand. To keep all participants engaged, we interrupted the story every eight minutes and asked some questions about the content of the story, a rather uncommon motivation to carefully listen to an audiobook in real life.

In sum, this highlights that, even when using naturally narrated stories, ecological validity is not perfect. Some of the limitations directly arise from the goals of our research. For example, the story was embedded in background noise to emulate a challenging listening scenario in which listeners were more likely to make use of the semantic predictability of speech (Rysop et al., 2021). Another example is that we aimed to present speech with rich and naturally *in*consistent context, which also gives the speaker more freedom to tell the story in her own, possibly peculiar, way. This illustrates that the naturalness of speech stimuli varies along a continuum, with different research questions warranting the use of more or less natural speech stimuli.

2.2 Operationalization of word predictability[1]

When listening to a story, all spoken words $\{w_1, w_2, \dots, w_{p-1}\}$ form the context for the prediction of the upcoming word w_p. The predictability of a word can be described by means of different metrics.

Cloze probability

The cloze procedure is a simple technique to determine the predictability of a word by its preceding context (Taylor, 1953). In detail, participants read a text (e.g., "Please pass the") and complete missing words with the first word that comes to mind (e.g., "salt"). The cloze probability is calculated as the proportion of participants who correctly filled in the blank.

Surprisal

The surprisal of a word is the amount of information that cannot be explained away by its preceding context (Hale, 2016):

$$surprisal(p) = -\log P(w_p \middle| w_1, \dots, w_{p-1})$$

The surprisal associated with a word is high when a low probability of being the next word was assigned to that word.

Entropy

Given the probability of being the next word for each word in a large vocabulary, entropy reflects the amount of uncertainty across the whole probability distribution (Shannon, 1948):

$$entropy(p) = -\sum_{w_{p+1} \in W} P(w_{p+1} \middle| w_1, \dots, w_p)\ \log P(w_{p+1} \middle| w_1, \dots, w_p)$$

When high probabilities are assigned to only one or few words in the vocabulary, entropy is low. However, entropy is high when semantic context is not informative enough to narrow down predictions to a limited set of words, resulting in more similar probabilities across the vocabulary.

As all information necessary to determine the entropy of a word is already available to participants before word presentation, entropy of word $\boldsymbol{w_{p+1}}$ is ascribed to the previous word $\boldsymbol{w_p}$.

[1] This section was partly adapted from Schmitt et al. (2020).

Whereas word surprisal quantifies the availability of information on the actual next word, entropy quantifies the overall difficulty of making any definite prediction.

Similarity

The similarity of a word to its preceding context $A = \{w_i, w_j, \dots\}$ is calculated by correlating the vector representation of a word with the average vector representation of its preceding content words (Frank & Willems, 2017):

$$similarity(p) = corr(\sum_{w_i \in A} \overrightarrow{w_i}, \overrightarrow{w_p})$$

A higher positive correlation indicates that a word is more similar to its preceding context.

2.3 Functional magnetic resonance imaging

On the one hand, we aimed to investigate functional interactions between brain regions, which calls for a recording modality with a high spatial resolution. On the other hand, we aimed to investigate brain responses to speech stimuli naturally unfolding over time, which calls for a recording modality with a high temporal resolution. How can we reconcile the two apparently opposing demands in fMRI?

Acquisition of BOLD responses to speech

In MRI, three-dimensional structural images of the body are acquired by measuring the behaviour of protons in a strong magnetic field when applying a disturbance through an electromagnetic wave (for a detailed description, see e.g., Schild, 1990; Vlaardingerbroek & den Boer, 2003).

In practice, the participant lies in the bore of an MR scanner, which forms a strong external magnetic field around the head. Put simply, this causes the protons in the head to align parallel or anti-parallel to the magnetic field. As any resulting magnetic field is longitudinal to the external magnetic field, it cannot be measured. To overcome this problem, an electromagnetic wave (or radio frequency pulse) is applied. This has two important effects: First, some protons are excited into a non-parallel state of higher energy, thereby decreasing longitudinal magnetization. Second, protons get to spin in phase (i.e., resonance), resulting in transversal magnetization.

When turning the radio frequency pulse off, protons return to their original low-energy state and fall off phase. This relaxation increases longitudinal magnetization and decreases transversal magnetization. As both processes bring shifts in energy, a coil can measure them. The time it takes the protons to return to their initial longitudinal (i.e., T1) and transversal magnetization (i.e., T2) depends on the surrounding chemical environment (or tissue), so that relaxation times can be used to distinguish between different types of tissue.

In fMRI, we can study localized brain activity over the time course of an experiment (for a detailed description, see e.g., Logothetis, 2008). An established functional imaging method is the blood-oxygen-level-dependent (BOLD) contrast, which measures blood oxygenation. More precisely, neural energy consumption in response to a stimulus initially reduces the blood oxygen level. However, the brain will compensate for higher energy demands by increasing the regional cerebral blood flow (i.e., neurovascular coupling), resulting in an increase of blood oxygenation within a couple of seconds. As T2 becomes longer for deoxygenated hemoglobin, we can distinguish between oxygenated and deoxygenated blood. The hemodynamic response is the temporal response profile of oxygenation to a stimulus.

An fMRI recording yields a four-dimensional image. In comparison to neurophysiological recordings, the spatial resolution of functional data in Study 1 and Study 2 was high (voxel size = 2.5 × 2.5 × 2.5 mm), while the temporal resolution was low (TR = 947 and 2,000 ms).

Functional MRI capitalizes on an indirect link from blood oxygenation to neural activity, so we rather infer than measure neural activity. This link has been substantiated by studies showing that the BOLD signal correlates with neural activity (Logothetis et al., 2001; Mukamel, 2005) and, more specifically, is proportional to the square root of neural activity (Bao et al., 2015). As for the exact electrophysiological components, the BOLD signal correlates more strongly with local field potentials reflective of peri-synaptic activity than with spiking activity (Ekstrom, 2010), suggesting that it is indicative of local rather than long-range signal transmission in the brain (Logothetis & Wandell, 2004). In particular, the BOLD signal reflects asynchronous neural responses manifesting in broadband oscillatory power rather than synchronous neural responses manifesting in narrowband gamma power (Hermes et al., 2017).

Analysis of BOLD responses to speech

One concern when studying speech comprehension with fMRI is that pulse sequences produce acoustic noise potentially interfering with the speech signal. Indeed, scanner noise reduces the sensitivity to stimuli in auditory cortex and increases demands for executive and attentional processing (Peelle, 2014).

In traditional experiments with single trials, it is oftentimes possible to sparsely sample only specific time points in a trial (Hall et al., 1999), a technique often used in speech studies (e.g., Rodd et al., 2005; Peelle, 2014; Scharinger et al., 2016). For example, Erb and colleagues (2013) measured BOLD activity after single-trial sentence presentation and response collection, so that the auditory signal was not affected by scanner noise. Although BOLD activity was sampled approximately 3.5 s after speech offset, activity in auditory cortex indicated that sparse sampling still captured effects of auditory stimulation. This highlights that BOLD responses to speech can be acquired without interference with the auditory stimulus. However, this approach comes at two costs. First, the number of volumes recorded per condition is small, thereby limiting reliability of results. Second, such an approach requires suspending stimulus presentation at regular intervals and therefore is not applicable when interested in responses to continuous natural speech on a word-by-word basis.

When comparing sparse with continuous scanning, previous studies showed that responses to tones in monkey auditory cortex (Petkov et al., 2009) and to speech in human superior temporal plane (Schmidt et al., 2008) were stronger with sparse imaging. Nevertheless, continuous fMRI sequences introduce a highly regular, energetic masker that is readily accepted by researchers during speech presentation in order to achieve a higher temporal resolution (e.g., Keller, 2001; Zempleni et al., 2007; Clos et al., 2014).

Another concern when studying speech comprehension with fMRI is that speech rapidly evolves over time, while the BOLD signal is sampled at a temporal resolution too low to capture these dynamics. Therefore, a common approach in continuously sampled fMRI studies is to manipulate whole sentences along some feature dimension and then contrast these experimental conditions. For example, Rogalsky (2008) continuously measured BOLD activity while participants judged the plausibility of sentences in a 2 (sentence type) × 3 (secondary task) design. The authors created regressors by coding for the sentence presentation times of each experimental condition and convolving these time courses with the hemodynamic response function. Finally, these regressors were submitted to a multiple linear regression analysis on the participant level and resulting voxel-wise coefficients of single conditions were compared via *t*-tests on the group level.

One potential limitation of this approach is that it largely dismisses any temporal dynamics of speech arising on short scales like words or phonemes. However, there are also studies in which the temporal dynamics of speech are of minor relevance. For example, in Study 1 we used sentences that were embedded in speech-shaped noise and either had a final word of low or high predictability. The SNR across single sentences was fixed, so that interactions with speech processing were stable over the sentence. Further, sentences were designed in a way that predictability either constantly increased or remained stable over the time course of a sentence. This implies that neural responses in a sentence had the same direction (e.g., relative increase of BOLD response to all words), so that they were treated as one response. However, in natural speech paradigms, I would expect predictability to suddenly raise or drop when transitioning from one word to the next, so responses must be modelled separately.

A similar approach can be used to analyse responses to natural speech, like in Study 2 and Study 4. However, there are two crucial differences: (1) Neural responses are modelled on the level of single words. This is either done by placing box-car functions on word presentation times or by placing stick functions on word onsets. (2) As we do not have separate experimental conditions in natural speech but variables change their value on a word-by-word basis, another difference is that box-car or stick functions are scaled to word-specific values of these variables. For example, Willems and colleagues (2016) derived word surprisal and entropy of a story presented to participants and modelled the duration of each word as a box-car function scaled to word-specific surprisal or entropy, respectively. After convolving surprisal and entropy time courses with the hemodynamic response function, an encoding model can be estimated to derive voxel-specific coefficients for surprisal and entropy (Naselaris et al., 2011).

This highlights that, depending on what we aim to model, we can adjust our methods to either model BOLD time series on a coarse temporal scale like sentences or on a more fine-grained temporal scale like words.

3 Experimental results

3.1 Neural modelling of the semantic predictability gain under challenging listening conditions[2]

3.1.1 Abstract

When speech intelligibility is reduced, listeners exploit constraints posed by semantic context to facilitate comprehension. The left angular gyrus (AG) has been argued to drive this semantic predictability gain. Taking a network perspective, we ask how the connectivity within language-specific and domain-general networks flexibly adapts to the predictability and intelligibility of speech. During continuous functional magnetic resonance imaging (fMRI), participants repeated sentences, which varied in semantic predictability of the final word and in acoustic intelligibility. At the neural level, highly predictable sentences led to stronger activation of left-hemispheric semantic regions including subregions of the AG (PGa, PGp) and posterior middle temporal gyrus when speech became more intelligible. The behavioural predictability gain of single participants mapped onto the same regions but was complemented by increased activity in frontal and medial regions. Effective connectivity from PGa to PGp increased for more intelligible sentences. In contrast, inhibitory influence from pre-supplementary motor area to left insula was strongest when predictability and intelligibility of sentences were either lowest or highest. This interactive effect was negatively correlated with behavioural predictability gain. Together, these results suggest that successful comprehension in noisy listening conditions relies on an interplay of semantic regions and concurrent inhibition of cognitive control regions when semantic cues are available.

[2] This section was adapted from Rysop et al. (2021).

3.1.2 Introduction

In everyday life, we are remarkably successful in following a conversation even when background noise degrades the speech signal. One important strategy facilitating comprehension in demanding listening scenarios is the prediction of upcoming speech. For example, the sentence fragment “She made the bed with new” provides rich semantic context to inform an accurate prediction of the sentence continuation “sheets”. However, poor semantic context like “We are very pleased with the new” provides little information to build up the prediction of the word “sheets”. This *semantic predictability gain* is a well-documented behavioural phenomenon. It can be observed in healthy young and older listeners (Obleser et al., 2007; Obleser & Kotz, 2010; Vaden et al., 2013, 2015), in hearing-impaired listeners (Holmes et al., 2018), as well as in cochlear implant users (Winn, 2016).

To study the interactive effects of intelligibility and semantic predictability in the listening brain, previous neuroimaging studies used sentences with different occurrence probabilities of the final word while varying the spectral resolution of speech using noise-vocoding (Shannon et al., 1995). In these studies, the left AG has been identified as a key region sensitive to strong semantic constraints in degraded speech (Obleser et al., 2007; Obleser & Kotz, 2010; Golestani et al., 2013). When highly predictable speech was degraded to a moderate extent, activity in AG increased. However, when degradation rendered speech either easily or, if it all, hardly comprehensible, activity in AG dropped irrespective of semantic constraints. Finally, Hartwigsen and colleagues (2015) used focal perturbations induced by transcranial magnetic stimulation to demonstrate that the ventral portion of left AG is functionally relevant for the behavioural predictability gain. Together, these studies provide converging evidence for a role of the left AG in the top-down use of semantic information when comprehension is challenged by degraded speech signals.

The AG is considered a heterogeneous region and has structural connections to language-specific fronto-temporal regions as well as domain-general cingulo-opercular regions (Binder et al., 2009; Seghier, 2013). Based on its cytoarchitectonics, AG can be subdivided into an anterior part (PGa) and a posterior part (PGp; Caspers et al., 2006, 2008). These subregions have been associated with different functional roles: PGa has been associated with the automatic and domain-general allocation of goal-directed attention and episodic memory. In contrast, PGp was (inconsistently) linked to more controlled and complex semantic-specific processes irrespective of control demands (Noonan et al., 2013; Bonnici et al., 2016; Bzdok et al., 2016; Jung et al., 2017).

These findings indicate that the AG might play a central role in language-specific but also in domain-general cognitive operations. Yet, the functional interplay of PGa and PGp within left AG as well as their differential roles in the processing of semantic context in degraded speech are unknown.

As comprehension of degraded speech poses a challenging task on listeners, the brain recruits additional resources beyond the semantic network (Peelle, 2018). These additional resources are thought to be provided by domain-general regions that typically do not show a functional specialization, such as the cingulo-opercular network. The cingulo-opercular network has been implicated in adaptive control processes subserving the flexible allocation of attentional resources and typically comprises the bilateral frontal operculae and adjacent anterior insulae, as well as the dorsal anterior cingulate cortex extending into the pre-supplementary motor area (pre-SMA; Dosenbach et al., 2008; Duncan, 2010; Camilleri et al., 2018). Indeed, previous studies on degraded speech processing have found not only fronto-temporo-parietal regions of the semantic network (Adank, 2012; Alain et al., 2018) but also regions of the cingulo-opercular network (Erb et al., 2013; Vaden et al., 2015, 2017). Additionally, using a word recognition task under challenging listening conditions, Vaden and colleagues (2013) demonstrated that the magnitude of activation within the cingulo-opercular network predicted the word recognition success in the following trial. This upregulation might reflect the activation of additional cognitive resources when no external linguistic cues are available (Peelle, 2018). In contrast, the recruitment of the semantic network is thought to improve comprehension most effectively when semantic context is available. Here, we aim to identify network-level descriptors of the semantic predictability gain to comprehension and investigate how changing task demands shape the adaptive interplay of language-specific and domain-general resources in successful speech comprehension.

Previous studies manipulating predictability and intelligibility of speech suffered from small samples and mainly relied on few levels of noise fixed across participants. Moreover, most of these studies used sparse temporal sampling, precluding the investigation of effective connectivity. Consequently, it remains unclear how the identified regions influence each other during successful comprehension of speech in noise. We aimed to overcome some of the previous shortcomings by performing a continuous event-related functional magnetic resonance imaging (fMRI) experiment with six individualized, varying levels of intelligibility, ranging from largely unintelligible to easily intelligible. Thereby, we accounted for individual differences in auditory perception and assured comparable task performance across participants.

Based on previous studies, we expected a facilitatory behavioural effect of semantic predictability on speech comprehension, which should be strongest at intermediate levels of intelligibility (i.e., when the auditory signal is somewhat degraded but still intelligible) and least pronounced for unintelligible and intelligible sentences. We hypothesized to find increased activation in left AG and other semantic regions as a neural correlate of the semantic predictability gain at intermediate levels of intelligibility. This interaction should be driven by increased effective connectivity between the AG and other semantic regions. Moreover, we expected increasing activity and within-network connectivity in the cingulo-opercular network to improve comprehension in challenging listening conditions (i.e., intermediate intelligibility and low predictability).

As a main finding of our study, highly predictable sentences at intermediate intelligibility levels drove activity in core regions of the semantic network and the behavioural comprehension gain was correlated with higher activity in extended regions of the semantic network. For low predictable sentences, the inhibitory influence between regions within the cingulo-opercular network was less pronounced with increasing intelligibility. In contrast, higher predictability led to stronger inhibitory connectivity within this network when intelligibility increased. The individual degree of the inhibitory influence of predictability and intelligibility on the connection from the pre-SMA to the left insula predicted the individual predictability gain.

3.1.3 Materials and methods

Participants

Thirty healthy young German native speakers took part in this study. After excluding three participants due to excessive head movement during scanning (i.e. range of motion > 1.5 times the voxel size; see Supplementary Figure 3.1 for an illustration of head movements) and one participant due to reduced task-related activity across the whole brain in the global *F*-test, our final sample yielded 26 participants (*Ra* = 19–29 years, *M* = 25 years; 15 females). All participants were right-handed (mean lateralization index > 90; Oldfield, 1971) and reported no history of neurological or psychiatric disorders as well as no hearing difficulties or disorders.

Participants gave written informed consent prior to participation and received reimbursement of €10 per hour of testing. The study was performed according to the guidelines of the Declaration of Helsinki and approved by the local ethics committee at the Medical Faculty of the University of Leipzig.

Experimental Procedures

Participants took part in one fMRI session. During this session, we first assessed the individual ability of participants to comprehend speech in noise by tracking the speech reception threshold (SRT). Second, participants performed a repetition task on sentences varying in intelligibility and predictability (Figure 3.1). The task had a 2 x 6 full factorial design (predictability: high vs. low; intelligibility: -9, -4, -1, +1, +4, +9 dB sentence intensity relative to the individual SRT) with 12 experimental conditions and 18 trials per condition.

Adaptive tracking procedure of the speech reception threshold. To account for inter-individual differences in speech perception acuity in the main experiment, we determined each participant's SRT beforehand, that is, the SNR required to correctly repeat a sentence with a probability of 50%. In an adaptive up-down staircase procedure (Kollmeier et al., 1988), participants listened to 20 sentences with highly predictable final words energetically masked by speech-shaped noise. After the presentation of each sentence, participants repeated the whole sentence as accurately as possible and the investigator directly rated the response as either correct or incorrect. We rated a response as incorrect if any word in the repeated sentence was missing, incomplete or inaccurately inflected. A correct repetition was followed by an SNR decrease (i.e., the following sentence was less intelligible), an incorrect repetition was followed by an SNR increase (i.e., the following sentence was more intelligible). The tracking procedure started with an initial SNR of 5 dB and a trial-to-trial step size of 6 dB. A turning point (i.e., an increase in SNR is followed by a decrease or vice versa) reduced the step size by a factor of 0.85. We determined SRTs as the average of SNR levels presented on those trials leading to the final five turn points. SRTs had an average of +1.7 dB SNR (*SD* = 2.3, *Ra* = -2.1 to +6.4 dB). All sentences were presented in the MR scanner while running the fMRI sequence later used in the experiment. Individual SRTs were used as a reference for the manipulation of intelligibility in the main experiment to account for inter-individual differences in speech-in-noise comprehension and thereby effectively modulate behavioural performance.

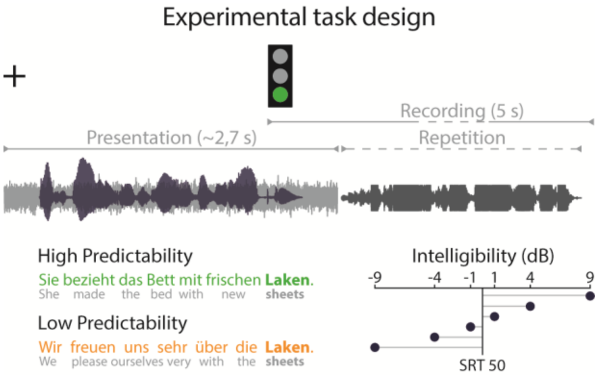

Experimental task design
+
Recording (5 s)
Presentation (~2,7 s)
Repetition
High Predictability
Sie bezieht das Bett mit frischen Laken.
She made the bed with new sheets
Low Predictability
Wir freuen uns sehr über die Laken.
We please ourselves very with the sheets
Intelligibility (dB)
-9
-4
-1
1
4
9
SRT 50

Figure 3.1. Design of the sentence repetition task. In each trial, participants listened to a sentence (black waveform) embedded in distracting background noise (grey waveform) while viewing a fixation cross on the screen. With the onset of the sentence's final keyword, a green traffic light prompted participants to orally repeat the whole sentence as accurately as possible within a 5 sec recording period. The preceding sentence context was either predictive (green) or non-predictive of the keyword (orange). Sentences varied orthogonally in intelligibility (ratio of speech intensity to SRT).

Stimulus material. To manipulate predictability of speech in the sentence repetition task, we used the German version of the speech in noise (SPIN) corpus (Kalikow et al., 1977; for a detailed description of the German version, see Erb et al., 2012). The German SPIN corpus contains pairs of spoken sentences which have the same final word (i.e. keyword) but differ in cloze probability of the keyword (i.e. the expectancy of a word given the preceding sentence; Taylor, 1953). One sentence of a pair provides poor semantic context for the prediction of the keyword ("We are very pleased with the new sheets."; low predictability), whereas the complementary sentence provides rich context for the same keyword ("She made the bed with new sheets."; high predictability). More specifically, the predictability of the keyword varies with the number of pointer words in a sentence that semantically link to the sentence-final word (e.g., "made", "bed", and "new" in the above example). Keywords from sentences with rich context had a mean cloze probability of 0.85 (*SD* = 0.14), low predictable keywords had a mean cloze probability of 0.1 (*SD* = 0.02).

In addition to the 100 original sentence pairs, we extended the corpus by another 36 sentences which were constructed and recorded during the development of the original German SPIN corpus. Initially, these sentences were not included in the corpus, as a smaller set of sentences was sufficient for the purposes of the developers. To obtain predictability ratings for the keywords in the new sentences, 10 participants who did not take part in the fMRI experiment (*Ra* = 24–28 years; *M* = 26 years; 4 females) performed a written sentence completion test (low predictability: < 5 participants reported the correct keyword; high predictability: > 5 participants). Together, 8 new sentence pairs and the original German SPIN sentence pairs yielded 216 experimental sentences (*M* = 2143 ms, *SD* = 256 ms). The remaining 20 new sentences were used in the adaptive tracking procedure prior to the experiment (see Supplementary Figure 3.2 for a detailed illustration of how sentences were assigned to the experiment and the adaptive tracking procedure).

Intelligibility of speech was manipulated by varying the intensity of the experimental sentences relative to a constant signal intensity of speech-shaped noise in the background. We created spectrally speech-shaped noise by filtering white noise with the long-term average spectrum of all experimental sentences (Nilsson et al., 1994). The noise stream preceded and ensued single sentences by 250 ms. Our six experimental intelligibility levels were symmetrically spaced around individual SRTs and logarithmically increased towards the extremes. With our intelligibility levels, we aimed to cover the full range of behavioural performance in each participant and sample intermediate intelligibility in smaller steps as SNR changes in this range affect performance more strongly.

Sentence repetition task. Participants performed an overt sentence repetition task during fMRI. They listened to single sentences while viewing a white fixation cross on a black screen. With the onset of the keyword, a green traffic light appeared on the screen and indicated the start of a 5000 ms recording period. During the recording period, participants orally repeated the sentence as accurately as possible. Whenever participants did not comprehend the complete sentence, they repeated the sentence fragments they could grasp. If participants could not repeat any word, they said that they did not understand the sentence. The intertrial interval was randomly set between 2000 and 7000 ms and served as an implicit baseline for the fMRI analysis. Sentences were presented in a pseudorandomized order avoiding the repetition of one intelligibility level more than three times in a row to prevent adaptation to specific intelligibility levels. The experiment was split in six blocks of approximately 8 minutes which were intermitted by pauses of 20 s each.

All sentences were presented at a comfortable volume. Sound was played via MR-Confon headphones (Magdeburg, Germany) and recorded via a FOMRI-III microphone (Optoacoustics, Yehuda, Israel). Stimulus presentation was controlled by Presentation software (version 18.0, Neurobehavioral Systems, Berkeley, USA, https://www.neurobs.com). The experiment lasted 50 minutes.

MRI acquisition. Whole brain fMRI data were acquired in one continuous run with a 3 Tesla Siemens Prisma Scanner and 32-channel head coil, using a dual gradient-echo planar imaging multiband sequence (Feinberg et al., 2010) with the following scanning parameters: TR = 2,000 ms; TE = 12 ms, 33 ms; flip angle = 90°; voxel size = 2.5 x 2.5 x 2.5 mm with an interslice gap of 0.25 mm; FOV = 204 mm; multiband acceleration factor = 2. In sum, 1,500 volumes were acquired per participant, each consisting of 60 slices in axial direction and interleaved order. To increase coverage of anterior temporal lobe regions, slices were tilted by 10° off the AC-PC line. Field maps were acquired for later distortion correction (TR = 620 ms; TE = 4 ms, 6.46 ms). Additionally, high-resolution T1-weighted images were either obtained from the in-house database if available or were acquired prior to the functional scans with an MPRAGE sequence (whole brain coverage, TR = 1300 ms, TE = 2.98 ms, voxel size = 1 x 1 x 1 mm, matrix size = 256 x 240 mm, flip angle = 9°).

Data Analysis

Behavioural Data Analysis. All spoken response recordings from the main experiment were cleaned from scanner noise using the noise reduction function in Audacity (version 2.2.2, https://www.audacityteam.org). For transcriptions, we split the recordings in two independent sets and assigned each recording set to one of two raters. To validate the quality of transcripts, a third rater additionally transcribed half of the recordings in each set. Cohen's kappa coefficient (Cohen, 1960) indicated very good inter-rater reliability for keywords (rater 1 and 3: $\kappa = .94$; rater 2 and 3: $\kappa = .97$). Finally, one rater manually determined onset and duration of oral responses.

As only predictability of the final keyword but not its preceding context is explicitly manipulated in the German SPIN corpus, we limited behavioural analyses to keywords. After rating keywords as incorrect (including missing, incomplete and inaccurately inflected keywords) or correct, we calculated the proportion of correctly repeated keywords for each participant and condition.

In the behavioural analysis, we quantified the effects of predictability and intelligibility on speech comprehension. For each participant, we fitted psychometric curves to the proportion of correct keywords across intelligibility levels, separately for sentences with low and high predictability. In detail, we fitted cumulative Gaussian sigmoid functions as beta-binomial observer models using the Psignifit toolbox (Fründ et al., 2011) in MATLAB (version R2018b, MathWorks). The beta-binomial model extends the standard binomial model by an overdispersion parameter which allows to carry out statistical inference on data affected by performance fluctuations challenging the serial independence of trials (Schütt et al., 2016). To estimate all five parameters describing the function (guess and lapse rate, threshold, width and overdispersion), we applied Psignifit's default priors for experiments with binary single-trial outcome (here: correctly vs. incorrectly repeated keyword; Schütt et al., 2016).

Goodness of psychometric fits was assessed by comparing the empirical deviance (i.e., two times the log-likelihood ratio of the saturated model to the fitted model) with a Monte Carlo simulated deviance distribution (Wichmann & Hill, 2001). For each psychometric curve, a reference distribution was created by 1) randomly drawing from a binomial distribution with n = 18 trials and p = fitted probability of a correct response for each intelligibility level, 2) calculating the deviance across intelligibility levels, and 3) repeating this procedure for 10,000 samples. All empirical deviances fell within the 97.5% confidence interval of their respective reference distribution, indicating that psychometric curves properly represented observed behavioural data in every participant. There was no significant difference between the deviance of fits for sentences with low vs. high predictability (t_{25} = 0.21, p = 0.839, r = 0.21, BF_{10} = 0.21). Parameter estimates including credible intervals for single participants can be found in Supplementary Figure 3.3. To illustrate goodness of fit, we additionally calculated R^2 based on the Kullback-Leibler divergence (R^2_{KL}) which represents the reduction of uncertainty by the fitted model relative to a constant model (Cameron & Windmeijer, 1997). On average, psychometric curves of sentences with high predictability yielded an R^2_{KL} of 0.96 (SD = 0.04, Ra = 0.84–0.998), sentences with low predictability an R^2_{KL} of 0.94 (SD = 0.05, Ra = 0.79–0.997).

As all 12 experimental conditions become subsumed in two psychometric curves, interaction effects can be evaluated by comparing parameter estimates between curves for sentences with low and high predictability. The threshold represents the intelligibility level at which participants correctly repeat half of the keywords. Keeping all other parameters constant, a threshold decrease indicates a comprehension gain at intermediate but not high and low intelligibility levels. We hypothesized that the interaction effect of intelligibility and predictability manifests in a smaller threshold for sentences with high predictability when compared to low predictability. Additionally, we expected that an increase in intelligibility leads to a stronger comprehension gain for sentences with high relative to low predictability. The sensitivity to changes in intelligibility is reflected by the slope which is inversely related to the width and here describes the steepness of a psychometric curve at a proportion correct of 0.5. To explore whether also other parameters of psychometric curves were modulated by the predictability of sentences, we compared guess and lapse rate (i.e., lower and upper asymptote) as well as width between sentences with high and low predictability. Parameter estimates of the two psychometric curves were contrasted using a paired-sample *t*-test (or a Wilcoxon signed-rank test whenever the assumption of normality was not met). Additionally, we regressed out SRTs from psychometric parameter estimates and reran all *t*-tests between sentences with low and high predictability to control for potential confounding effects of overall inter-individual differences in auditory perception.

We quantified the strength of evidence in favour of our hypotheses by calculating the Bayes Factor (BF; Jeffreys, 1961) for all *t*-tests and correlations carried out in the behavioural analyses using the default settings in JASP (i.e., comparing against a Cauchy prior with a scale parameter of 0.707; version 0.8.6.0). Whereas BF_{10} = 1 supports neither the alternative nor the null hypothesis, BF_{10} > 3 indicates increasingly substantial support for the alternative hypothesis (Kass & Raftery, 1995). For BF_{10} = 3, the data are 3 times more likely to have occurred under the alternative hypothesis than the null hypothesis. As an additional effect size measure, we report Pearson’s correlation coefficient *r* (Rosenthal & Rubin, 1994).

Functional MRI Data. Functional MRI data were analysed using the Statistical Parametric Mapping software (SPM12, version 7219, Wellcome Department of Imaging Neuroscience, London, UK) and MATLAB (version R2017b). Preprocessing steps were applied in line with SPM’s default settings and comprised realignment, distortion correction, segmentation, co-registration to the individual high-resolution structural T1-scan and spatial normalization to the standard template by the Montreal Neurological Institute (MNI) while keeping the original voxel size. Volumes were smoothed using a 5 mm full width at half maximum Gaussian kernel to allow statistical inference based on Gaussian random-field theory (Friston et al., 1994). Realignment parameters were obtained from the first echo. Temporal SNR maps were calculated for each echo based on the first 30 volumes of each participant. Echoes were combined using the information from the corresponding temporal SNR maps as weighting factors and realigned to the first scan. This approach has been proposed to increase SNR in brain regions typically suffering from signal loss (e.g., anterior temporal lobe regions; for a similar approach, see Halai et al., 2014). Motion parameters were calculated from the parameters obtained by realignment. Additionally, framewise displacement as described by Power and colleagues (2012) was calculated as an index of head movement. Volumes that exceeded a framewise displacement of 0.9 (Siegel et al., 2014) were included as nuisance regressors.

At the single-participant level, we set up a general linear model (GLM). For each experimental condition, onset and duration of the sentence presentation period were modelled with a stick function and convolved with SPM's canonical hemodynamic response function. Additionally, we modelled onset and duration of the sentence repetition periods using individual speech onset times and speech durations. We further included motion parameters obtained from the realignment step and framewise displacement as regressors of no interest. Data were high-pass filtered with a cut-off at 128 s and serial correlations were taken into account using a first-order autoregressive model. Each experimental regressor was contrasted against the implicit baseline (intertrial interval), thus yielding 12 contrast images of interest, one for each experimental condition. Additionally, an *F*-contrast containing the experimental regressors was set up to capture experimental effects of interest and exclude effects of no interest (i.e. sentence repetition periods, motion-related activity).

At the group level, these contrasts were submitted to a random-effects flexible factorial design with the experimental factors predictability (high, low) and intelligibility (-9 to +9 dB) entered as interaction term, and an additional factor modelling the subjects' constant. First, we set up a *t*-contrast modelling a linear increase in intelligibility (-9 > -4 > -1 > +1 > +4 > +9 dB SNR relative to SRT) to reveal brain regions tuning to increasing clarity of speech irrespective of predictability (i.e. main effect of intelligibility).

Further, we were interested in those brain regions showing a linear or quadratic interaction of intelligibility and predictability. The interaction contrasts were set up as separate t-contrasts for linearly (high > low, -9 > -4 > -1 > +1 > +4 > +9 dB SNR relative to SRT) and quadratically modelled intelligibility (high > low, -9 < -4 < -1 = +1 > +4 > +9 dB). These parametric interaction contrasts were set up in both directions (high > low; low > high).

Single-participant's parameter estimates were extracted from significant peak voxels and transformed into percent signal change using the rfx toolbox (Gläscher, 2009) for illustration purposes. Figures were created using MRIcroGL (https://www.mricro.com, version v1.0.20180623).

All effects for the whole-brain group analysis are reported at a voxel-level family-wise error rate (FWE) correcting threshold of $p < 0.05$ to control for multiple comparisons. We do not report clusters with less than 10 voxels, as we consider such small clusters biologically implausible. Local maxima are reported in MNI space. We used the SPM Anatomy Toolbox (Eickhoff et al., 2005, version 2.2c) and the Harvard-Oxford cortical structural atlas (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) to classify the anatomical correspondence of significant voxels and clusters.

Brain–Behaviour Correlation. We identified brain regions reflecting the behavioural predictability gain across intelligibility levels by correlating neural response patterns with individual task performance. On the behavioural level, we extracted the proportion of correctly repeated keywords in each experimental condition and participant as modelled by psychometric curves.

To quantify how strongly the proportion correct in each intelligibility level was affected by the predictability of sentences for single participants, we weighted the interaction contrast by behaviour. The interaction contrast was built from the Kronecker product of differential predictability and intelligibility effects. A differential effect is the transposed orthonormal basis of differences between columns of an l-by-l identity matrix (predictability: $l = 2$, intelligibility: $l = 6$; for a detailed description see Henson, 2015). Multiplying this F-contrast with the condition-wise task performance of each participant yielded the proportion of variance in performance explained by the interaction, i.e. the predictability gain (or loss) in each intelligibility level.

On the neural level, we limited our analysis to all voxels consistently implicated in task-related activity as indicated by a significant F-contrast ($p_{\text{uncorrected}} < 0.05$) calculated across single-participant parameter estimates of all experimental conditions (see mask in Supplementary Figure 3.4). For each voxel within the mask, we correlated single-participant behavioural predictability gain and neural parameter estimates across conditions, using a custom Matlab script. After applying Fisher's z-transformation to all Pearson product-moment correlation coefficients, single-participant maps were submitted to the Local Indicators of Spatial Association (LISA) group-level one-sample t-test (Lohmann et al., 2018). LISA is a threshold-free framework that allows to find consistent effects in small brain regions by applying a non-linear filter to statistical maps before controlling for multiple comparisons at a false discovery rate (FDR) < 0.05. Voxel-wise FDR scores were obtained from a Bayesian two-component mixture model in which filtered statistical values are compared to a null distribution based on 5000 permutations (i.e., randomly switching signs of statistical maps in participants).

Dynamic Causal Modelling (DCM)

To investigate condition-specific interactions between brain areas, we conducted two separate bilinear one-state deterministic Dynamic Causal Modelling (DCM) analyses using the DCM 12.5 implementation in SPM 12. DCM is a forward-model-based method that models the dynamics of hidden neuronal states within a predefined set of regions and relates them to the measured BOLD signal (Friston et al., 2003). Using this Bayesian framework, modulations of effective connectivity by experimental conditions can be assessed.

Seed region selection. According to our main hypothesis, we were interested in the neural network dynamics underlying the predictability gain in speech in noise processing and the role of left angular gyrus. In the GLM analysis, we found two distinct networks implicated in the neural interaction of predictability and intelligibility: the semantic network and the cingulo-opercular network (CO). Here, we test effective connectivity in a subset of regions within each network. The semantic model was based on the GLM interaction contrast (high > low, linearly modelled intelligibility) and included left posterior middle temporal gyrus (pMTG), PGa and PGp as seed regions. This model specification allowed to investigate differential contributions of subregions within AG, namely PGa and PGp. Based on the opposite GLM interaction contrast (low > high, linearly modelled intelligibility), we specified the CO model including pre-SMA/paracingulate gyrus as well as left and right anterior insula as seed regions.

For each subject, we identified the peak voxel nearest to the group maximum of each seed region within a sphere of 10 mm radius using the appropriate first-level contrast. Individual timeseries of each region (summarized as the first eigenvariate) were extracted from all voxels within a sphere of 6 mm radius centered on the individual maximum and exceeding the liberal threshold of $p_{\text{uncorrected}} < 0.05$. The extracted timeseries were adjusted to the respective effects-of-interest *F*-contrast to exclude variance of no interest. The exact anatomical localization of individual seed regions was verified using SPM's Anatomy Toolbox (Eickhoff et al., 2005). For spatially adjacent PGa and PGp, we made sure that seed regions were not overlapping within each subject.

The summarized timeseries together with information about the onsets of the relevant conditions served as input for the DCM. The design matrix for the DCM analyses differed from the design matrix used in the GLM analysis in the following way: seven regressors were defined, modelling the onsets of (1) all experimental stimuli, (2) high predictable sentences at low intelligibility levels (-9, -4 dB SNR relative to SRT), (3) high predictable sentences at medium intelligibility levels (-1, +1 dB SNR relative to SRT), (4) high predictable sentences at high intelligibility levels (+4, +9 dB SNR relative to SRT). The remaining regressors modelled the onsets of low predictable sentences at low, medium and high intelligibility, respectively. The first regressor served as sensory input (i.e. driving input) of the stimuli into the model. Regressors (2) to (7) were used as modulatory inputs and encoded predictability and intelligibility of the stimulus. Considering that each experimental condition comprised only 18 trials, we aimed to increase the SNR of our DCM parameter estimates by binning neighbouring intelligibility levels into groups of low, medium and high intelligibility. As inputs were not mean-centered, intrinsic parameter estimates can be interpreted as average connection strength in the absence of experimental manipulations (Marreiros et al., 2008).

DCM model architecture. At the first-level, 84 models were specified for each subject. All models consisted of three types of parameters: (i) intrinsic connections between two nodes (unidirectional or bidirectional), (ii) changes in coupling strength between regions dependent on the experimental manipulation (modulatory effects) and (iii) direct influence of external input on a region (driving input).

For both DCM analyses, we modelled full reciprocal intrinsic connections as well as (inhibitory) self-connections. The driving input was set to either one region at a time, or all possible combinations between regions (resulting in 7 possible combinations). The following potential modulatory influences were considered: either one connection (except for the self-connections) was modulated separately or two afferent or efferent connections were modulated by all experimental conditions at a time (12 possible combinations; see Supplementary Figure 3.5 for a schematic overview of the model space). Since family-level inference has been found especially useful in case of large model spaces (Penny et al., 2010), the resulting model space of 7 x 12 models was partitioned into 12 families, grouping models according to the modulatory input (i.e. within a family, the modulating connection was kept constant while the driving input was varied). We used the same model architecture for both DCM analyses.

All 84 models per participant were inverted using a Variational Bayes approach to estimate parameters that provide the best trade-off between accuracy and complexity quantified by free energy (Friston et al., 2006, 2007). At the second level, a random-effects Bayesian Model

Selection procedure was applied to identify the most probable family of models given the fMRI data quantified by the respective exceedance probability (Stephan et al., 2009; Penny et al., 2010). Single-subject parameter estimates were averaged across all models within the winning family using Bayesian Model Averaging. Intrinsic parameter estimates were passed to a one-sample *t*-test. Modulatory parameter estimates were submitted to a two-way repeated measures analysis of variance (ANOVA) with the within-subject factors predictability (high, low) and intelligibility (low, medium, high) to investigate the interaction effect on modulated connections between regions. ANOVAs were calculated in JASP (version 0.9.1); Greenhouse-Geisser correction was applied if Mauchly's test indicated that the assumption of sphericity was not met. The strength of ANOVA effects was quantified with partial eta squared (η_p^2).

Finally, we investigated the within-subject relationship of behavioural comprehension and functional connectivity in the brain. For single participants, we resorted to the condition-wise proportion of behavioural performance explained by the interaction of predictability and intelligibility already used to relate BOLD activity to behavioural performance. To match the number of DCM parameter estimates, we averaged behavioural data according to the binning scheme described above, thus yielding six behavioural values (predictability: high, low; intelligibility: high, medium, low). Those connections with a significant intelligibility-by-predictability interaction of modulatory parameters were correlated with behavioural performance across binned conditions at the single-participant level. Fisher's *z*-transformed Pearson product-moment correlation coefficients were tested against zero with a one-sample *t*-test. This analysis was run in Matlab using a custom script.

3.1.4 Results

Semantic predictability benefits speech comprehension at intermediate levels of intelligibility

On average, participants repeated 52.33% of keywords correctly (*SD* = 8.09). The threshold of psychometric curves fitted to the proportion of correctly repeated keywords across intelligibility levels decreased for sentences with high predictability when compared to low predictability (t_{25} = -9.87, p < 0.001, r = 0.57, BF_{10} > 1,000; Figure 3.2A). This threshold reflected the facilitatory effect that high semantic predictability had on speech comprehension at intermediate intelligibility levels. Moreover, the lapse rate was smaller for sentences with high compared to low predictability (Z = -2.78, p = 0.005, r = 0.55), indicating a predictability gain that persisted (to a smaller degree) beyond intermediate intelligibility in the highest intelligibility level.

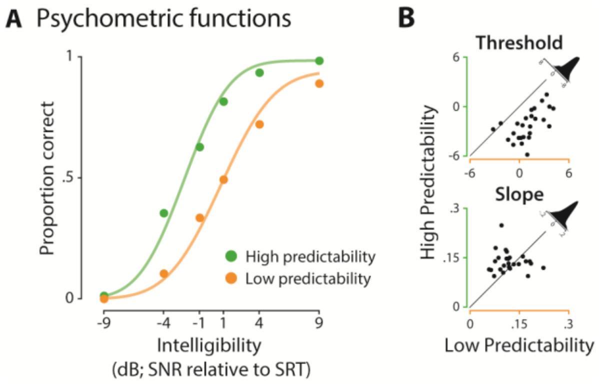


Figure 3.2. Behavioural results of the sentence repetition task. (A) Psychometric functions (coloured lines) were fitted to the proportion of correctly repeated keywords (coloured dots) across intelligibility levels for sentences with low (orange) and high predictability (green); grand average is displayed for illustration purposes. **(B)** The threshold of sentences with low predictability shifted towards more intelligible levels when compared to sentences with high predictability (p < 0.001; top). There was no evidence for a slope difference between sentences with low and high predictability (p = 0.122; bottom). Density plots illustrate the difference between the proportions correct of sentences with high vs. low predictability.

There was no significant difference between sentences with low and high predictability for slope (t_{25} = 1.6, p = 0.122, r = -0.11, BF_{10} = 0.64), width (t_{25} = 1.84, p = 0.077, r = 0.11, BF_{10} = 0.896) and guess rate (Z = 0.27, p = 0.79, r = 0.05) of psychometric curves. Importantly, even though SRTs from the adaptive tracking procedure were strongly correlated with the overall proportion of correctly repeated keywords (r = 0.89, p < 0.001, BF_{10} > 1,000; see Supplementary Figure 3.6), effects of predictability on psychometric curves were unaffected by these inter-individual differences in auditory perception. When controlling for the potentially confounding influence of individual SRTs on parameter estimates, effects of threshold (t_{25} = -9.31, p < 0.001, r = 0.13, BF_{10} > 1,000) and lapse rate (t_{25} = -2.09, p = 0.047, r = 0.06, BF_{10} = 1.33) remained significant and all other effects remained non-significant.

A fronto-temporo-parietal network is tuned towards intelligible speech

First, we were interested in brain regions showing stronger activity for increasing intelligibility of speech stimuli, irrespective of sentence predictability. The parametric variation of intelligibility showed increased activation in a set of regions comprising anterior-to-posterior bilateral superior temporal gyri (STG), left precentral and postcentral gyrus and left inferior frontal gyrus (IFG) as well as a large left-hemispheric temporo-parietal cluster and left precuneus (Figure 3.3). These regions have previously been reported to support auditory speech comprehension and suggested to be tuned to perceptual clarity (Rauschecker & Scott, 2009).

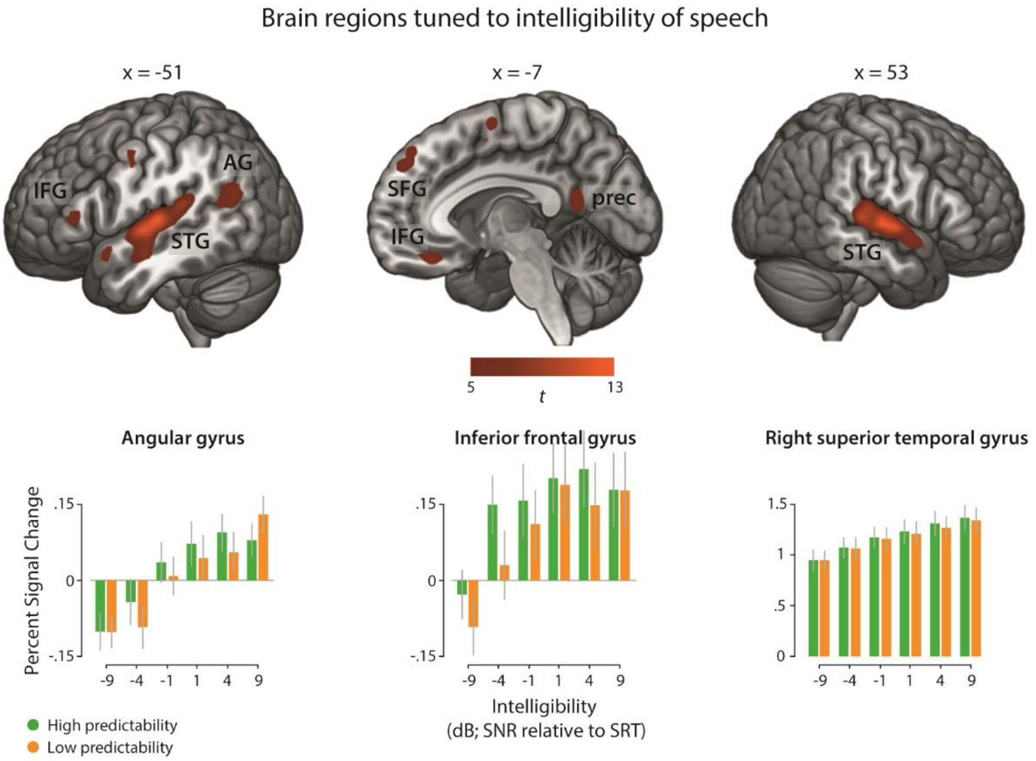


Figure 3.3. Main effect of increasing speech intelligibility across the whole cortex. Top: Activation map thresholded at $p < 0.05$ (FWE-corrected). STG: superior temporal gyrus, IFG: inferior frontal gyrus, AG: angular gyrus, SFG: superior frontal gyrus, prec: precuneus. Bottom: Average parameter estimates (percent signal change) for each experimental condition at peak voxels of left AG (MNI: x = -47, y = -67, z = 25), left IFG (MNI: x = -54, y = 33, z = 8) and right STG (MNI: x = 53, y = -12, z = 2). Error bars represent the standard error of the mean (*SEM*).

Increasing intelligibility of high predictable sentences enhances recruitment of left-hemispheric temporo-parietal regions

Further, we were interested in brain regions that were differentially affected by predictability at different levels of intelligibility. Higher activation for sentences with high compared to low predictability under linearly increasing intelligibility was found in left-hemispheric regions encompassing pMTG, middle to posterior cingulate cortex, two separate clusters in left AG (PGa and PGp), inferior temporal gyrus, precuneus and supramarginal gyrus (SMG; Figure 3.4, Table 3.1). Similar but weaker activation patterns were found for the quadratically modulated interaction contrast. Interestingly, the two separate left-hemispheric AG clusters fell into different cytoarchitectonic subregions: the larger cluster was mainly located in the anterior dorsal portion of AG (PGa & PFm) at the boundary to SMG and the smaller cluster was mainly located in the posterior portion of AG (PGp) in close proximity to the middle occipital gyrus (MOG). The AG clusters and the pMTG cluster showed a pattern of deactivation relative to the fixation baseline with relatively less deactivation for high predictable as compared to low predictable sentences (Figure 3.4B). As both, the AG and pMTG are considered key regions of semantic processing, their time series were submitted to later connectivity analyses.

Increasing intelligibility of low predictable sentences enhances recruitment of domain-general regions

The opposite interaction contrast revealed greater activity for low compared to high predictable sentences under linearly increasing intelligibility and encompassed pre-SMA/paracingulate gyrus, left-hemispheric IFG (pars triangularis; BA 45) and the bilateral anterior insula (see Figure 3.4). These regions overlap with the cingulo-opercular network (Dosenbach et al., 2008), which is frequently reported in conditions associated with high task demands. To investigate context-dependent changes within this network, the connectivity between pre-SMA/paracingulate gyrus and bilateral insulae was analysed in a later DCM analysis. Note that the interaction contrasts modelling intelligibility quadratically (i.e., inverted u-shape with the highest activation at intermediate levels of intelligibility) resulted in a similar pattern of brain regions and therefore are not shown here.

Table 3.1. Results from the context-dependent interaction.

Location	Hemisphere	MNI			*t*	Cluster size
		x	y	z		
High > low predictable sentences						
Middle cingulate cortex	R	3	–34	38	7.49	159
Middle cingulate cortex	L	–4	–30	42	7.17	subcluster
Middle temporal gyrus	**L**	**–60**	**–52**	**–5**	**6.85**	**150**
Middle temporal gyrus	L	–54	–62	2	6.6	subcluster
Angular gyrus (PFm/PGa)	**L**	**–42**	**–62**	**50**	**6.47**	**135**
Angular gyrus (PGa)	L	–42	–70	42	6.31	subcluster
Angular gyrus (PGp)	L	–34	–77	42	6.02	subcluster
Angular Gyrus (PGp)/ middle occipital gyrus	**L**	**–34**	**–77**	**30**	**6.34**	**51**
Angular gyrus (PGp)	L	–42	–77	28	5.63	subcluster
Supramarginal gyrus	L	–60	–30	30	5.91	26
Inferior temporal gyrus	L	–47	–52	–25	5.82	21
Precuneus	L	–10	–47	55	5.62	15
Low > high predictable sentences						
Presupplementary motor area/paracingulate gyrus	**R/L**	**0**	**16**	**50**	**7.72**	**138**
Inferior frontal gyrus (p. orbitalis, area 45)	L	–42	26	–10	6.47	106
Insular cortex	**L**	**–30**	**26**	**–2**	**6.43**	**subcluster**
Inferior frontal gyrus (p. triangularis, area 45)	L	–50	23	5	5.9	subcluster
Inferior frontal gyrus (p. triangularis, area 45)	L	–44	20	0	5.88	subcluster
Insular cortex	**R**	**30**	**26**	**–2**	**6.13**	**24**
Inferior frontal gyrus (p. triangularis, BA 45)	L	–57	20	15	5.71	16
Inferior frontal gyrus (p. triangularis, BA 45)	L	–50	20	18	5.07	subcluster

High > low predictable sentences: Regions showing greater activity for high as compared to low predictable sentences under increasing intelligibility; Low > high predictable sentences: Regions showing greater activity for low as compared to high predictable sentences under increasing intelligibility. Regions are reported at a peak-level FWE-correcting threshold at an alpha of .05. Seed regions for DCM analyses are highlighted in bold.

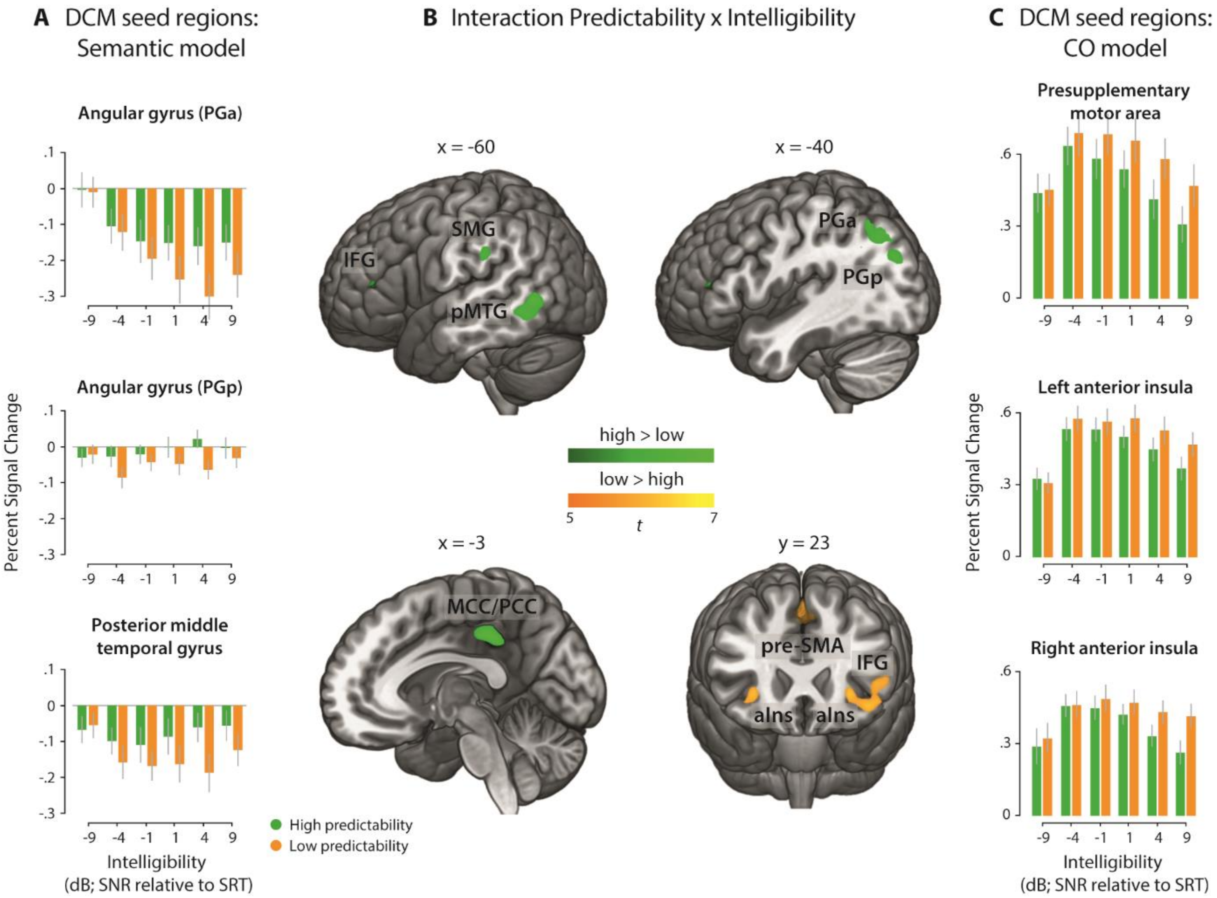


Figure 3.4. Group-level fMRI results illustrating brain regions that are sensitive to the interaction of intelligibility and predictability. (A) Parameter estimates illustrate the average interaction effect, extracted from peak voxels of DCM seed regions for the semantic network. **(B)** Brain regions showing a significant interaction; IFG: inferior frontal gyrus, SMG: supramarginal gyrus, pMTG: posterior middle temporal gyrus, PGa: anterior angular gyrus, PGp: posterior angular gyrus, MCC/PCC: middle and posterior cingulate cortex, pre-SMA: pre-supplementary motor area, aIns: anterior insula; green: high > low predictable sentences, orange: low > high predictable sentences. The activation map is thresholded at $p_{\text{FWE-corrected}} < 0.05$. **(C)** Parameter estimates illustrate the interaction effect, extracted from peak voxels of DCM seed regions for the cingulo-opercular (CO) network. Error bars represent ±1 *SEM*.

The extended semantic system scales to the behavioural predictability gain dependent on speech intelligibility

The proportion of individual task performance explained by the interaction of predictability and intelligibility in each experimental condition correlated positively with BOLD activity in a broad semantic network (Figure 3.5; Binder et al., 2009). At intermediate levels of intelligibility, neural activity increased for sentences with high predictability in those participants exploiting a stronger behavioural predictability gain.

In line with the linear interaction effect (see Table 3.1, high > low predictability), modelling the behavioural interaction effect revealed significant clusters in parietal and lateral temporal cortex. A broad left-hemispheric and a smaller right-hemispheric cluster were observed in PGp (Table 3.2). Additional clusters were found in bilateral pMTG.

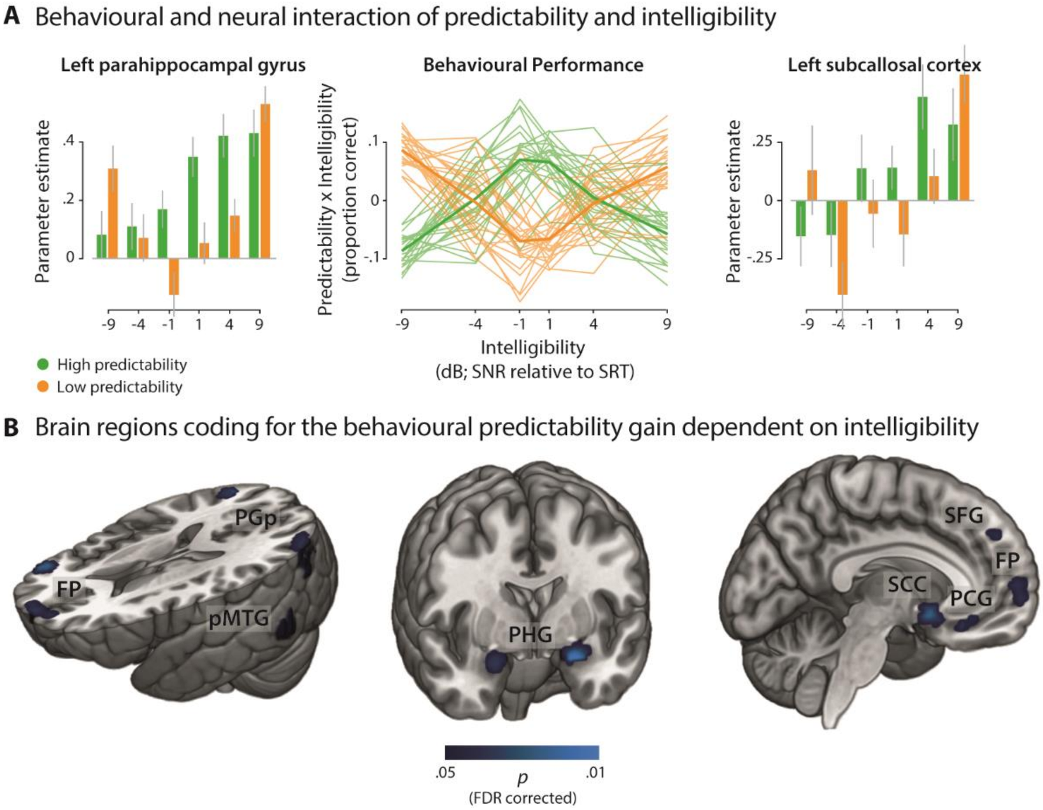


Figure 3.5. Projection of activity onto the behavioural predictability gain across levels of intelligibility. (A) The condition-specific proportion of behavioural performance explained by the interaction of intelligibility and predictability in each participant (middle line plot) was correlated with individual parameter estimates of single-voxel BOLD activity (left and right bar plots; exemplary grand average parameter estimates for two peak voxels). Thin lines in the line plot represent single participants; fat lines represent grand average. Error bars represent ±1 *SEM*. **(B)** Group statistics revealed brain areas modulated by the single-participant interaction effect of intelligibility and predictability on speech comprehension. *p*-Values were FDR-corrected within a mask of voxels responsive to the listening task. FP: frontal pole; PGp: posterior angular gyrus; pMTG: posterior middle temporal gyrus; PHG: parahippocampal gyrus; SCC: subcallosal cortex; SFG: superior frontal gyrus; PCG: paracingulate gyrus (for an exhaustive list of clusters see Table 3.2).

More importantly, the behavioural interaction effect mapped onto frontal and ventromedial brain regions not implicated in the linear interaction effect. The frontal pole encompassed one broad medial cluster spanning from the left to the right hemisphere as well as a right-hemispheric cluster at the border with IFG and a left-hemispheric cluster extending into the superior frontal gyrus (Table 3.2). Near the inferior-posterior frontal pole, we observed bilateral clusters in IFG. The limbic lobes contained a bilateral cluster in the posterior subcallosal cortex that was complemented by a bilateral cluster in inferior paracingulate gyrus extending into anterior subcallosal cortex as well as cingulate gyrus and frontal medial cortex. Additional clusters comprised the anterior division of bilateral superior parahippocampal gyrus and left medial-posterior insular cortex. In the medial temporal lobe, we found a left-hemispheric cluster in the posterior division of the temporal fusiform gyrus. No significant cluster in the brain showed a negative correlation with the interaction effect on task performance.

Table 3.2. Brain areas sensitive to the interaction effect of intelligibility and predictability on task performance.

Location	**Hemisphere**	**MNI**			***t***	**Cluster size**
		x	y	z		
Brain Behaviour Correlation						
Frontal pole	L	−5	56	5	3.53	161
	R	48	38	3	3.96	103
	L	−12	56	35	3.71	26
Angular gyrus (PGp)	L	−50	−70	33	3.85	100
	R	46	−67	25	3.45	55
Subcallosal cortex	L	−2	6	−13	4.22	78
Middle temporal gyrus (temporo-occipital)	L	−60	−55	−5	3.65	80
	R	58	−45	−5	4.04	25
	L	−57	−65	10	2.76	2
Parahippocampal gyrus (anterior)	R	21	1	−18	3.74	55
	L	−25	−2	−20	4.31	51
Paracingulate gyrus	R	1	41	−8	3.96	54
Insular cortex	L	−33	−17	5	3.74	17
Temporal fusiform gyrus (posterior)	L	−42	−30	−23	3.53	5
Inferior frontal gyrus (p. orbitalis)	R	28	31	−15	3.94	13
	L	−37	36	−10	3.71	3

Each cluster ($p_{FDR} < 0.05$) is described by hemisphere, MNI coordinates and *t*-value of its peak voxel as well as the number of voxels it contains.

Connectivity between subregions of the angular gyrus increases with higher intelligibility

Next, to investigate context-dependent changes in effective connectivity between regions identified in the mass univariate GLM analysis, we performed two separate DCM analyses, one within the semantic system and one within the cingulo-opercular network.

The family of models with a modulatory influence on the connection from PGp to pMTG and PGa was identified as the winning family by means of Bayesian model selection (exceedance probability: 0.65; see Supplementary Figure 3.7A for an overview of all exceedance probabilities). All intrinsic connections, except for self-connections, were positive (see Supplementary Table 3.1), although only the outgoing connections from pMTG and the self-connections reached statistical significance. Within the winning family, there was a significant main effect of intelligibility on the connection from PGp to PGa ($F_{1.56,\ 39.07} = 3.57$, $p = 0.048$, $\eta_p^2 = 0.13$, Greenhouse-Geisser corrected, Mauchly's sphericity test: $p = 0.02$). The modulatory influence of intelligibility increased connectivity from PGp to PGa when speech became more intelligible. The main effect of predictability ($F_{1,25} = 1.86$, $p = .185$, $\eta_p^2 = .07$) and the interaction effect ($F_{2,50} = 0.71$, $p = .496$, $\eta_p^2 = 0.03$) were not significant. There was no significant effect on the connection from PGp to pMTG (main effect predictability: $F_{1,25} = 0.02$, $p = .899$, $\eta_p^2 = .001$; main effect intelligibility: $F_{2,50} = 0.35$, $p = .705$, $\eta_p^2 = .01$; interaction effect: $F_{2,50} = 1.55$, $p = .222$, $\eta_p^2 = 0.06$).

Inhibitory connectivity within the cingulo-opercular network increases with the individual predictability gain at intermediate intelligibility

For the cingulo-opercular network model, the family of models with a modulatory effect on the connections from pre-SMA and right insula to left insula was identified as the winning family, with an exceedance probability of 0.73 (see Supplementary Figure 3.7B).

In the winning model, there was a significant interaction in the connection from pre-SMA to left insula ($F_{2,50} = 4.74$, $p = 0.013$, $\eta_p^2 = 0.016$): inhibitory influence was less pronounced for low predictable sentences with increasing intelligibility but stronger for high predictable sentences with increasing intelligibility. The respective main effects of intelligibility ($F_{2,50} = 0.55$, $p = .583$, $\eta_p^2 = .02$) and predictability ($F_{1,25} = 0.11$, $p = .747$, $\eta_p^2 = .004$) were not significant (see Figure 3.6A and Supplementary Table 3.2). There was no significant effect on the connection from right insula to left insula (main effect intelligibility: $F_{2,\ 50} = 0.36$, $p = .702$, $\eta_p^2 = .01$; main effect predictability: $F_{1,25} = 1.64$, $p = .212$, $\eta_p^2 = .06$; interaction effect: $F_{2,50} = 2.07$, $p = .137$, $\eta_p^2 = 0.08$).

Following up the interaction effect in connectivity between pre-SMA and left insula with a post-hoc analysis, we investigated how individual connectivity patterns match the individual pattern of the behavioural predictability gain across intelligibility levels. We found a significant negative correlation (t_{25} = -3.919, p = 0.004, BF_{10} = 53.17), indicating that the individual predictability gain at medium intelligibility was associated with stronger inhibitory influence from pre-SMA to the left insula for highly predictable speech (see Figure 3.6B).

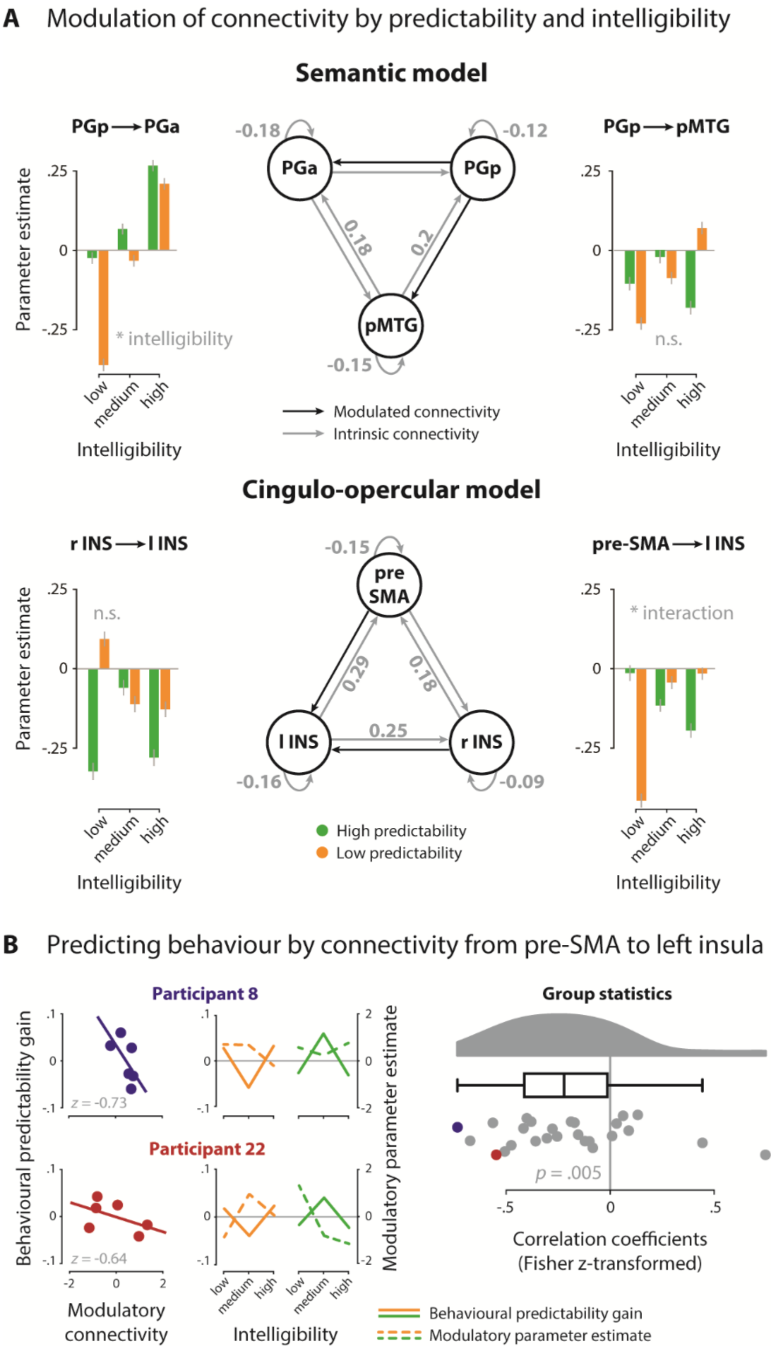


Figure 3.6. DCM results. (A) Network architecture of the semantic (top) and cingulo-opercular network (bottom); grey arrows with parameter estimates indicate significant intrinsic connections; black arrows indicate modulatory connections. Bar graphs show the modulation of connections by intelligibility (low, medium, high; binned) and predictability (low, high). The connection from PGp to PGa was significantly modulated by the main effect of intelligibility, the connection from pre-SMA to left insula was sensitive to the interaction of predictability and intelligibility. Bar height represents grand average, error bars represent ±1 *SEM*. **(B)** Scatterplots (left) show the negative correlation of the condition-wise predictability gain with the condition-wise modulatory parameter estimates from pre-SMA to left insula in two exemplary participants. Complementary, line plots (middle) illustrate patterns of behavioural performance (solid line) and modulation strength (dashed line) resolved for intelligibility (low, medium, high) and predictability (low: orange; high: green). Raincloud plot (right) shows the distribution of individual *z*-transformed correlations across all participants; dots representing participants from scatter and line plots are highlighted in colour. *n.s.* = not significant, * *p* < 0.05. PGa: anterior angular gyrus; PGp: posterior angular gyrus; pMTG: posterior middle temporal gyrus; pre-SMA: pre-supplementary motor area; l INS: left anterior insula; r INS: right anterior insula.

3.1.5 Discussion

In the present fMRI study, we investigated the neural underpinnings of the semantic predictability gain during speech processing in noisy listening situations at the network level. First, we showed that highly predictable but acoustically degraded speech engages left-hemispheric semantic regions and modulates the individual comprehension gain in behaviour via additional frontal and medial semantic regions. Second, we found that highly intelligible speech strengthens effective connectivity between two subregions of a key semantic network node in the left angular gyrus. Third, we showed for the cingulo-opercular network that inhibitory influence from pre-SMA to left insula was stronger for increasing intelligibility of highly predictable speech but was less pronounced for low predictability. Notably, individual behavioural response patterns were negatively correlated with individual connectivity patterns. Specifically, the higher the behavioural comprehension gain at intermediate intelligibility was, the stronger the inhibitory connectivity strength within individual subjects. These findings suggest that semantic predictability facilitates speech comprehension in challenging listening conditions via upregulation of the semantic network and active inhibition between domain-general regions.

Our sentence repetition task was designed to overcome the experimental shortcomings of previous studies and investigate speech comprehension in noise on the basis of rich neural and behavioural data. Hitherto, fMRI studies on the neural underpinnings of degraded speech processing had mainly used sparse temporal sampling to avoid interference of scanner noise with experimentally manipulated SNR of speech stimuli (Obleser & Kotz, 2010; Erb et al., 2013; Golestani et al., 2013). As we aimed to investigate speech processing directly during (compared to after) listening and to improve sampling density for our connectivity analyses, BOLD activity in the present study was continuously recorded throughout the whole experiment. In line with studies using temporal sparse sampling (Davis & Johnsrude, 2003; Rauschecker & Scott, 2009; Abrams et al., 2013), we found activity in regions associated with auditory comprehensibility (e.g., bilateral STG as well as left IFG) to be stronger with increasing intelligibility of speech. This replication of the intelligibility effect speaks for continuous scanning as a feasible technique to investigate *online* speech processing.

Crucially, in contrast to previous studies with only few and fixed experimental levels of speech degradation, we covered the full range of intelligibility from hardly to easily comprehensible speech in the individual participant. Note that our experimental design contained neither a "noise only" nor a clear speech condition. At the behavioural level, we were nevertheless able to fit full psychometric curves (i.e., speech comprehension as a function of intelligibility, see Figure 3.2). We found the psychometric curve to shift towards less intelligible levels for sentences with high compared to low semantic predictability. This threshold shift is an established effect in more challenging comprehension tasks (Shannon et al., 2004) and reflects a comprehension benefit for predictable speech at intermediate intelligibility.

As left AG has been found to show a strong differentiation between sentences with high and low predictability at intermediate intelligibility (Obleser et al., 2007; Obleser & Kotz, 2010) and to exhibit functional relevance in mediating the comprehension gain (Hartwigsen et al., 2015), this region has been proposed to serve top-down activation of semantic concepts that facilitate speech comprehension (Obleser & Kotz, 2010). However, as an unexpected finding, two separate AG clusters showed stronger activation for sentences with high compared to low predictability under linearly increasing intelligibility: a larger cluster mainly falling into PGa at the boundary to the supramarginal gyrus, and a smaller cluster at the boundary between PGp and middle occipital gyrus. Importantly, these AG subregions have been discussed to subserve different functions (Seghier et al., 2010; Noonan et al., 2013; Bonnici et al., 2016). PGa is implicated in domain-general attention and converges with coordinates commonly reported as AG activation during speech in noise processing (Obleser et al., 2007; Obleser & Kotz, 2010; Clos et al., 2014; Guediche et al., 2014). In contrast, PGp activation has often been labelled as MOG (McGettigan et al., 2012; Guediche et al., 2016). In these studies, MOG was interpreted to provide free resources otherwise recruited by visual processing to the auditory domain during resource-demanding speech in noise comprehension. Other studies have associated PGp activation with "pure" semantic processing (Seghier et al., 2010; Bonnici et al., 2016). Our findings implicate that both AG subregions contribute to successful speech comprehension under challenging listening conditions when semantic cues are available.

To further characterize the functional contributions of anterior and posterior AG, we analysed the connectivity between these subregions and observed that coupling from posterior to anterior AG was strengthened when speech became more intelligible. Surprisingly, this connection was not modulated by the predictability of speech. This finding suggests that the interplay of AG subregions might support the integration of lexical information in general, but does not specifically facilitate integration of semantic information. Further, this finding speaks against the view that the anterior portion (PGa) is associated with goal-directed attention processes as one would expect attentional demands to be highest at intermediate levels of intelligibility and not at the most intelligible levels. However, it is important to bear in mind that the effective connectivity results are restricted to the limited set of regions included in the network models. Aside from AG, we largely found the same regions (e.g., left pMTG, left SMG, left inferior temporal gyrus, posterior cingulate gyrus and precuneus) that were previously implicated in processing of sentences with high predictability under increasing intelligibility (Obleser et al., 2007; Obleser & Kotz, 2010; Golestani et al., 2013). These regions overlap with the previously described core semantic (control) regions (Binder et al., 2009; Jefferies, 2013) and form a functional network during degraded speech processing (Obleser et al., 2007).

To complement the linear interaction contrast, we also modelled task performance in the brain and expected to find the left-hemispheric semantic network strengthened by this contrast fine-tuned to inter-subject variation. Indeed, the semantic core regions extended to their homologous regions in the right hemisphere. Strikingly, additional semantic regions scaled in activation with the behavioural predictability gain at intermediate intelligibility. These regions largely pertained to two subsets of the semantic network: *lateral and ventral temporal cortex* includes left temporal fusiform gyrus and parahippocampal gyrus, whereas *left ventromedial prefrontal cortex* includes subcallosal cortex and frontal pole (Binder et al., 2009). The temporal cortex has been shown to communicate with the hippocampus via a connection from temporal fusiform gyrus to parahippocampal gyrus evident in structural (Powell et al., 2004) and functional imaging (Teipel et al., 2010). It has been suggested that this pathway is used to encode semantic information (Levy et al., 2004; Martin, 2007). Critically, ventral temporal activation has been implicated in the processing of semantically unambiguous speech in a memory-demanding word recognition task (Rodd et al., 2005) and processing of visually presented words that were correctly repeated during later recall (Strange et al., 2002). In the sentence repetition task employed here, memory formation is crucial to correctly repeat the sentence after listening. The association of more accurate recall with stronger ventral temporal activation for predictable but degraded sentences suggests that memory engagement is most crucial for successful speech comprehension when semantic information is available for integration and must be protected against competing but irrelevant information (i.e., background noise). In line with this notion, prefrontal cortex regions implicated in executive control functioning (Duncan & Owen, 2000) show increased activation when efficient management of cognitive resources promises the strongest gain in performance. While we will refrain from overinterpreting these exploratory results, it is worth highlighting that the role of executive control and memory formation might have been often overlooked in previous studies on speech comprehension despite their face validity when it comes to understanding individual hearing difficulties.

As a last key finding of our study, challenging speech comprehension in the absence of semantic constraints is accompanied by an increase of activity in the pre-SMA and bilateral anterior insulae, which are regions associated with the cingulo-opercular network. This finding is broadly in line with previous reports on activation of cingulo-opercular regions during effortful speech processing (Eckert et al., 2009; Adank, 2012; Hervais-Adelman et al., 2012; Erb & Obleser, 2013; Vaden et al., 2013; Clos et al., 2014). The cingulo-opercular network has been associated with domain-general control or executive processes such as task monitoring (e.g., Dosenbach et al., 2008; Duncan, 2010; Vaden et al., 2013, 2015), executive control (Erb & Obleser, 2013) and attentional control (Fitzhugh et al., 2019). Specifically, in the speech in noise literature, the increased reliance on this network in difficult listening conditions has been interpreted as an adaptive control mechanism to enable speech comprehension when little or no semantic information is available.

Within these cingulo-opercular regions, we found an inverted u-shaped response, with highest activation at intermediate levels of intelligibility and less activation for the conditions that were either hardly or easily comprehensible. This activation pattern suggests that domain-general regions contribute to effortful speech processing as long as *any* information can be extracted. Analogously, recruitment of these regions is less important when the auditory signal is too bad or when speech is clear. On the other hand, the insula has been shown to mirror the inverse quadratic response time effect across intelligibility levels, thereby reflecting the involvement of the insula in response selection (Binder et al., 2004). Even though we did not use a discrimination task but a sentence repetition paradigm, the number of lexical candidates for the response might be highest when speech is hardly comprehensible and not guided by semantic information, thereby explaining an increase in insula activity. However, this does not explain the consistent response pattern across different regions implicated in the cingulo-opercular network that are typically not involved in response selection. Moreover, in a recent meta-analysis, bilateral insulae were linked to higher-order cognitive aspects of speech comprehension and production, highlighting that insula function goes beyond mere involvement in motor aspects of speech production (Oh et al., 2014).

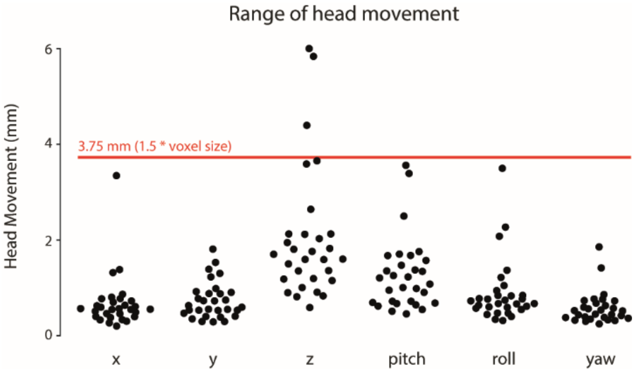
Further, we observed that connectivity between these regions was differentially modulated by high and low predictability. These findings suggest that active inhibition is increased between domain-general cognitive control regions when predictability and intelligibility were highest or lowest. Interestingly, effective connectivity from pre-SMA to the left insula was associated with individual task performance, that is, the inhibitory modulation of this connection increased with the individual degree of the predictability gain at intermediate intelligibility (and decreased with the relative behavioural loss from low predictability). This result converges with and extends previous task-based fMRI studies that found increased activity in the bilateral insulae when auditory input was degraded (Hervais-Adelman et al., 2012; Erb et al., 2013). While older adults with well-preserved hearing abilities and younger participants recruited the anterior insula under challenging listening conditions, recruitment of this region was already observed in clear speech conditions for older adults with hearing loss (Erb & Obleser, 2013). Together, the previous and present findings suggest that recruitment of the left anterior insula is beneficial in challenging conditions (e.g., when auditory input is deteriorated) but becomes unnecessary as soon as the task is easily or not at all solvable. Moreover, the present findings suggest that better behavioural performance may require an active inhibition of this region to ensure efficient processing.

Recently, the pre-SMA has been suggested to be relevant during high task demands (Hertrich et al., 2016) and to play a role in the wider semantic control network (Hallam et al., 2016). Specifically, Hallam and colleagues (2016) found increased activity in pre-SMA and pMTG after inhibitory transcranial magnetic stimulation (TMS) to left IFG, which was interpreted as upregulation of the relative contribution to semantic control in the presence of a TMS-induced disruption. Further, Dietrich and colleagues (2018) reported decreased performance in a sentence repetition task after inhibitory TMS over the pre-SMA, which was selectively evident for the most challenging task condition. Note that pre-SMA activation has also been reported in sentence comprehension tasks in the degraded speech literature before (e.g., Clos et al., 2014: sentence matching task). Therefore, it is unlikely that our pre-SMA finding mainly reflects the choice of the task alone. Taken together, our results replicate the previously reported involvement of domain-general regions during effortful speech processing and extend these findings by demonstrating predictability-dependent modulation of functional connectivity within this network.

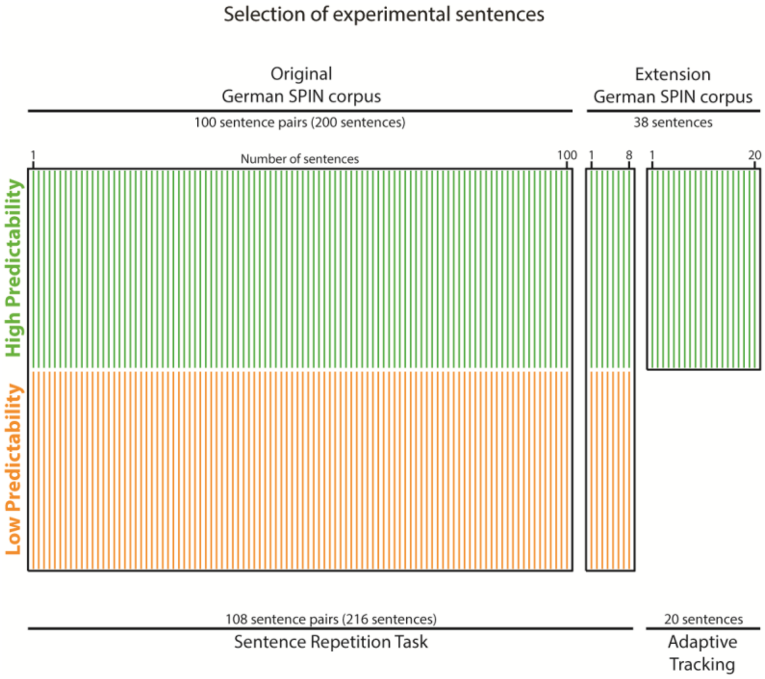
Conclusion

Our results demonstrate that predictability in challenging listening conditions not only modulates the semantic network but critically extends to modulations in the domain-general network in a behaviourally relevant fashion. We demonstrated that intelligibility modulated connectivity between two subregions of left AG, underscoring the functional heterogeneity of this region. Further, the degree of inhibitory modulation of connectivity within the domain-general cingulo-opercular network was associated with the individual speech comprehension gain in behaviour. This highlights the importance to further investigate the dynamic interplay between the semantic and cingulo-opercular network in successful speech comprehension under adverse listening conditions.

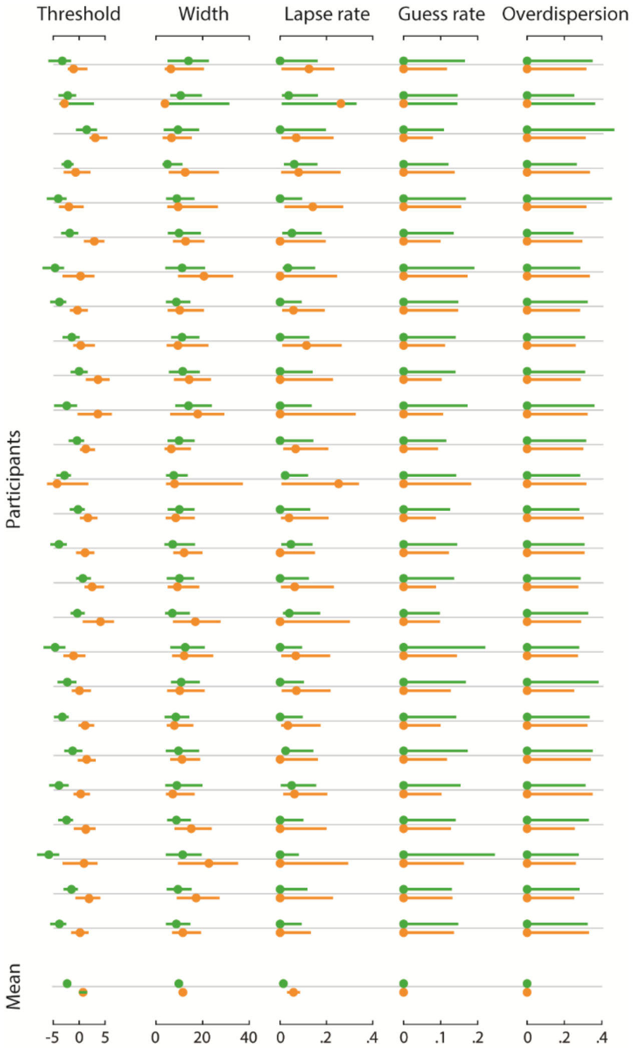
3.1.6 Supplementary materials



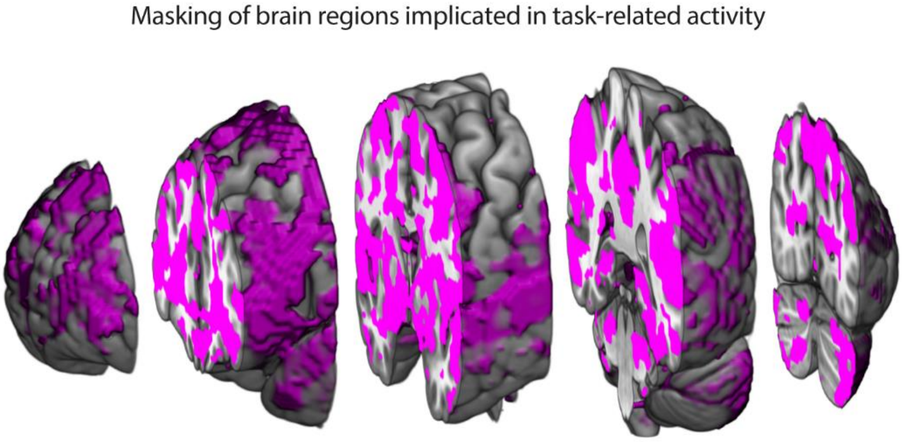
Supplementary Figure 3.1. Overview of the average head movement for each participant during the fMRI experiment. The six movement parameters were obtained from the rigid-body transformation of the realignment procedure during preprocessing and averaged for each participant. Black dots indicate individual averages for each movement parameter and the red line indicates the movement threshold that was set to 1.5 x voxel size. Participants who exceeded this threshold were excluded from further analyses.



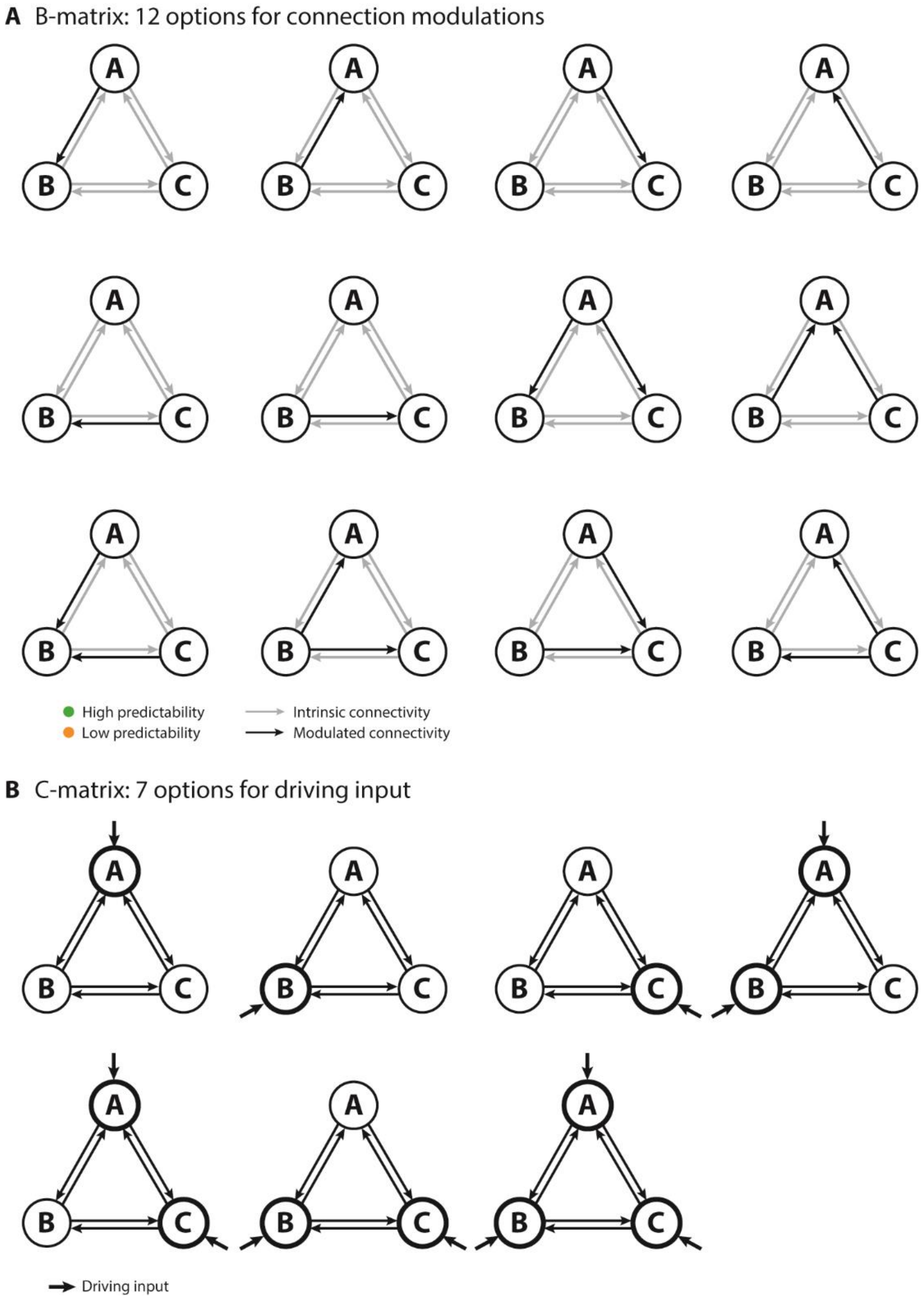
Supplementary Figure 3.2. Schematic illustration of the stimulus material. The original German SPIN corpus (Erb et al., 2012) comprises 200 sentences with pairs of high and low predictable keywords. Together with 8 new sentence pairs, we presented 216 sentences in the sentence repetition task. We used another 20 newly rated sentences with highly predictable keywords for the adaptive tracking procedure prior to the experiment.



Supplementary Figure 3.3. Rows represent single-participant parameter estimates (dots) and 95 % Bayesian credible intervals (horizontal lines) of psychometric curves for sentences with low (orange) and high predictability (green). The bottom row shows parameter estimates averaged across participants with frequentist confidence intervals.


Masking of brain regions implicated in task-related activity

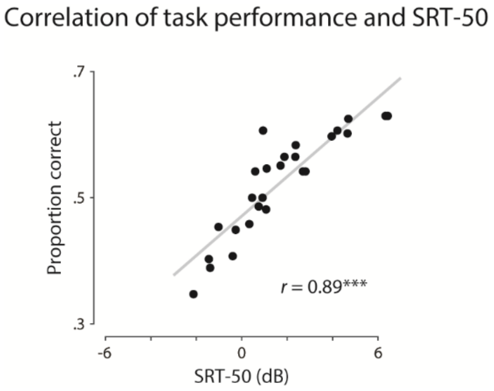
Supplementary Figure 3.4. A brain mask was used to limit the correlational analysis of behavioural and neural responses to those voxels implicated in task-related activity. The mask included all voxels yielding a significant *F*-contrast across single-participant parameter estimates of all experimental conditions ($p_{\text{uncorrected}} < 0.05$).



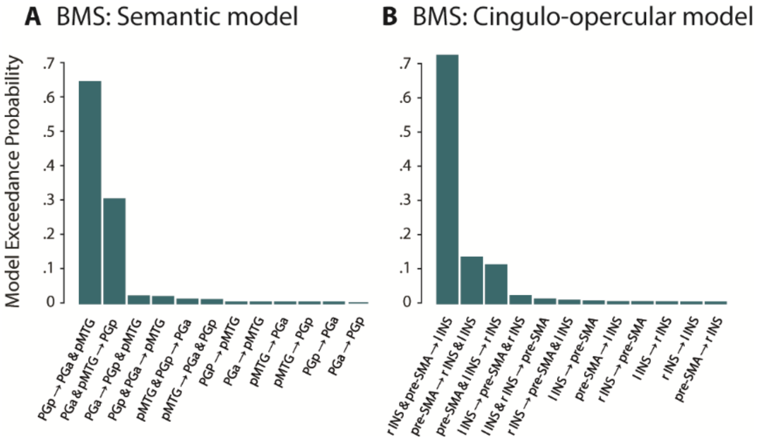
A B-matrix: 12 options for connection modulations

B C-matrix: 7 options for driving input

Supplementary Figure 3.5. Schematic overview of the model space for the semantic as well as the cingulo-opercular DCMs. **(A)** Modulation of intrinsic connectivity (arrows) by high (green) and low (orange) predictability. Modulated connections are highlighted in black. **(B)** Fat arrows indicate driving input. A, B and C represent three distinct brain regions.



Supplementary Figure 3.6. The SRT-50 determined in the adaptive tracking procedure prior to the experiment was strongly correlated with the proportion of correctly repeated keywords across all conditions ($r = 0.89$, $p < 0.001$, $BF_{10} > 1,000$).



Supplementary Figure 3.7. Exceedance probabilities for all DCM model families estimated using Bayesian Model Selection (BMS). **(A)** Model families within the semantic network DCM. **(B)** Model families within the cingulo-opercular network DCM.

Supplementary Table 3.1. Endogenous and modulatory parameter estimates (Hz) of the semantic network DCM.

Connection			*M*	*SD*	*t*	*p*	*d*	BF$_{10}$
Endogenous Parameters								
PGa	→	PGp	0.07	0.38	1.00	0.326	0.20	0.33
PGa	→	MTG	0.04	0.36	0.59	0.562	0.12	0.24
PGp	→	PGa	0.04	0.31	0.69	0.499	0.14	0.28
PGp	→	MTG	0.11	0.40	1.41	0.170	0.28	0.50
MTG	→	**PGa**	**0.18**	**0.29**	**3.22**	**0.003**	**0.63**	**11.61**
MTG	→	**PGp**	**0.20**	**0.29**	**3.55**	**0.002**	**0.70**	**23.13**
PGa	↺		**–0.18**	**0.11**	**–8.24**	**< 0.001**	**–1.62**	**>100.00**
PGp	↺		**–0.12**	**0.16**	**–3.69**	**0.001**	**–0.72**	**31.91**
MTG	↺		**–0.15**	**0.16**	**–4.84**	**< 0.001**	**–0.95**	**>100.00**

Modulatory Parameters						
			Predictability	Intelligibility	*M*	*SD*
PGp	→	PGa	High	Low	–0.02	1.12
PGp	→	PGa	High	Medium	0.07	0.71
PGp	→	PGa	High	High	0.27	0.99
PGp	→	MTG	High	Low	–0.11	0.85
PGp	→	MTG	High	Medium	–0.02	0.85
PGp	→	MTG	High	High	–0.18	1.13
PGp	→	PGa	Low	Low	–0.37	0.93
PGp	→	PGa	Low	Medium	–0.33	0.83
PGp	→	PGa	Low	High	0.21	0.82
PGp	→	MTG	Low	Low	–0.24	0.94
PGp	→	MTG	Low	Medium	–0.89	0.71
PGp	→	MTG	Low	High	0.07	0.79

Parameters were estimated using Bayesian model averaging (BMA) across models of the winning family and significance was assessed by means of *t*-tests. Significant parameters are highlighted in bold. Effect sizes are reported as Cohen's *d* and Bayes Factor (BF$_{10}$).

Supplementary Table 3.2. Results of the endogenous and modulatory parameter estimates (Hz) of the cingulo-opercular DCM.

Connection			*M*	*SD*	*t*	*p*	*d*	**BF_{10}**
Endogenous Parameters								
pre-SMA	→	lIns	−0.04	0.30	−0.67	0.511	−0.13	0.22
pre-SMA	→	rIns	−0.09	0.24	−1.84	0.078	−0.36	0.90
lIns	→	**pre-SMA**	**0.29**	**0.25**	**6.08**	**< 0.001**	**1.19**	**>100.00**
lIns	→	**rIns**	**0.25**	**0.38**	**3.36**	**0.003**	**0.66**	**15.28**
rIns	→	**pre-SMA**	**0.18**	**0.32**	**2.90**	**0.008**	**0.57**	**5.96**
rIns	→	lIns	0.09	0.31	1.43	0.164	0.28	0.51
pre-SMA	↺		**−0.15**	**0.15**	**−4.89**	**< 0.001**	**−0.96**	**>100.00**
lIns	↺		**−0.16**	**0.12**	**−6.83**	**< 0.001**	**−1.34**	**>100.00**
rIns	↺		**−0.09**	**0.16**	**−2.74**	**0.011**	**−0.54**	**4.29**

Modulatory Parameters						
			Predictability	Intelligibility	*M*	*SD*
pre-SMA	→	**lIns**	High	Low	−0.01	0.89
pre-SMA	→	**lIns**	High	Medium	−0.12	0.63
pre-SMA	→	**lIns**	High	High	−0.20	1.02
rIns	→	**lIns**	High	Low	−0.33	0.78
rIns	→	lIns	High	Medium	−0.06	0.78
rIns	→	lIns	High	High	−0.28	0.77
pre-SMA	→	**lIns**	Low	Low	−0.42	0.77
pre-SMA	→	**lIns**	Low	Medium	−0.05	0.65
pre-SMA	→	**lIns**	Low	High	−0.02	0.56
rIns	→	**lIns**	Low	Low	0.09	0.92
rIns	→	lIns	Low	Medium	−0.11	0.77
rIns	→	lIns	Low	High	−0.13	0.61

Parameters were estimated using Bayesian model averaging (BMA) across models of the winning family and significance was assessed by means of *t*-tests. Significant parameters are highlighted in bold. Effect sizes are reported as Cohen's *d* and Bayes Factor (BF_{10}).

3.2 Temporo-parietal BOLD activity encodes semantic similarity at multiple timescales of natural speech[3]

3.2.1 Introduction

In natural speech comprehension, rich semantic context guides a listener's expectation on upcoming speech. However, previous research focused mostly on speech material with short timescales of context (e.g., isolated sentences). Here, we ask how the human brain orchestrates the multitude of semantic timescales underlying natural speech to build up predictions on upcoming speech.

Hasson, Chen & Honey (2015) found that larger timescales of speech like paragraphs are processed in higher parietal cortex, whereas short timescales like words are processed in lower temporal cortex. As frameworks of predictive coding (Rao & Ballard, 1999; Friston, 2005) propose that predictions are likewise formed along a processing hierarchy, we hypothesized that listeners exploit semantic context at multiple timescales of speech to inform speech prediction. More specifically, we expected that the timescales of semantic prediction organize along an auditory dorsal processing hierarchy.

3.2.2 Methods

Sixty-three participants (18–78 years) took part in an fMRI listening study. Here, we analyzed a subset of 30 younger participants (16 female, 18–31 years). All participants were right-handed healthy German native speakers. Data were reanalysed in Study 4, for a detailed description of experimental procedures, see section 3.4.

In the experiment, 30 participants listened to a one-hour narrative incorporating a multitude of timescales while confronted with a competing stream of resynthesized natural sounds at an SNR of 0 dB. The narrative was told by a female German speaker in eight separate blocks (9,446 words, including 4,451 content words). Each block was followed by three multiple-choice questions on the plot of the story.

In a first step, we adapted a measure of the semantic similarity between a word and its preceding context, which has been shown to modulate neural responses of speech

[3] This section was adapted from Schmitt et al. (2019).

comprehension (Frank & Willems, 2017; Broderick et al., 2018). To validate that this measure is a suitable proxy of word predictability, we compared the semantic similarity between sentence pairs with low vs. high constraint of context for the same final keyword. Sentences were adapted from an established stimulus set (Erb et al., 2012). For each word in a sentence, we calculated the similarity to its context by correlating the high-dimensional vector representation of its meaning (or embedding; Mikolov et al., 2013) with the average embedding of all the words preceding it. Final keywords of high predictability were more similar to their preceding sentence context when compared to low predictability (p < 0.0001; Figure 3.7), suggesting that word predictability is at least partially explained by semantic similarity.

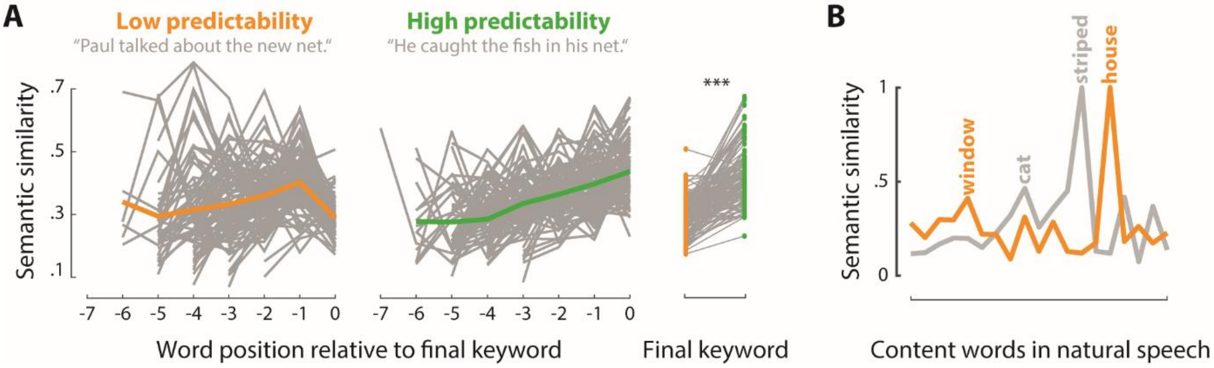


Figure 3.7. Evaluation of the semantic similarity metric. (A) Left: Semantic similarity between each word in a sentence and the average embedding of its preceding sentence context, separately for sentences with a final keyword of low (orange) vs. high predictability (green); grey lines indicate single sentences, fat lines indicate group average. Right: Semantic similarity of final keyword differentiates predictable (green) from unpredictable sentences (orange); *** p < 0.0001. **(B)** Semantic similarity of two exemplary content words to their neighbouring words in a natural story varies word by word. Similarity peaks at 1 indicate a time lag of 0 (i.e., target word).

To test for the hierarchical organization of semantic similarity at different timescales, we modelled semantic similarity at five timescales corresponding to a logarithmic increase in context length (i.e., 1–24 words) by computing the similarity between the embedding of each content word in the story and each timescale's average word embedding.

During the listening task, we acquired continuous whole brain 3 Tesla fMRI data (2.5 mm isotropic voxels, TR = 947 ms, TE = 28 ms). Initially, we calculated the intersubject correlation of the BOLD signal (Nastase et al., 2019) across the whole cortex to determine brain regions consistently engaged in speech processing. All further analyses were limited to those parcels (Glasser et al., 2016) along the temporo-parietal pathway with at least 60 % of vertices in the top 30 % of correlation coefficients. Next, we projected the BOLD signal onto the timescales of semantic similarity using vertex-wise ridge regression within a fourfold cross-validation scheme

(6 training blocks, 2 testing blocks). Individual best-timescale maps were derived from the encoding model, smoothed and submitted to a group-level cluster-permutation test.

3.2.3 Results

Along the auditory dorsal pathway, we found two distinct clusters ($p < 0.0001$): Increased activity in the posterior portion of superior temporal gyrus was coupled to short timescales similar to the next word, whereas parietal regions like the temporo-parietal junction and angular gyrus were most responsive to similar long timescales (Figure 3.8).

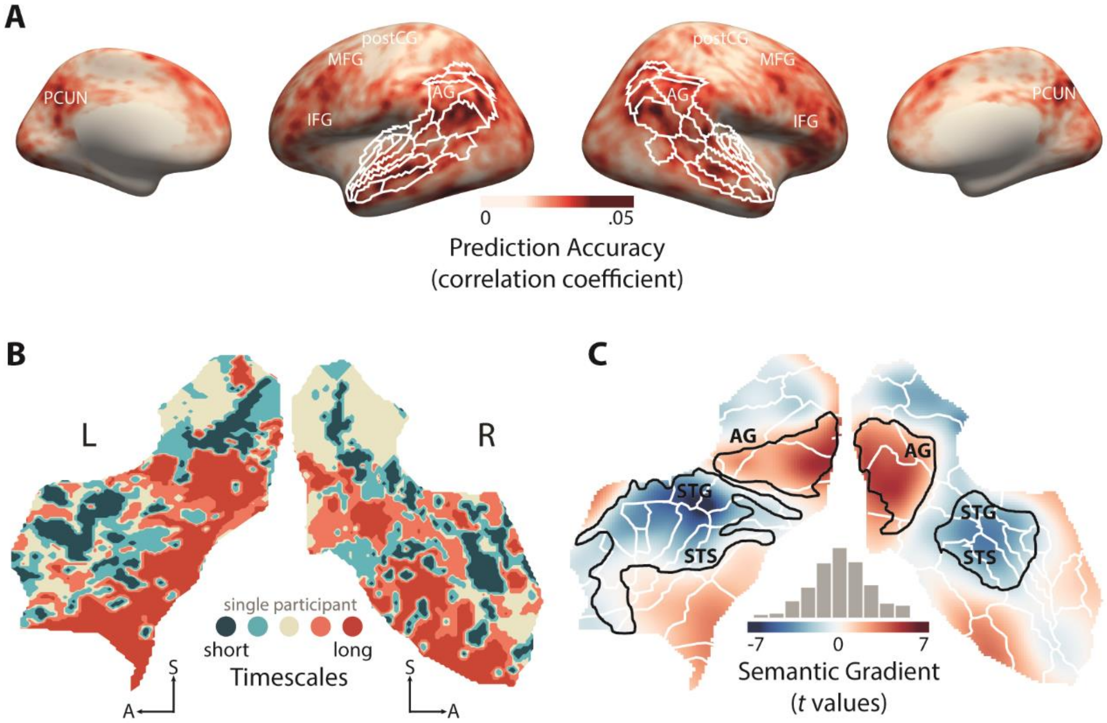


Figure 3.8. Encoding the timescales of semantic similarity. (A) Prediction accuracy of encoding models; white outlines indicate temporo-parietal region of interest. **(B)** Exemplary single-participant map of vertex-wise winning timescales. **(C)** Grand-average statistics calculated on best-timescale maps; blue regions indicate increased activity to informative short timescales, red regions indicate increased activity to informative long timescales, black outlines indicate significant clusters.

3.2.4 Conclusion

In this study, we showed that posterior temporal regions code for semantic similarity at short timescales, whereas parietal convergence zones code for long timescales. However, we found no hierarchy of similarity at different timescales, which would require a gradual increase of context length from temporal to parietal regions. One explanation for the absence of such a hierarchical effect might be that semantic similarity is not an appropriate measure of word predictability in the first place. For a detailed discussion of this result, see section 4.3. This result was replicated in Study 4 using another set of embeddings (see section 3.4).

3.3 From human to machine prediction: Modelling the predictiveness of context on multiple timescales[4]

3.3.1 Introduction

Our metric of semantic similarity turned out not to pick up on the predictability of language at multiple timescales (see Study 2). Nevertheless, we expected that, given a valid metric, the timescales of speech prediction evolve along a temporo-parietal processing hierarchy. The difficulty in deriving an adequate metric is twofold. First, the sheer number of words in our natural listening task was too large to derive a "handcrafted" predictability measure like cloze probability. Second, we were interested in the constraint of context at multiple timescales. While it is in general possible to manipulate how informative context is at specific timescales by scrambling speech at different timescales, this approach was not feasible in our case. First, scrambling speech renders specific timescales fully uninformative for the prediction of words but we were interested in the natural dynamic range of predictability. Second, we expected that the speech prediction hierarchy is sensitive to event-based compared to continuously updated context representations (see Study 4).

To this aim, we trained two artificial neural networks, which yielded word-by-word predictions at multiple timescales. I will briefly outline the architecture of artificial neural networks as well as how networks were trained and how predictions were read out. Resulting predictability metrics were applied to fMRI data in Study 4.

We trained two versions of a long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) with five layers to predict the next word in a story given a sequence of semantic context: a "continuously updating LSTM" where information is fed to a higher layer with each upcoming word, and a competing "sparsely updating HM-LSTM" where information is fed to a higher layer only at the end of a timescale (Chung et al., 2016). The predictiveness of context at each of the multiple timescales was read out from single layers of both language models for each word in the story presented to participants in experiments. Ultimately, we tested how closely these derivatives of different network architectures match signatures of behavioural and neural prediction processes in Study 4.

[4] This section was partly adapted from Schmitt et al. (2020).

3.3.2 Representing words in vector space

In natural language processing, it is common to represent a word by its linguistic features in the form of high-dimensional vectors (or embeddings). As the German language is morphologically rich and flexibly combines words into new compounds, there are many rare words for which language models cannot learn good (if any) vector representations on the word level. Therefore, we mapped all texts used for training, validating and testing our language models to pre-trained *sub*word vectors publicly available in the BPEmb collection (Heinzerling & Strube, 2017). These embeddings allow for the representation of *any* word by a combination of 100-dimensional subwords from a finite vocabulary of 100,000 subwords. We further reduced this vocabulary to those subwords that appeared at least once in any of the texts used for training, validating or testing our language models (i.e., number of subwords in vocabulary v = 91,645).

In short, the BPEmb vocabulary is based on a text corpus, which was segmented into its subwords using byte-pair encoding. That is, smaller subword units (e.g., letters) most frequently co-occurring in the corpus were iteratively merged into larger units (e.g., syllables) and added to the vocabulary until the predefined maximum of merge operations was reached (i.e., vocabulary size). The corresponding embeddings were trained with the GloVe algorithm (Pennington et al., 2014). Importantly, the length of subwords ranged from single letters to complete words. For example, the inflected verb “fischte” [“fished”; 3rd person singular, simple past, active voice, indicative of “to fish”] consists of one subword embedding representing its word stem “fisch” and another embedding representing its suffix “te”, whereas the more frequent word “Wasser” [“water”] is represented by only one embedding.

Matching our texts to subwords and their respective embeddings in the BPEmb vocabulary, yielded the embedded text $t \in R^{w \times e}$ where w is the number of words and $e = 100$ is the number of vector dimensions. On average, a word in the story was represented by 1.07 subwords (*Ra* = 1–6, *SD* = 0.33). As single words were encoded by only one subword in 94.25 % of cases, we will refer to subwords as words from here on.

3.3.3 Architecture of language models

When listening to a story, a fused representation of all spoken words $\{w_1, w_2, \ldots, w_p\}$ is maintained in memory and used as context information to make a prediction about the upcoming word w_{p+1}. In natural language processing, this memory formation is implemented via recurrent connections between the states of adjacent neural network cells. The hidden state h_{p-1} stores all relevant context and is sequentially passed to the next cell where it is updated with information from word w_p. More specifically, the recurrent input and the bottom-up input are combined into the cell input vector g_p:

$$g_p = tanh\left(W_g w_p + U_g h_{p-1} + b_g\right),$$

where *tanh* is the activation function, $W_g \in R^{n \times e}$ and $U_g \in R^{n \times n}$ are trainable weight matrices, n is the number of neurons (or units), $b \in R^{n \times 1}$ is a bias term, $w_p \in R^{e \times 1}$ and $h_{p-1} \in R^{n \times 1}$.

As such a simple recurrent neural network (RNN) tends to memorize only the most recent past, the more complex LSTM (Hochreiter and Schmidhuber, 1997) became a standard model in time series forecasting. In an LSTM cell, the state is split in two vectors: The cell state c_p acts as long-term memory, whereas the hidden state h_p incorporates information relevant to the cell output (i.e., the prediction of the next word). The integration of new information and the information flow between the two memory systems is controlled by three gating mechanisms. First, the cell state is updated. The forget gate f_p determines which information from the previous cell state c_{p-1} has become irrelevant and should be removed by:

$$f_p = \sigma\left(W_f w_p + U_f h_{p-1} + b_f\right),$$

where σ is the sigmoid activation function. The input gate i_p determines which information from candidate state g_p should be added to the cell state by:

$$i_p = \sigma\left(W_i w_p + U_i h_{p-1} + b_i\right).$$

The new cell state c_p is created by:

$$c_p = f_p \odot c_{p-1} + i_p \odot g_p,$$

where $c_{p-1} \in R^{n \times 1}$. Second, the hidden state is updated. The output gate o_p determines which information from long-term memory c_p might become relevant shortly and should be added to the hidden state by:

$$o_p = \sigma\left(W_o w_p + U_o h_{p-1} + b_o\right).$$

The new hidden state h_p is created by:

$$h_p = o_p \odot tanh(c_p).$$

When stacking multiple LSTM cells on top of each other, semantic context gets hierarchically organized in the model, with lower layers coding for short-term dependencies and higher layers coding for long-term dependencies between words. The bottom-up input to the first layer remains to be the embedded word w_p. However, the lower layer's hidden state h_p^{l-1} becomes the input to a cell from the second layer on. Importantly, the hidden state and cell state are updated at each layer with every new bottom-up input to the model.

A competing model that has been shown to slightly outperform the continuously updating ("vanilla") LSTM in character-level language modelling is the hierarchical multiscale LSTM (HM-LSTM; Chung et al., 2016). This model, referred to as "sparsely-updating HM-LSTM", employs a revised updating rule where information from the lower layer is only fed forward at the end of an event (i.e., a sequence of words closely related to each other). z_p^l is introduced which marks the end of a timescale:

$$\tilde{z}_p^l = hard\ sigm\left(z_p^{l-p} W_z h_p^{l-1} + U_z h_{p-1}^l + b_z^l\right),$$

$$z_p^l = \begin{cases} 1 & \text{if } \tilde{z}_p^l > 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where *hard sigm* is the hard sigmoid activation function. If $z_p^{l-1} = 1$, the hidden state h_p^l and cell state c_p^l are computed like in a vanilla LSTM cell ("update mechanism"). Otherwise, the hidden state h_p^l and cell state c_p^l are simply the copy of h_{p-1}^l and c_{p-1}^l ("copy mechanism"), respectively.

Importantly, The HM-LSTM allows for a sparse updating rate, with lower layers operating on short timescales and higher layers operating on longer timescales. Here, we used the simplified version of the HM-LSTM (Kádár et al., 2018) with no top-down connections.

3.3.4 Prediction of the next word

LSTM and HM-LSTM cells form the representations of semantic information relevant to speech prediction, whereas the actual prediction of the next word takes place in the output module. Here, hidden states at word position p are combined across the different layers of the language model by:

$$h_p^r = LReLU\left(\sum_{l=1}^{L} W_r^l h_p^l\right),$$

where $LReLU$ is the leaky rectified linear unit activation function and L is the number of layers. The combined hidden state h_p^r is mapped to a fully connected dense layer of as many neurons as there are words in the vocabulary and squashed to values in the interval [0,1], which sum to 1:

$$d_p = softmax(W_d h_p^r + b_d),$$

where *softmax* is the squashing function, $W_d \in R^{v \times n}$ and $b_d \in R^{v \times 1}$. Each neuron in vector d_p indexes one particular word in vocabulary v and denotes its probability of being the next word. Finally, the word referring to the highest probability in the distribution is chosen by:

$$s_p = argmax(d_p),$$

where s_p is the predicted next word in a story.

3.3.5 Training and evaluation of language models

The objective of our language models was to minimize the difference between the “predicted” probability distribution d_p (i.e., a vector of probabilities ranging from 0 to 1) and the “actual” probability distribution corresponding to the next word in a text (i.e., a vector of zeros with a one-hot encoded target word). To this end, we trained models on mini-batches of 16 independent text sequences à 500 words and monitored model performance by means of categorical cross-entropy between the “predicted” and “actual” probability distribution of each word in a sequence. Based on model performance, trainable parameters were updated after each mini batch using the Adam algorithm for stochastic gradient optimization (Kingma & Ba, 2017).

Our text corpus comprised more than 130 million words including 4,400 political speeches (Barbaresi, 2018) as well as 1,200 fictional and popular scientific books. All texts had at least 500 words; metadata, page numbers, references and punctuations (except for hyphenated compound words) were removed from documents. A held-out set of 10 randomly selected documents was used for validation after each epoch of training (i.e., going through the complete training set once) and allowed us to detect overfitting on the training set. Training automatically stopped after model performance did not increase over two epochs for the validation data set.

Using a context window of 500 words, we aimed at roughly modelling timescales of the length of common linguistic units in written language (i.e., words, phrases, sentences, and paragraphs). Therefore, we only used a small range of values from three to seven to find the number of layers—intended to represent distinct timescales—best suited to make good predictions. Additionally, we tuned the number of units in LSTM and HM-LSTM cells of language models, using values from 50 to 500 in steps of 50. Hyperparameters were evaluated on a single epoch using grid search and the best combination of hyperparameters was chosen based on performance on the validation set. Our final language models had five LSTM or HM-LSTM layers à 300 units and an output module. The LSTM model had 31,428,745 and the HM-LSTM model had 31,431,570 trainable parameters.

All other architectural choices were based on results from systematic ablation tests on HM-LSTM cells carried out by Kádár and colleagues (2018). Accordingly, the optimizer's initial learning rate of 0.001 was reduced by a factor of 0.02 when performance on the validation set did not improve over one epoch. Gradients were clipped at a value of 1. We applied layer normalization to all inputs after multiplication with their respective weight matrices (Ba et al., 2016). Further, we added an l^2-norm penalty term for weight size to the loss function ($\lambda = 0.0005$). The dense layer in the output module was excluded from normalization and regularization. Models were trained and evaluated with custom scripts in TensorFlow 2.1 (Abadi et al., 2015). For a detailed discussion of LSTM and HM-LSTM, see section 4.2.

3.4 Predicting speech from a cortical hierarchy of event-based timescales[5]

3.4.1 Abstract

How can anticipatory neural processes structure the temporal unfolding of context in our natural environment? We here provide evidence for a neural coding scheme that sparsely updates contextual representations at the boundary of events and gives rise to a hierarchical, multi-layered organization of predictive language comprehension. Training artificial neural networks to predict the next word in a story at five stacked timescales and then using model-based functional MRI, we observe a sparse, event-based "surprisal hierarchy". The hierarchy evolved along a temporo-parietal pathway, with model-based surprisal at longest timescales represented in inferior parietal regions. Along this hierarchy, surprisal at any given timescale gated bottom-up and top-down connectivity to neighbouring timescales. In contrast, surprisal derived from a continuously updated context influenced temporo-parietal activity only at short timescales. Representing context in the form of increasingly coarse events constitutes a network architecture for making predictions that is both computationally efficient and semantically rich.

[5] This section was partly adapted from Schmitt et al. (2020).

3.4.2 Introduction

While the past predicts the future, not all context the past provides is equally informative: it might be outdated, contradictory, or even irrelevant. Nevertheless, the brain as a "prediction machine" (Clark, 2013a, 2013b) is seemingly equipped with a versatile repertoire of computations to overcome these contextual ambiguities. A prominent example is speech, where a slip of the tongue may render the most recent context uninformative, but we can still predict the next word from its remaining context. At much longer time scales, we can re-use context that suddenly proves informative, as a speaker returns to a topic discussed earlier.

Using natural language comprehension as a working model, we here ask: How does the brain dynamically organize, evaluate, and update these complex contextual dependencies over time to make accurate predictions?

A robust principle in cerebral cortex is the decomposition of temporal context into its constituent timescales along a hierarchy from lower to higher-order areas, which is evident across species (Murray et al., 2014; Zhang & Yartsev, 2019), recording modalities (La Camera et al., 2006; Burt et al., 2018), sensory modalities (Lakatos et al., 2005; Mattar et al., 2016), and cognitive functions (Lamme & Roelfsema, 2000). For instance, sensory cortices closely track rapid fluctuations of stimulus features and operate on short timescales (e.g., Buračas et al., 1998). By contrast, association cortices integrate stimuli over an extended period and operate on longer timescales (e.g., Runyan et al., 2017).

Conceptually, such hierarchies of "temporal receptive windows" are often subsumed under the framework of predictive coding (Friston, 2005): A nested set of timescale-specific generative models informs predictions on upcoming sensory input and is updated based on the actual input (Keller & Mrsic-Flogel, 2018). In particular, context is thought to shape the prediction of incoming stimuli via feedback connections. These connections would link each timescale to its immediate shorter timescale, while the prediction error is propagated forward through the hierarchy (Clark, 2013b; Kiebel et al., 2008). Indeed, hierarchical specialization has been shown empirically to emerge from structural and functional large-scale connectivity across cortex (Chaudhuri et al., 2015; Demirtaş et al., 2019). More precisely, feedforward and feedback connections (Bastos et al., 2015; Cocchi et al., 2016) shown to carry prediction errors and predictions (Wacongne et al., 2011; Schwiedrzik & Freiwald, 2017), respectively, are a hallmark of hierarchical predictive coding.

However, studies on the neural underpinnings of predictive coding have primarily used artificial stimuli of short temporal context (but see Donhauser & Baillet, 2020) and employed local vs. global violations of expectations, effectively manifesting a two-level cortical hierarchy (but see Chao et al., 2018). We thus lack understanding whether the hierarchical organization of prediction processes extends to natural environments unfolding their temporal dependencies over a multitude of interrelated timescales.

With respect to functional organization in cortex, temporo-parietal areas are sensitive to a rich set of hierarchies and timescales in speech (Honey et al., 2012; Stephens et al., 2013; Mesgarani et al., 2014; Huth et al., 2016; de Heer et al., 2017). Most relevant to the present work, semantic context in a spoken story has been shown to map onto a gradient extending from early auditory cortex representative of words up to intraparietal sulcus, representative of paragraphs (Lerner et al., 2011). This timescale-specific representation of context is reminiscent of the multi-layered generative models proposed to underlie predictive coding (Bornkessel-Schlesewsky et al., 2015; Kuperberg & Jaeger, 2016). Compatible with this notion, previous studies on speech comprehension found evidence for neural coding of prediction errors at the level of syllables (Arnal et al., 2011), words (Blank & Davis, 2016), or discourse (Kandylaki et al., 2016).

Yet the interactions between multiple representational levels of speech in predicting upcoming words remain unclear. Here, we ask whether the processing hierarchy enabling natural speech comprehension is also implicated in evaluating the predictiveness of timescale-specific semantic context and integrating informative context into predictions.

We do not know how context that unfolds at a particular timescale would be updated cortically when the listener receives new bottom-up input. One attractively simple candidate architecture is the *continuously updating processing hierarchy*. In a recent study, Chien and Honey (2020) showed that neural responses to a story rapidly aligned across participants in areas with shorter, but only later in areas with longer receptive windows. This response pattern was best explained by a computational model, which immediately integrates upcoming input with context representations at all timescales. An important implication of such continuous updates is that all context representations are continuously tuned to current processing demands.

A competing candidate architecture, however, is the *sparsely updating processing hierarchy*. For example, it is known that scenes in a movie are encoded as event-specific neural responses (Zadbood et al., 2017) and that more parietal receptive windows represent increasingly coarse events in movies (Baldassano et al., 2017). Such an event hierarchy is effectively based on the boundaries of events: It calls for stable working memory representations that are *sparsely* recombined with preceding events at higher processing stages only at the end of an event. The simultaneous representation of distinct events in working memory allows to draw on diverse context when making predictions. We here hypothesize that such a sparsely updating network architecture is a more appropriate model for prediction processes in the brain.

In the present study, we recorded BOLD responses while participants listened to a narrated story, which provides rich semantic context and captures the full dynamic range of speech predictability (Hamilton & Huth, 2018). Following the rationale that neural computations can be inferred by comparing the fit of neural data to outputs from artificial neural networks with different architectures (Cohen et al., 2017; Cichy & Kaiser, 2019), we derived context-specific surprisal associated with each word in the story from single layers of long short-term memory (LSTM)-based language models with either a continuous or a sparse updating rule.

Here, we show that the event-based organization of semantic context provides a valid model of predictive processing in the brain. We show that a “surprisal hierarchy” of increasingly coarse event timescales evolves along the temporo-parietal pathway, with stronger connectivity to neighbouring timescales in states of higher word surprisal. Surprisal derived from continuously updated context had a (non-hierarchical) effect on temporo-parietal activity only at short timescales. Our results suggest that representing context in the form of increasingly coarse events constitutes a network architecture that is both, computationally efficient and semantically rich for making predictions.

3.4.3 Results

Thirty-four participants listened to a one-hour narrated story while their hemodynamic brain responses were recorded using functional magnetic resonance imaging (fMRI). To emulate a challenging listening scenario, the story audio was presented against a competing stream of resynthesized natural sounds (for an analysis focussing on this acoustic representation see Erb et al., 2020).

The surprisal associated with each word in the story was modelled at multiple timescales of semantic context by two artificial neural networks, one with a continuous updating rule (LSTM) and another one with a sparse updating rule (HM-LSTM; Figure 3.9A).

First, we encoded surprisal at multiple timescales into univariate neural responses and fit a gradient to temporo-parietal peak locations of timescales. Second, we decoded timescale surprisal from patterns of neural responses in single parcels and compared decoding accuracies between language models. Finally, we investigated how surprisal gates the information flow between brain regions sensitive to different timescales.

All encoding and decoding models were estimated separately per each language model, using ridge regression with four-fold cross-validation.

Two competing language models of hierarchical speech prediction

We trained two artificial neural networks on more than 130 million words of running text to predict an upcoming word by its preceding semantic context. More specifically, language models consisted of long short-term memory cells (LSTM; Hochreiter & Schmidhuber, 1997), which incorporate context that might become relevant at *some* time (*cell state*) or that is relevant already to the prediction of the *next* word (*hidden state*). By stacking five LSTM layers, our models operated on different timescales of context, with higher layers coding for long-term dependencies between words.

In the continuously updating (or "vanilla") LSTM, recurrent memory states are updated at each layer with every new bottom-up word input (Figure 3.9B). A second model, the *hierarchical multiscale* LSTM (Chung et al., 2016), referred to as "sparsely updating HM-LSTM", employs a revised updating rule where information from a lower layer is only fed forward at the end of its representing timescale (Figure 3.9C). This allows for less frequent updates between layers and stronger separation between contextual information represented at different layers.

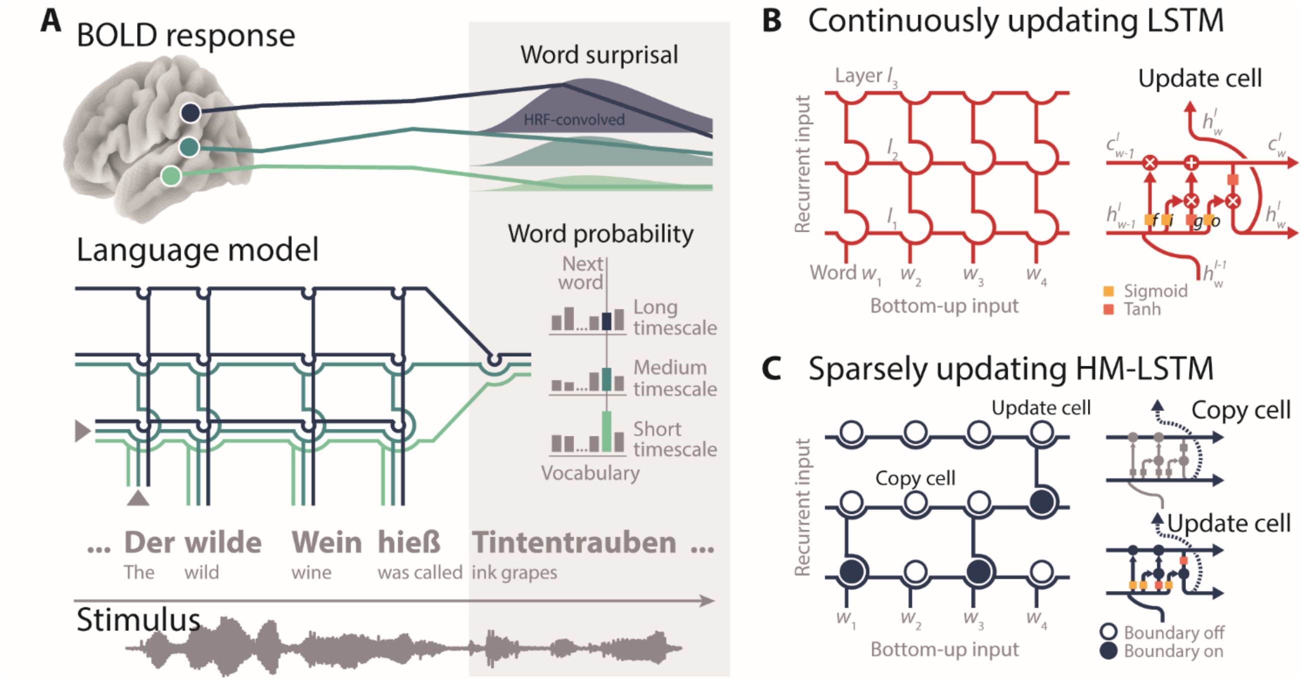


Figure 3.9. Modelling neural speech prediction with artificial neural networks. (A) Participants listened to a story (grey waveform) during fMRI. Based on its semantic context ("The wild wine was called"), a language model predicted each word in the story ("ink grapes"). The probability of the next word was read out from each layer of the model separately, with higher layers accumulating information across longer semantic timescales. Word probabilities were transformed to surprisal, convolved with the hemodynamic response function and mapped onto temporo-parietal BOLD time series. **(B)** Two language models were trained. With each new word-level input, the "continuously updating" long short-term memory (LSTM; Hochreiter & Schmidhuber, 1997) combines "old" recurrent long-term (c^l_{w-1}) and short-term memory states (h^l_{w-1}) with "new" bottom-up semantic input (h^{l-1}_w) at each layer *l*. This allowed semantic information to continuously flow to higher layers with each incoming word. f: forget gate, i: input gate, g: candidate state, o: output gate. **(C)** The "sparsely updating" hierarchical multiscale LSTM (HM-LSTM)(Chung et al., 2016) was designed to learn the hierarchical structure of text. An upper layer keeps its representation of context unchanged (copy mechanism) until a boundary indicates the end of a timescale on the lower layer and information is passed to the upper layer (update mechanism). Networks were unrolled for illustration only.

Three model-derived metrics of predictiveness at multiple timescales

For each word in the entire presented story (> 9,000 words), we determined its predictability given the semantic context of the preceding 500 words. Hidden states were combined across layers and mapped to an output module, which denotes the probability of occurring next for every word in a large vocabulary of candidate words. The word with the highest probability was selected as the *predicted* next word. Overall, the LSTM (proportion correct across words: 0.13) and the HM-LSTM (0.12; Supplementary Figure 3.8) were on par in accurately predicting the next word in the story.

To derive the predictability of words based on layer-specific context (or, for our purpose, timescale), we allowed information to freely flow through pre-trained networks, yet only mapped the hidden state of one layer to the output module by setting all other network weights to zero. Outputs from these "lesioned" language models represented the five timescales.

As the primary metric of predictiveness, we calculated the degree of surprisal associated with the occurrence of a word given its context (i.e., negative logarithm of the probability assigned to the *actual* next word). The surprisal evoked by an incoming word indexes the amount of information that was not predictable from the context represented at a specific timescale (Hale, 2001; Levy, 2008). Of note, surprisal was considerably higher for longer timescales in the LSTM ($p < 0.001$, Cohen's $d = 2.43$; compared to slopes drawn from surprisal shuffled across timescales) but remained stable across timescales in the HM-LSTM ($p = 0.955$, $d = 0.05$; direct comparison LSTM vs. HM-LSTM: $p < 0.001$, $d = 2.7$; Figure 3.10).

To determine the temporal integration window of each timescale, we scrambled input to the network at different granularities corresponding to a binary logarithmic increase in the length of intact context (i.e., 1–256 words). The LSTM showed no such increase of temporal integration windows at higher layers (LSTM: $p = 0.219$, $d = 0.11$). In contrast, in the HM-LSTM, surprisal decreased more strongly at longer compared to shorter timescales as more intact context became available (HM-LSTM: $p = 0.027$, $d = 0.73$; LSTM vs. HM-LSTM: $p < 0.001$, $d = 0.76$; Figure 3.10).

Our secondary metrics expressed the predictability of a word in relation to other words, that is, (1) the entropy of the probability distribution predicted for individual words (indicative of the difficulty to make a definite prediction) and (2) the dissimilarity of vector representations (or embeddings) coding for the constituent linguistic features of the predicted and actual next word (Product-moment correlation; indicative of conceptual (un-)relatedness).

We derived surprisal, entropy and dissimilarity associated with single words from "lesioned" models at each of five timescale and from "full" models across all timescales, separately for each language model. All features were convolved with the hemodynamic response function, and we will collectively refer to them as "features of predictiveness" from here on.

Higher model-derived surprisal of words slows down reading

To test the behavioural relevance of model-based predictiveness, another 26 participants performed a self-paced reading task where they read the transcribed story word-by-word on a noncumulative display and pressed a button as soon as they had finished reading.

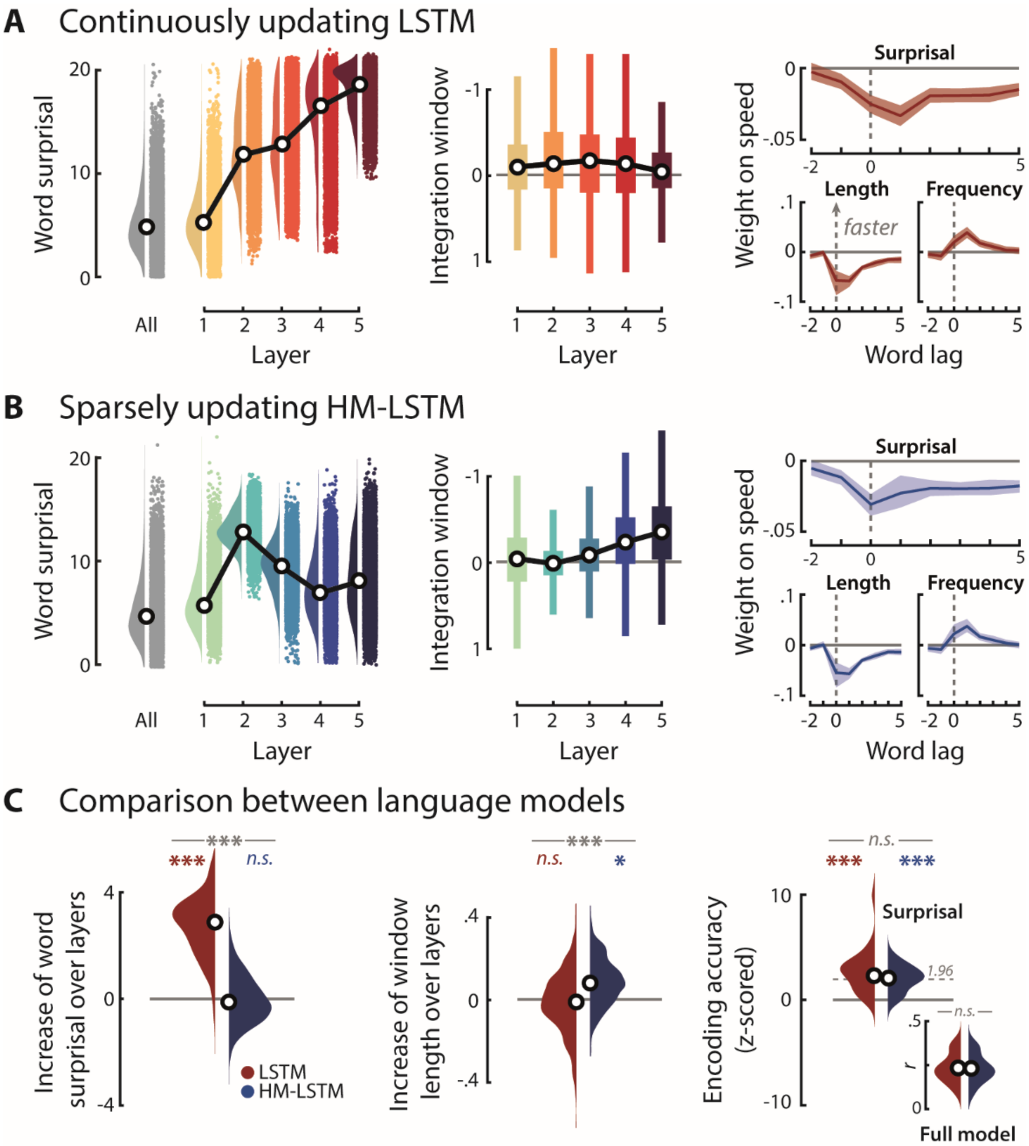


Figure 3.10. Evaluating model-derived surprisal. (A) Left: Word surprisal derived from the "full" LSTM model including all layers (grey distribution) and from single layers of "lesioned" LSTM models (coloured distributions); black circles represent grand-median surprisal. Middle: Input to the LSTM was scrambled at different granularities corresponding to an increase in the length of intact context (i.e., 1–256 words). For each layer of the LSTM, linear functions were fit to word surprisal across these context windows. A negative slope parameter indicates a stronger benefit (or lower surprisal) from longer context (i.e., larger integration window). Right: Speed in a self-paced reading task was modelled as a function of time-lagged predictiveness and a set of nuisance regressors. Weight profiles illustrate the temporal dynamics of the surprisal effect in the full model (top) in comparison to word length (bottom, left) and word frequency (bottom, right); positive weights indicate an increase in response speed; error bands represent ±*SEM*. **(B)** Same as in A, but for the sparsely updating HM-LSTM. **(C)** Left: We fit linear functions to word surprisal across layers and compared resulting slope parameters to null distributions drawn from shuffled layers and between LSTM (red) and HM-LSTM (blue). Middle: Linear fit to integration windows across layers, indicating the benefit of higher layers from longer context. Right: Encoding accuracy in the self-paced reading task uniquely explained by the predictiveness of context (standardized to scrambled features of predictiveness); dotted grey line indicates critical significance level for single participants. Inset shows non-standardized encoding accuracies. *** $p < 0.001$, * $p < 0.05$, *n.s.*: not significant.

When regressing response speed onto time-lagged features of predictiveness and a set of nuisance regressors (e.g., word length and frequency), we found that—as expected—reading speed slowed down for words determined as more surprising by language models given the full context across all timescales (Figure 3.10A).

Further, we predicted response speed on held-out testing data and z-scored the resulting encoding accuracy (i.e., Product-moment correlation of predicted and actual response speed) to a null distribution drawn from scrambled features of predictiveness while only keeping nuisance regressors intact. This yielded the unique contribution of the predictiveness of words (i.e., surprisal, entropy and dissimilarity) to reading speed, which was significant for both language models (LSTM: $p < 0.001$, $d = 1.51$; HM-LSTM: $p < 0.001$, $d = 1.64$, LSTM vs. HM-LSTM: $p = 0.975$, $d = 0.35$). Together, these findings suggest that both language models picked up on processes of speech prediction that shape behaviour.

Selecting temporo-parietal regions of interest involved in speech processing

We hypothesized that the speech prediction hierarchy is represented as a gradient along the temporo-parietal pathway. This rather coarse region of interest was further refined to only include regions implicated in processing of the listening task.

To this aim, we calculated pairwise intersubject correlations (Nastase et al., 2019), which revealed consistent cortical activity across participants in a broad bilateral language network. Responses were most prominently shared in auditory association cortex and lateral temporal cortex as well as premotor cortex, paracentral lobule and mid cingulate cortex (Figure 3.11A). Crucially, as sound textures presented in the competing stream were randomly ordered across participants, this approach allowed us to extract shared responses specific to the speech stream.

All further analyses were limited to those parcels in temporal and parietal cortex (Glasser et al., 2016) that showed significant median intersubject correlations in more than 80 % of vertices ($p_{\text{FDR}} < 0.01$; ranked against a bootstrapped null distribution). The cortical sheet of the six parcels determined as regions of interest (ROI) was flattened, resulting in a two-dimensional plane spanned by an anterior-posterior and inferior-superior axis (Figure 3.11B).

We expected gradients of speech prediction to unfold along the inferior-superior axis, that is, from temporal to parietal areas.

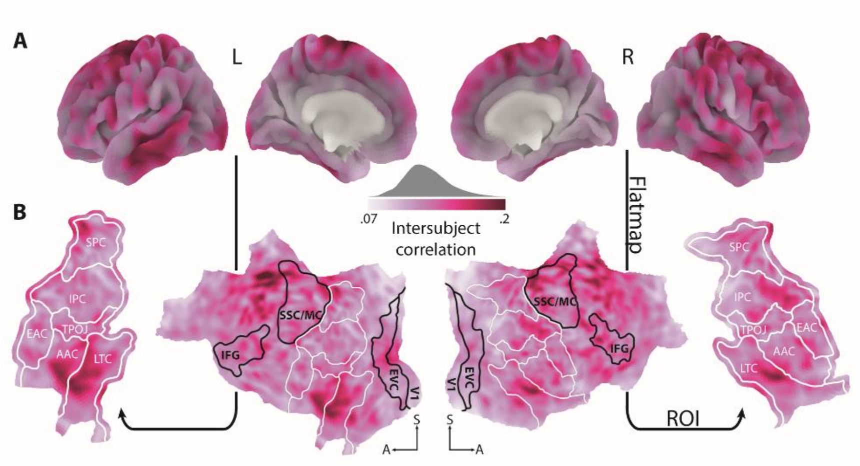


Figure 3.11. Selection of regions of interest. (A) When listening to a story against background noise, pairwise intersubject correlations showed stronger synchronization of BOLD activity in cortical areas implicated in the language network. **(B)** The cortical surface was flattened. All temporal and parietal parcels (Glasser et al., 2016) highlighted by white outlines were included as regions of interest (ROI) in the following analyses. Black outlined parcels serve as reference point, only. EAC: early auditory cortex, AAC: auditory association cortex, LTC: lateral temporal cortex, TPOJ: temporo-parieto-occipital junction, IPC: inferior parietal cortex, SPC: superior parietal cortex, V1 : primary auditory cortex, EVC: early visual cortex, IFG: inferior frontal gyrus, SSC/MC: somatosensory and motor cortex. Maps were smoothed with an 8 mm FWHM Gaussian kernel for illustration only.

Differential tuning to continuously and sparsely updated timescales of surprisal in temporo-parietal cortex

After the predictiveness of speech was encoded into neural activity of single vertices, we extracted temporo-parietal weight maps of word surprisal at each timescale for both language models.

When performing spatial clustering on these weight maps (p_{vertex} and $p_{\text{cluster}} < 0.05$; compared to scrambled surprisal by means of a cluster-based permutation test), we found large positive clusters in both hemispheres for shorter timescales of the LSTM (Figure 3.12A, yellow outlines) but, if at all, only focal clusters for longer timescales (Figure 3.12A, red outlines). Hence, temporo-parietal activity primarily increased in response to words that were less predictable by the context provided at shorter, continuously updated timescales. In contrast, clusters of distinct polarity, location and extent were observed for all timescales of the HM-LSTM (Figure 3.12B), suggesting that even longer timescales had the potency to modulate temporo-parietal activity when they were sparsely updated.

Sparsely updated timescales of surprisal evolve along a temporo-parietal processing hierarchy

To probe the organization of timescales along a temporo-parietal gradient, we collapsed across the anterior-posterior axis of weight maps and selected the local maximum with the largest positive value on the inferior-superior axis. Fitting a linear function to those peak coordinates of timescales, we found flat slope parameters indicating random ordering of LSTM timescales in both hemispheres (left: p = 0.458, d = -0.15; right: p = 0.716, d = -0.07; compared to slopes drawn from coordinates scrambled across timescales, Figure 3.12C). Conversely, we found steep positive slopes for the HM-LSTM in both hemispheres (left: p < 0.001, d = 0.72; right: p < 0.001, d = 0.75; Figure 3.12C), reflecting the representation of longer timescales in more parietal regions. On average, the left hemisphere represented sparsely updated timescales 12 mm superior (along the unfolded cortical surface) to their directly preceding timescale. Most relevant, this finding was underpinned by a significant difference of slope parameters between the LSTM and HM-LSTM (left: p = 0.005, d = 0.9; right: p < 0.001, d = 0.89; Figure 3.12C), demonstrating a temporo-parietal processing hierarchy of word surprisal that preferably operates on sparsely updated timescales.

The absence of a gradient for continuously updated timescales was corroborated when specifically targeting the dorsal processing stream. To this aim, we confined the first timescale to peak in temporal regions and all other timescales to peak superior to the first timescale in more parietal regions. Slope effects along the dorsal stream largely complied with those found along the inferior-superior axis (LSTM: all p ≥ 0.134; HM-LSTM: all p ≤ 0.006; LSTM vs. HM-LSTM: all p ≥ 0.103), thereby ruling out the possibility that the presence of a competing ventral stream obscured the consistent ordering of timescales.

Additionally, rotating weight maps by -45° before collapsing across the first dimension showed that sparsely updated timescales were not only processed along an inferior-superior but also an anterior-posterior gradient in the left hemisphere (LSTM: p = 0.921, d = 0.02; HM-LSTM: p = 0.001, d = 0.65; LSTM vs. HM-LSTM: p = 0.011, d = 0.55). As right-hemispheric maps already had a strong initial rotation off the inferior-superior axis, rotating these maps merely confirmed that longer timescales are processed in more superior regions (LSTM: p = 0.355, d = -0.18; HM-LSTM: p < 0.001, d = 1.39; LSTM vs. HM-LSTM: p = 0.039, d = 1.45).

Unlike for the sparsely updated timescales of surprisal, neither the timescales of entropy (all p ≥ 0.583) nor dissimilarity (all p ≥ 0.623) organized along a dorsal gradient (Supplementary

Figure 3.9). Further, effects of HM-LSTM timescale surprisal were dissociable from a simple measure of semantic incongruence between words in the story and their context at five timescales logarithmically increasing in length (all $p \geq 0.5$; Product-moment correlation of target and average context embedding). This highlights the specificity of the observed gradient to prediction processes in general and word surprisal in particular.

A segregated stream of continuously updated timescales of surprisal?

To determine the contribution of predictiveness to overall encoding accuracy on held-out data, we z-scored accuracies relative to null distributions drawn from scrambled features of predictiveness while keeping additional (spectro-temporal) acoustic and linguistic nuisance regressors intact. Interestingly, the LSTM produced—in comparison to the HM-LSTM—better predictions in early auditory cortex and supramarginal gyrus (p_{vertex} and $p_{\text{cluster}} < 0.05$; cluster-based permutation paired-sample *t*-test; Figure 3.12C). On the other hand, predictions of the HM-LSTM seemed slightly more accurate than for the LSTM along middle temporal gyrus, temporo-parieto-occipital junction and angular gyrus, even though not statistically significant. Taking into account the broad clusters found earlier specifically for shorter (but not longer) LSTM timescales, this poses the question whether continuously updated timescales take full effect only in earlier medial temporal and anterior parietal processing stages, whereas the sparsely updating processing hierarchy evolves along a separate lateral temporal and posterior parietal route.

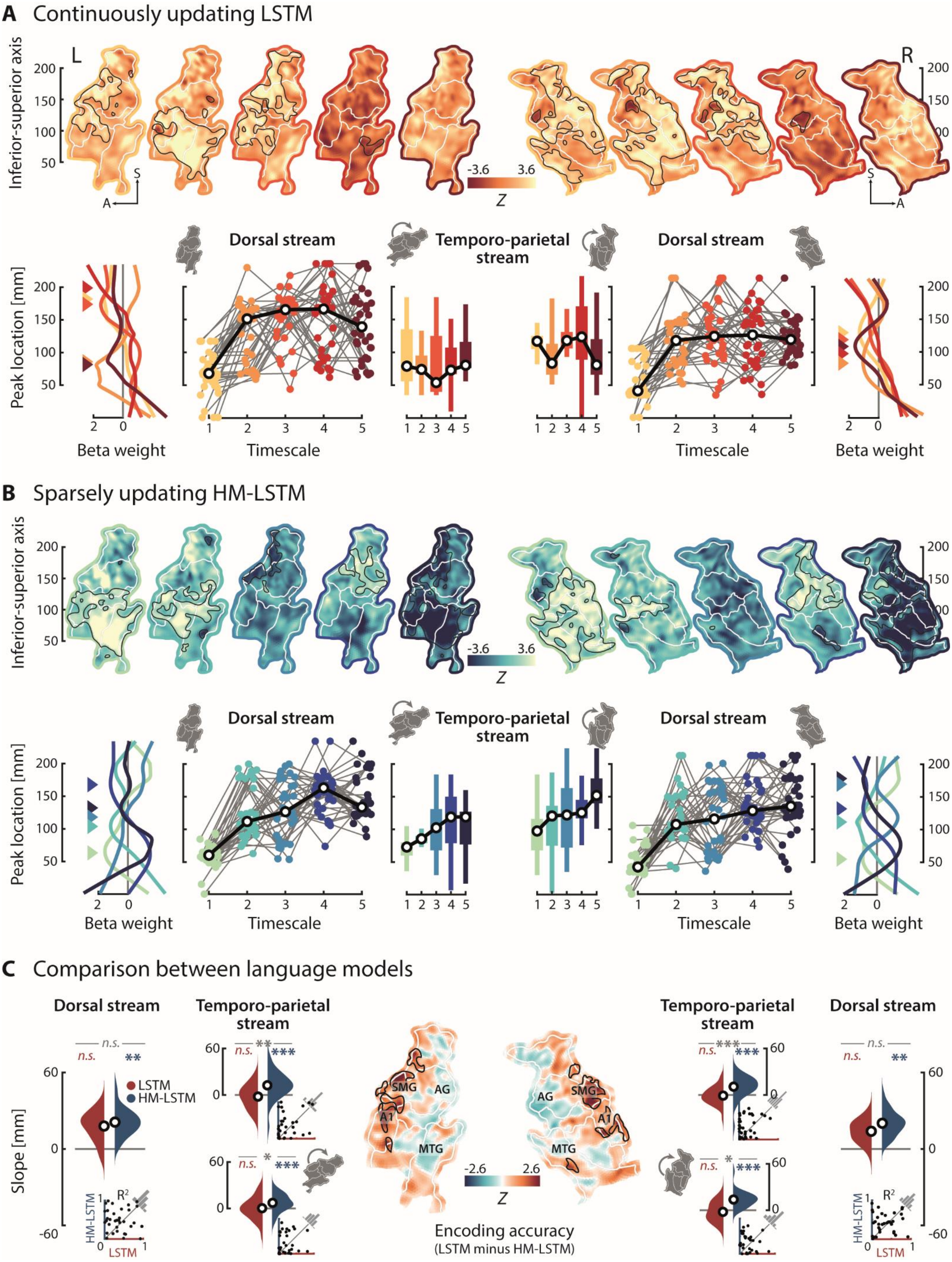

A Continuously updating LSTM
Inferior-superior axis
L
R
200
150
100
50
S
A
-3.6
3.6
Z
Dorsal stream
Temporo-parietal stream
Dorsal stream
Peak location [mm]
200
150
100
50
2
0
Beta weight
1
2
3
4
5
Timescale
Beta weight
B Sparsely updating HM-LSTM
Inferior-superior axis
200
150
100
50
-3.6
3.6
Z
Dorsal stream
Temporo-parietal stream
Dorsal stream
Peak location [mm]
Beta weight
Timescale
Timescale
Beta weight
C Comparison between language models
Dorsal stream
Temporo-parietal stream
Temporo-parietal stream
Dorsal stream
n.s.
n.s.
**
***
*
LSTM
HM-LSTM
Slope [mm]
60
0
-60
R²
HM-LSTM
LSTM
0
1
SMG
AG
A1
MTG
-2.6
2.6
Z
Encoding accuracy
(LSTM minus HM-LSTM)

Figure 3.12. Encoding the timescales of surprisal. (A) Temporo-parietal BOLD time series were mapped onto the predictiveness of speech derived at five timescales from the continuously updating LSTM. Top row: Maps show *z*-values from timescale-specific weights of surprisal tested against a null distribution drawn from scrambled surprisal; black outlines indicate significant clusters; white outlines indicate parcels; coloured outlines indicate short (light yellow) to long (dark red) timescales, separately for the left and right hemisphere. Bottom row: For each timescale, we determined its peak coordinate along the inferior-superior axis (coloured triangles), here shown for grand-average weight profiles for illustration only. Testing for a processing hierarchy along the dorsal stream, timescales were restricted to peak superior to the first timescale; coloured dots connected by grey lines represent peak coordinates of single participants; black circles represent grand-median peak coordinates. Testing for a processing hierarchy along the temporo-parietal stream, timescales were allowed to peak at any location. Maps were rotated by -45° to test simultaneous effects on the inferior-superior and anterior-posterior axis in the left hemisphere. Note that right-hemispheric maps already had an initial rotation off the inferior-superior axis, so rotating these maps resulted in testing for effects on the inferior-superior axis only. **(B)** Encoding maps and timescale-specific peak locations for the sparsely updating HM-LSTM. **(C)** Linear functions were fit to peak coordinates across timescales and resulting slope parameters were compared to null distributions drawn from scrambled coordinates and between language models, separately for the dorsal (not rotated) and temporo-parietal stream (top: not rotated; bottom: rotated) in both hemispheres. Black circles represent grand-average slope parameters; insets depict coefficients of determination for linear fits of single participants. Temporo-parietal encoding accuracies displayed on ROI maps were z-scored to null distributions drawn from scrambled features of predictiveness and compared between language models; black outlines indicate significant clusters; maps were smoothed with an 8 mm FWHM Gaussian kernel for illustration only; SMG: supramarginal gyrus, AG: angular gyrus, A1: primary auditory cortex; MTG: middle temporal gyrus. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, *n.s.*: not significant.

Parietal regions preferentially represent sparsely updated, long timescales

In a complementary decoding approach, we reconstructed the timescales of surprisal from patterns of neural activity in single regions of interest (i.e., temporo-parietal parcels). Reconstructed timescales were z-scored to scrambled features of predictiveness. Overall, bilateral auditory association cortex and lateral temporal cortex yielded highest decoding accuracies on held-out testing data for both language models (Figure 3.13A). More parietal regions showed comparably lower decoding accuracies. Nevertheless, these accuracies can be deemed meaningful, as the pattern mirrors the lower intersubject correlations in parietal compared to temporal regions (Figure 3.11), which are commonly found during natural listening (e.g., Boldt et al., 2013; Schmälzle et al., 2015; Regev et al., 2018). This indicates an overall greater variability of neural responses in parietal regions irrespective of the timescales of surprisal.

Contrasting decoding accuracies between language models, temporo-parietal regions of interest showed an overall preference for the shortest LSTM but longer HM-LSTM timescales (Figure 3.13B). This suggests that the predominance of the LSTM in early auditory cortex and supramarginal gyrus observed for encoding accuracies of the encoding model is specific to the shortest timescale, while lateral temporal and posterior parietal regions reflect longer timescales of the HM-LSTM, lending further support to the functional dissociation of two routes of predictive processing in speech prediction.

In particular, left-hemispheric early auditory cortex contained more information on medium, sparsely updated than continuously updated timescales, whereas inferior and superior parietal cortex preferentially represented long, sparsely updated timescales ($p_{\mathrm{FDR}} < 0.05$). This finding converges with the organization of the gradient described earlier for sparsely updated but not continuously updated timescales.

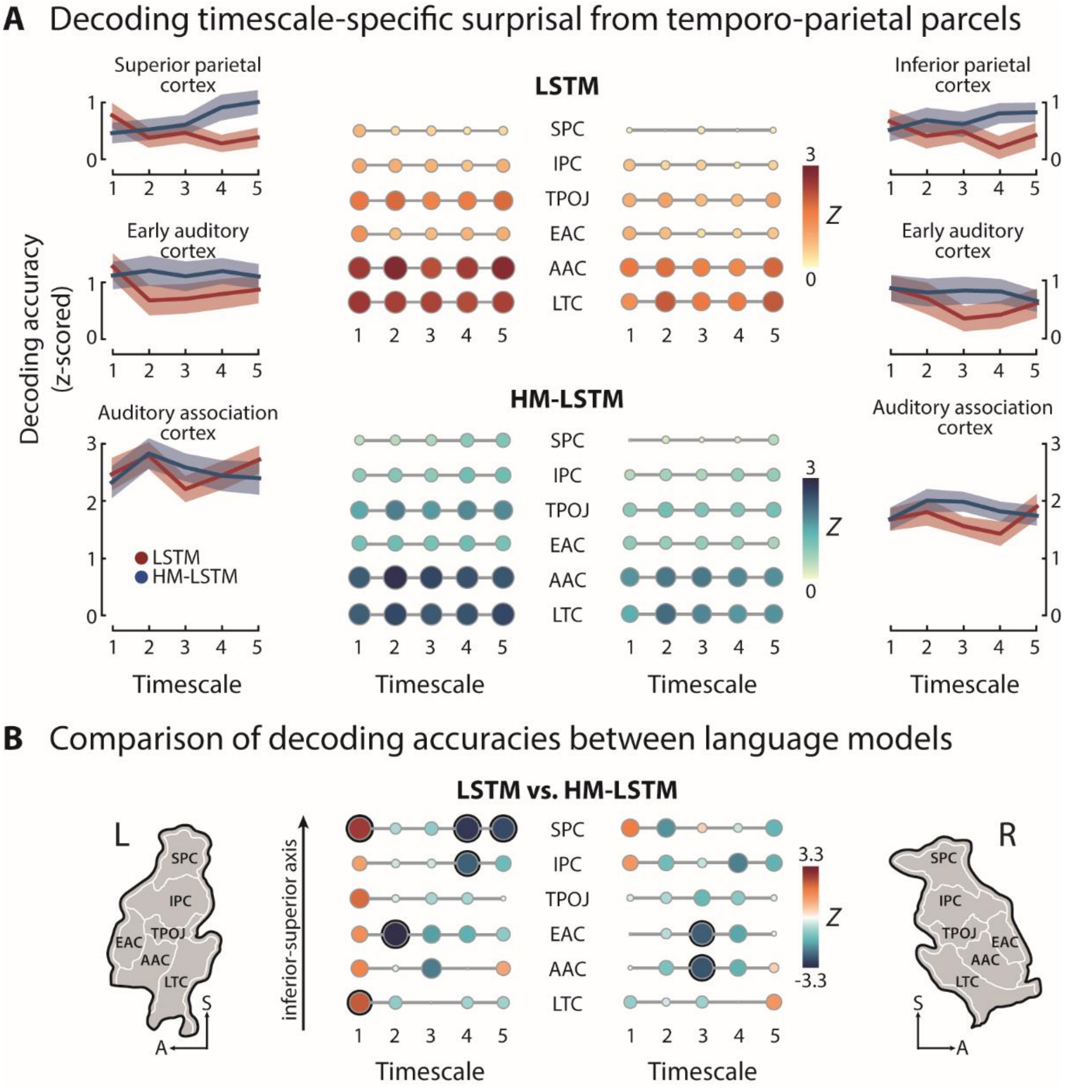


Figure 3.13. Decoding the timescales of surprisal. (A) Surprisal at different timescales was decoded from regions of interest. Matrices depict decoding accuracies determined on held-out testing data and z-scored to null distributions drawn from scrambled surprisal, separately for the LSTM (top) and HM-LSTM (bottom) in the left and right hemisphere; colour and size of circles scale to decoding accuracy. Of note, some z-scored decoding accuracies in more superior parcels fell below an average value of 1.96. However, z-scores were indicative of significance only on the level of single participants. Line plots illustrate patterns of decoding accuracies across timescales in select regions of interest; error bands represent ±*SEM*. **(B)** Decoding accuracies were contrasted between language models by means of a permutation test on the mean of differences; black circles indicate $p_{\mathrm{FDR}} < 0.05$; maps indicate location of parcels. EAC: early auditory cortex, AAC: auditory association cortex, LTC: lateral temporal cortex, TPOJ: temporo-parieto-occipital junction, IPC: inferior parietal cortex, SPC: superior parietal cortex.

Surprisal at sparsely updated timescales gates connectivity along the processing hierarchy

After establishing the temporo-parietal processing hierarchy, we examined the modulatory effect of surprisal on connectivity between peak locations of timescales taken from the encoding analysis. To this aim, we created psychophysiological interactions between the BOLD response at the peak location of one timescale and word surprisal at the same timescale. The BOLD response of each (target) timescale was mapped onto psychophysiological interactions of all other (predictor) timescales (Figure 3.14A).

We hypothesized that coupling between brain regions representing two neighbouring timescales increases when one timescale becomes unpredictive. Numerically, this can be expressed by setting the weights of neighbouring timescales to 1 and all other predictor weights to -1 (Figure 3.14B). This hypothesized pattern of weights was not matched by the weights observed for the LSTM (left: p = 0.83, d = 0.27; right: p = 0.348, d = 0.1; Euclidean distance compared to null distributions drawn from target BOLD activity shifted in time), which was expected given that the continuously updated timescales were not organized along a gradient in the first place. Critically, for sparsely updated timescales of the HM-LSTM, surprisal-modulated connectivity in the left hemisphere not only matched our hypothesis (left: p = 0.032, d = 0.5; right: p = 0.853, d = 0.23) but also matched our hypothesis better than the LSTM (left: p = 0.001, d = 0.79; right: p = 0.902, d = 0.3).

To specify the directionality of information flow, we averaged weights of top-down modulations (i.e., predictor weights of timescales longer than respective target timescales) and bottom-up modulations. We found no difference between the modulatory strength of these top-down and bottom-up connections (LSTM left: p = 0.184, d = 0.23; right: p = 0.839, d = 0.4; HM-LSTM left: p = 0.408, d = 0.4; right: p = 0.367, d = 0.16).

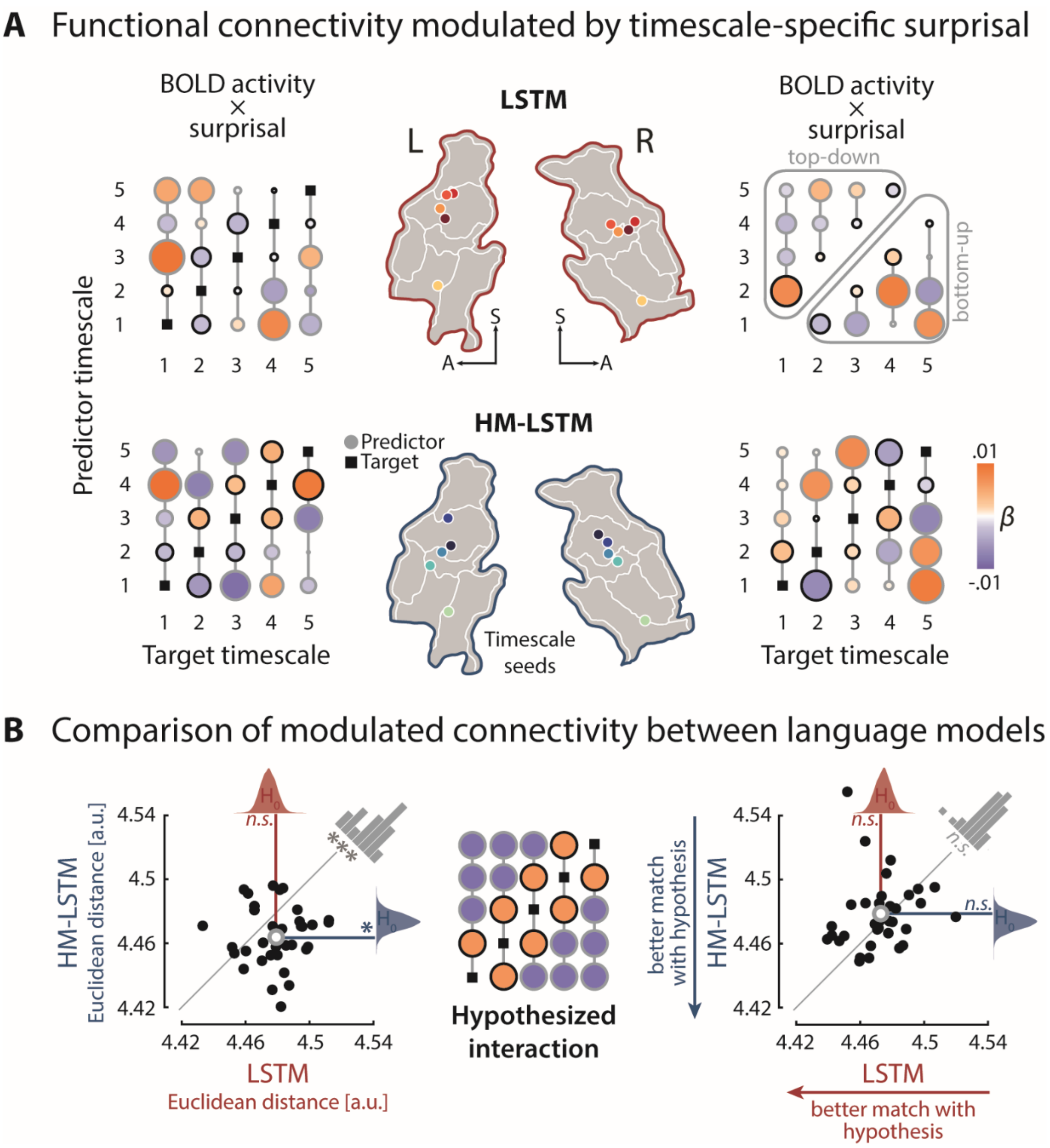


Figure 3.14. Surprisal-dependent modulation of effective connectivity. (A) A sphere of 6 mm was centred on median peak locations of timescales as defined in the encoding analysis (coloured circles on temporo-parietal maps) and BOLD responses were averaged within these timescale seeds. BOLD time series at one (target) timescale were regressed onto psychophysiological interactions of all other (predictor) timescales (i.e., pointwise product of timescale-specific BOLD and surprisal time series). For each target seed, we added a column vector of timescale-specific predictor weights to a 5-by-5 matrix with an empty main diagonal. Matrices were created separately for each language model (top: LSTM, bottom: HM-LSTM) and hemisphere. The upper triangle of a matrix indicates top-down, the lower triangle bottom-up information flow; black outlines of circles indicate timescale pairing for which we expected an increase of connectivity, a decrease of connectivity was expected for pairings with grey outlines (see hypothesized interaction pattern in 6B). **(B)** A hypothesized matrix of psychophysiological interactions was created, with positive weights on diagonals below and above the main diagonal, indicating increased connectivity between neighbouring timescales when surprisal is high. The Euclidean distance between observed and hypothesized matrices was compared to null distributions of distances drawn from target timescales shifted in time (coloured density plots), separately for the left and right hemisphere; black dots indicate distances of single participants; grey circles indicate mean distances. * $p < 0.05$; *** $p < 0.001$; n.s.: not significant.

3.4.4 Discussion

How are the complex temporal dependencies underlying natural speech processed in the brain to inform predictions of upcoming speech? In the present study, we emulated these prediction processes in two language models (artificial neural networks; LSTM vs. HM-LSTM), which critically differed in how often semantic context representations are updated at multiple, hierarchically organized timescales.

Surprisal as derived from both language models modulated reading times in the behavioural reading task to a similar degree, while hemodynamic brain responses to surprisal during the listening task differed between models: In early auditory cortex and supramarginal gyrus, the continuously updating LSTM predicted activity better than the sparsely updating HM-LSTM. In general, surprisal at the shortest timescale was decoded more precisely from temporo-parietal regions when derived from the continuously updating LSTM than from the HM-LSTM.

In contrast, and in line with our initial hypothesis, temporo-parietal regions hierarchically encoded the (sparsely updated) event-based surprisal provided by the layers of the HM-LSTM, with longer timescales represented in inferior parietal regions. Moreover, higher timescale-specific surprisal based on the HM-LSTM increased connectivity from receptive windows of a given timescale to their immediately neighbouring (shorter or longer) timescales.

Together, these results provide evidence for the neurobiological parsimony of an event-based processing hierarchy. In the present data, this was expressed in the simultaneous neural representation of surprisal at multiple timescales, and in surprisal dynamically gating the connectivity between these timescale-specific receptive windows.

The event-based organization of context as a foundation for language prediction

The spatial organization of timescale-specific receptive windows observed in the present study converges with previous results, where bilateral primary auditory cortex coded for relatively shorter timescales (e.g., words) and inferior parietal cortex coded for longer timescales (e.g., paragraphs; Lerner et al., 2011). Critically, this spatial overlap was found despite targeting different aspects of speech processing.

In our study, neural responses were expressed as a function of timescale-specific word surprisal, a proxy tapping into prediction processes. In other studies, receptive windows were based on the (in-)consistency of neural activity across participants in response to speech input at varying timescales (Hasson et al., 2008), which is typically linked to working memory formation. This implies that the same neural system most likely fulfils distinct functions: Temporal receptive windows have been suggested to store timescale-specific context in working memory, and, in parallel, exploit this context to process information in the present (Hasson et al., 2015). In line with this theoretical account, our results suggest that timescale-specific memory representations serve as the basis for the generative models shaping predictions of upcoming speech.

The here observed temporo-parietal gradient of surprisal at sparsely updated representations of context is specifically well in line with accounts of neural event segmentation (Zacks et al., 2007; Radvansky, 2012), and with the notion of hierarchical multiscale network architectures more generally (here, HM-LSTM). Taking a sentence from our listening task as an example, "The wild wine was called ink grapes" is embedded in a brief event where the narrator describes how the bluish black of the grapes in a backyard reminded her of the color of the night. At the same time, the sentence is embedded in a larger event of the author wandering around the single rooms and the garden of her parents' house, and an even larger event of the author reliving the memory of walking the streets in her Romanian hometown.

The HM-LSTM architecture resembles neural event segmentation in two decisive points. First, the boundary detector allows revealing the event structure of context, similar to an increase of neural activity indexing prediction errors at event boundaries (Zacks et al., 2001; Ditman et al., 2008; Whitney et al., 2009). Second, the sparse updates to higher processing stages at event boundaries allow retaining multiple, stable context representations in memory, similar to temporo-parieto-occipital receptive windows reflecting the hierarchical event structure during movie watching (Baldassano et al., 2017). We directly tie in with this result by showing that such hierarchical, event-based context enables neural prediction processes.

What are the computational and mechanistic implications of this contextual architecture for the prediction of speech? Somewhat paradoxically, event models have been referred to as "an added burden for an organism" (Richmond & Zacks, 2017). This argument is certainly plausible with regard to the size of the parameter space, which increases in an artificial or, likewise, biological neural network by introducing an additional boundary detector. At the expense of model parsimony, however, such an event-based network allows for less updates in comparison to continuously updating networks like the LSTM, where each new input to the model elicits computationally complex updates to all timescales.

The trade-off between computational costs of boundary detector versus update frequency is nicely illustrated by the fact that the sparsely updating HM-LSTM is considerably faster in making predictions than the continuously updating LSTM (Chung et al., 2016). Thus, from a functional perspective, keeping layered representations of multiple events in memory allows to efficiently draw on diverse information to make predictions on upcoming speech.

Context-dependent surprisal as a gating mechanism for predictions and prediction errors

The hallmark of prediction processes in our data is the increase in reading times and neural activity observed in response to more surprising input (Wacongne et al., 2011; Lieder et al., 2013; Kumar et al., 2017). There are different computational ways in which this "expectation suppression" (Todorovic & de Lange, 2012) can be realized, namely integration difficulty, neural sharpening, and predictive coding.

One take on expectation suppression is that surprising sensory input is more difficult to integrate into already existing representations of context because it conveys a relatively larger amount of new information (Brown & Hagoort, 1993; Hagoort et al., 2009). As the architecture of the sparsely updating HM-LSTM dictates that new information is integrated into timescale-specific representations only at event boundaries, integration difficulty should arise primarily at an event-by-event basis. However, surprisal indeed varies on a word-by-word basis. This conceptual mismatch renders it highly unlikely that integration difficulty accounts for our effects of event-based surprisal. The other two accounts both assume that expectation suppression is indicative of prediction processes but differ in how these processes are thought to be implemented in the brain.

Sharpening accounts argue that *un*expected components of sensory input are suppressed via feedback predictions (Lee & Mumford, 2003; Kok et al., 2012), resulting in an overall decrease of neural activity in response to expected input. Under the predictive coding account (Rao & Ballard, 1999; Friston, 2005), the brain filters out (or "dampens") expected components of sensory input, so remaining neural activity (mostly related to prediction errors) is overall smaller for more expected input. The similarity between hypothesized response patterns makes it notoriously hard to disentangle those accounts (Kok et al., 2012; Bell et al., 2016; Blank & Davis, 2016).

Notably, however, a distinguishing feature of predictive coding is the specificity of feedforward prediction error signals, which can be captured by modelling effective connectivity between receptive windows of timescales. In agreement with the hierarchical information flow laid out in predictive coding (Friston, 2005), surprisal in our study modulated connectivity via bidirectional links between neighbouring receptive windows of longer and shorter event-based timescales in the left hemisphere (Figure 3.14).

Surprisal in the event-based artificial neural network was modelled as the amount of information an input word conveys that cannot be explained away by the context (or generative model) represented at a specific timescale. Therefore, the increase of feedforward connectivity in response to higher surprisal precisely aligns with the concept of prediction errors in predictive coding (Rao & Ballard, 1999).

In addition, the increase of feedback connectivity in response to higher surprisal accords with an electrocorticography study in macaques by Chao and colleagues (2018). The study showed that prediction errors evoked in tone sequences trigger feedback signals from prefrontal to anterior temporal and early auditory cortex in alpha and beta frequency bands. Extending these previous results, our findings suggest that surprisal initiates bottom-up prediction errors, indicative of imprecise predictions, and top-down updates to predictions at processing stages of shorter events to facilitate perception of new words.

As an interim conclusion, our findings have two important implications for frameworks of prediction and prediction error: First, we show that a multi-layered hierarchy of predictive coding (e.g., Kiebel et al., 2008) applies well to higher-order semantic language processing. Second, predictive coding remains a viable account of neural processing, also when put to test using complex temporal dependencies underlying real-life stimuli.

Implications for a larger network perspective on the event-based prediction hierarchy

Dual stream models of language propose that speech processing is organized along a ventral and a dorsal stream (Hickok & Poeppel, 2004, 2007). In the present study, we found a hierarchy of speech prediction along the dorsal stream, which emanated from early auditory cortex and extended well into parietal cortex (Figure 3.12B).

This result may seem at odds with other studies showing an additional mirror-symmetric ventral gradient, in which more complex speech features are represented in more anterior temporal regions (DeWitt & Rauschecker, 2012). The ventral stream has been proposed to chunk speech features into increasingly abstract concepts irrespective of their temporal presentation order (Bornkessel-Schlesewsky & Schlesewsky, 2013). In contrast, we here modelled context representations by respecting the temporal order of words, that is, the HM-LSTM integrates incoming words into an event until words become too dissimilar to previous words and a new event is created. Hence, the ventral stream may contribute to hierarchical speech prediction by exploiting another, more nested facet of context.

The inferior frontal gyrus (IFG), alongside premotor cortex, is deemed the apex of the dorsal stream (Hickok & Poeppel, 2007), yet we here considered only the role of temporo-parietal cortex in speech prediction. Previous studies showed that activity in IFG relies on longer timescales of speech being intact (Wilson et al., 2008; Lerner et al., 2011), that connectivity between IFG and superior temporal gyrus is driven by expectations (Phillips et al., 2016; Garrido et al., 2018), and that right IFG is sensitive to the violation of non-local regularities (Meyniel & Dehaene, 2017; Cheung et al., 2018). While this suggests an interplay between frontal and temporo-parietal regions in hierarchical speech prediction, the precise anticipatory mechanisms IFG exerts cognitive control over are just as unclear as how top-down cognitive control and bottom-up sensory input are balanced along the hierarchy.

Beyond short-term semantic context, also long-term knowledge facilitates speech prediction. In theory, both memory systems can be couched into the larger framework of the dual reference frame system (Bottini & Doeller, 2020), where flexible sensory knowledge in parietal cortex interacts with stable conceptual knowledge in hippocampus. Consistent with the key characteristics of the speech prediction hierarchy, hippocampus codes for boundaries in the environment (Spiers et al., 2015; Brunec et al., 2018), hierarchically organizes memories (Alexander & Nitz, 2017) and engages in predictive coding (Johnson & Redish, 2007; Stachenfeld et al., 2017). As parietal cortex has been shown to interface with hippocampus at event boundaries of longer timescales during movie watching (Baldassano et al., 2017), we speculate that the hierarchy of speech prediction might extend from receptive windows in parietal cortex to hippocampus.

Importantly, the event-based prediction hierarchy relies on a set of neural computations—i.e., event segmentation, temporal receptive windows, predictive coding—available beyond the domain of language. Our results thereby encourage future studies to probe its generalizability to other species, sensory modalities, and cognitive functions.

Alternative mechanisms of predictive processing in lieu of event-based timescales

Although we only found a processing hierarchy for surprisal of sparsely updated timescales, temporo-parietal regions were nevertheless sensitive to continuously updated timescales. In particular, decoding accuracies suggested a predominance of the LSTM over the HM-LSTM at the shortest timescale and encoding accuracies suggested a predominance in medial temporal and anterior parietal regions (Figure 3.12C and 3.13B). This finding agrees with previous studies showing that participants track changes to situational dimensions of narratives both “globally” at the end of an event and “incrementally” within events (Kurby & Zacks, 2012) and that computational models with continuous updates to all hierarchical levels can explain the construction of context representations in temporo-parietal regions (Chien & Honey, 2020). Could continuously updated context representations, after all, play an integral role in successful speech prediction?

One potential explanation for the negligible neural effects at longer timescales is that the continuously updating language model relies primarily on shorter timescales in predicting the next word. This is supported by the considerably worse model performance observed for longer LSTM timescales (i.e., higher average word surprisal) compared to both shorter LSTM timescales and, despite comparable overall model performance, all HM-LSTM timescales (Supplementary Figure 3.8). Interestingly, the accuracy in predicting reading speed from word surprisal was the same between LSTM and HM-LSTM (Figure 3.10), suggesting that continuous updates make for an efficient mechanism to generate equally accurate predictions while relying on less timescales. The strength of such continuously updating models is that context representations are more integrated with what is currently relevant for prediction.

A unifying account might be that continuous and sparse updating mechanisms form one instead of two distinct processing streams. For example, Sainburg and colleagues (2019) showed that short-range dependencies of acoustic speech features follow sequential Markovian processes, whereas long-range dependencies follow hierarchical processes. This poses the question whether such interactions of different sequencing mechanisms also better match the semantic structure of speech. Future studies could test this by setting up hybrid language models with continuous updates on shorter and sparse updates on longer timescales.

Conclusion

The present study bridges the gap between the hierarchical, temporally structured organization of context in language comprehension on the one hand and the more general principles of hierarchical predictive processing in cerebral cortex on the other hand.

Combining continuously narrated speech, artificial neural networks, and functional MRI building on these networks' output allowed us, first, to sample the natural dynamic range of word-to-word changes in predictiveness over a multi-level hierarchy. Second, we were able to systematically compare the neural effects of different contextual updating mechanisms.

Our data demonstrate that the prediction processes in language comprehension build on an event-based organization of semantic context along the temporo-parietal pathway. Not least, we posit that such an event-based organization provides a blueprint for a semantically rich, yet computationally efficient network architecture of anticipatory processing in complex naturalistic environments.

3.4.5 Methods

Participants

Thirty-seven healthy, young students took part in the fMRI listening study. The final sample included *N* = 34 participants (18–32 years; *M* = 24.65; 18 female), as data from one participant was excluded from all analyses due to strong head movement throughout the recording (mean framewise displacement > 2 *SD* above group average; Power et al., 2012) and two experimental sessions were aborted because participants reported to not understand speech against noise. Another 26 students (19–32 years; *M* = 23.54; 17 female) took part in the behavioural self-paced reading study.

All participants were right-handed German native speakers who reported no neurological, psychiatric or hearing disorders. Participants gave written informed consent and received an expense allowance of €10 per hour of testing. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee of the University of Lübeck.

Stimulus materials

As a speech stimulus in the fMRI listening task, we used the first 64 minutes of an audio recording featuring Herta Müller, a Nobel laureate in Literature, reminiscing about her childhood as part of the German-speaking minority in the Romanian Banat ("Die Nacht ist aus Tinte gemacht", 2009).

To emulate an acoustically challenging scenario in which listeners are likely to make use of the semantic predictability of speech (Rysop et al., 2021), this recording was energetically masked by a stream of concatenated five-second sound textures at an SNR of 0 dB. Sound textures were synthesized from the spectro-temporal modulation content of 192 natural sounds (i.e., human and animal vocalizations, music, tools, nature scenes; McDermott & Simoncelli, 2011), so that the noise stream did not provide any semantic content potentially interfering with the prediction of upcoming speech. The order in which sound textures were arranged was randomized across participants. For more details on how sound textures in the present experiment were generated and how they were processed in auditory cortex, see Erb and colleagues (2020).

The monaural speech and noise streams were sampled to 44.1 kHz and custom filters specific to the left and right channel of the earphones used in the fMRI experiment were applied for frequency response equalization. Finally, speech-in-noise stimuli were divided into eight excerpts à 8 minutes, which served as independent runs in the experiment.

A trained human external agent literally transcribed the speech stream. The text transcript comprised 9,446 words, which were used as stimuli in the self-paced reading task and as input to our language models. To automatically determine onset and offset times of all spoken words and phonemes, we used the web service of the Bavarian Archive for Speech Signals (BAS; Kisler et al., 2017): First, the text transcript was transformed to a canonical phonetic transcript encoded in SAM-PA by the G2P module. Second, the most likely pronunciation for the phonetic transcript was determined within a Markov model and aligned to the speech recording by the MAUS module. Fourteen part-of-speech tags were assigned to the words in the text transcript using the pre-trained German language model de_core_news_sm (2.2.5) from spaCy (https://spacy.io). Based on these tags, words were classified as content or function words. Word frequencies were derived from the subtitle-based SUBTLEX-DE corpus (Brysbaert et al., 2011) and transformed to standardized Zipf values (van Heuven et al., 2014) operating on a logarithmic scale from about 1 (word with a frequency of 1 per 100 million words) to 7 (1 per 1,000 words). The Zipf value of a word not observed in the corpus was 1.59 (i.e., smallest possible value).

Experimental procedures

Behavioural self-paced reading task. While the transcribed story was presented word-by-word on a noncumulative display, participants had the task to read each word once at a comfortable pace and quickly press a button to reveal the next word as soon as they had finished reading. A timeout of 6 seconds was implemented. The time interval between word appearance and button press was logged as the reading time. After each run, participants answered three four-option multiple-choice questions on the plot of the story (performance: *Ra* = 58.33–100 % correct, *M* = 79.17, *SD* = 10.87) and took a self-paced break. In total, each participant completed four out of eight runs, which were randomly selected and presented in chronological order. Throughout the reading task, we recorded movement and pupil dilation of participants' right eye at a sampling rate of 250 Hz in one continuous shot with an eye tracker (EyeLink 1000, SR Research).

The experiment was controlled via the Psychophysics Toolbox (Brainard, 1997) in MATLAB (R2017b, MathWorks). All words were presented 20 % off from the left edge of the screen in white

Arial font on a grey background with a visual angle of approximately 18°. Participants used a response pad (URP48, The Black Box ToolKit) to navigate the experiment with their right index finger. The experimental session took approximately 40 minutes.

Functional MRI listening task. We instructed participants to carefully listen to the story while ignoring the competing stream of sound textures as well as the MRI scanner noise in the background. Each of the eight runs was initialized by 10 baseline MRI volumes after which a white fixation cross appeared in the middle of a grey screen and playback of the 8-minute audio recording started. MRI recording stopped with the end of playback and participants successively answered the same questions used in the self-paced reading task via a response pad with four buttons (HHSC-2x4-C, Current Designs). On average, participants answered 65.5 % of the questions correctly (*Ra* = 38–100 %, *SD* = 15.9 %). There was a 20-second break between consecutive runs.

The experiment was run in MATLAB (R2016b) using the Psychophysics Toolbox. Stimuli were presented at a subjectively comfortable sound pressure level via insert earphones (S14, SENSIMETRICS) covered with circumaural air cushions. The experimenters monitored whether participants kept their eyes open throughout the experiment via an eye tracker.

MRI data acquisition. MRI data were collected on a 3 Tesla Siemens MAGNETOM Skyra scanner using a 64-channel head coil. During the listening task, continuous whole-brain fMRI data were acquired in eight separate runs using an echo-planar imaging (EPI) sequence (repetition time (TR) = 947 ms, echo time (TE) = 28 ms, flip angle = 60°, voxel size = 2.5 × 2.5 × 2.5 mm, slice thickness = 2.5 mm, matrix size = 80 × 80, field of view = 200 × 200 mm, simultaneous multi-slice factor = 4). Fifty-two axial slices were scanned in interleaved order. For each run, 519 volumes were recorded.

Before each second run, field maps were acquired with a gradient echo (GRE) sequence (TR = 610 ms, TE_1 = 4.92 ms, TE_2 = 7.38 ms, flip angle = 60°, voxel size = 2.5 × 2.5 × 2.75 mm, matrix size = 80 × 80, axial slice number = 62, slice thickness = 2.5 mm, slice gap = 10 %).

In the end of an experimental session, anatomical images were acquired using a T1-weighted (T1w) MP-RAGE sequence (TR = 2,400 ms, TE = 3.16 ms, flip angle = 8°, voxel size = 1 × 1 × 1 mm, matrix size = 256 × 256, sagittal slice number = 176) and a T2-weighted (T2w) SPACE sequence (TR = 3,200 ms, TE = 449 ms, flip angle = 120°, voxel size = 1 × 1 × 1 mm, matrix size = 256 × 256, sagittal slice number = 176).

Deriving the predictiveness of timescales by "lesioning" the language models. We used each trained language model to determine the predictiveness of semantic context in the story presented to participants in the behavioural and fMRI experiment. For a detailed description of the architecture of language models and the trainings procedure, see section 3.3.3. First, predictiveness was read out from "full" models: We iteratively selected each word in the story as a target word and fed all 500 context words preceding the target word to our language models. Note that the context for target words in the very beginning of the story comprised less than 500 words. The "predicted" probability of each word in the vocabulary was extracted from distribution d_p in the output module.

Second, predictiveness was read out from "lesioned" models, where we allowed information to freely flow through networks, yet only considered semantic context represented at single layers to generate the "predicted" probability distribution. These timescale-resolved probabilities were created by setting weight matrix W_r^l of pre-trained models to zero for all layers of no interest, so that the hidden state of only one layer is passed to the softmax function and all other layers have no bearing on the final prediction. We iteratively set all but one layer to zero with each layer being the only one influencing predictions once, resulting in five lesioned outputs for each language model.

We derived three measures of predictiveness from probability distributions. Our primary measure was the degree of surprisal associated with the occurrence of a word given its context. Word surprisal is the negative logarithm of the probability assigned to the actual next word in a story.

Secondary measures of predictiveness were used to explore the specificity of the processing hierarchy to only some aspects of prediction processes. Word entropy reflects the amount of uncertainty across the whole probability distribution, which is the negative sum of probabilities multiplied by their natural logarithm. When high probabilities are assigned to only one or few words in the vocabulary, entropy is low. On the other hand, entropy is high when semantic context is not informative enough to narrow predictions down to a limited set of words, resulting in similar probabilities for all candidate words. As all information necessary to determine the entropy of a word is already available to participants before word presentation, entropy of word w_p was ascribed to the previous word w_{p-1}. Whereas word surprisal quantifies the availability of information on the actual next word, word entropy quantifies the overall difficulty of making any definite prediction. Another secondary measure of predictiveness was the relatedness of the predicted next word to the actual next word. This word similarity is expressed as the correlation of respective word embeddings. A high positive Product-moment correlation indicates that the prediction is semantically close to the target word, even though the model prediction might have been incorrect.

All three measures were calculated for each word in the story, separately for full models and five lesioned models. This yielded an 18-dimensional feature space of predictiveness for the LSTM as well as the HM-LSTM model, which was linked to BOLD activity and reading times in our analysis.

Additionally, we created a metric to dissociate neural effects of predictiveness from more low-level effects of semantic dissimilarity between target words and their preceding context. To this end, we correlated the embedding of each function word in the story with the average embedding of a context window, and subtracted resulting Product-moment correlation coefficients from 1 (Broderick et al., 2018). This measure of contextual dissimilarity was calculated at five timescales corresponding to a logarithmic increase in context length (i.e., 2, 4, 8, 16, and 32 words).

To determine temporal integration windows of layers, we scrambled input to language models at nine levels of granularity corresponding to a binary logarithmic increase in the length of intact context (i.e., context windows of 1–256 words). For each layer, we fit linear functions to word surprisal across context windows and extracted slope parameters indicating how much a layer benefits from longer context being available when predicting the next word. On the second level, we fit linear functions to these layer-specific integration windows to determine the context benefit of higher layers over shorter layers. Resulting model-specific slopes were compared to a null distribution of slopes computed by shuffling integration windows across layers (n = 10,000). Additionally, slopes were compared between language models by means of a Monte Carlo approximated permutation test (n = 10,000) on the difference of means.

Convolving features with the hemodynamic response function. We used three classes of features to model brain responses: 18 features of the predictiveness of timescales (per language model), 3 linguistic features, and 9 acoustic features. While we were primarily interested in modelling effects of predictiveness, linguistic and acoustic features were used as nuisance regressors potentially covarying with predictiveness. Linguistic features included information on when words were presented (coded as 1), whether they were content or function words (coded as 1 and -1), and which frequency they had. In a study by Erb and colleagues (2020), we decomposed the speech-in-noise stimuli into a 288-dimensional acoustic space of spectral, temporal and spectrotemporal modulations, which was derived from a filter bank modelling auditory processing (Chi et al., 2005). Here, we reduced the number of acoustic features to the first 9 principal components, which explained more than 80 % of variance in the original acoustic space. All features were z-scored per run.

A set of 500 scrambled features of predictiveness was generated, which was used to estimate null distributions of predictive processing. We applied the fast Fourier transform to single features, randomly shifted the phase of frequency components, and inverted the transform to project the data back into the time domain. This preserved power spectra of features but disrupted the temporal alignment of frequencies. See Supplementary Text 1 for a detailed description of convolving features with the hemodynamic response function (HRF).

Data analysis

See Supplementary Text 2 for a detailed description of structural and functional MRI data preprocessing.

Selection of regions of interest. We hypothesized that the speech prediction hierarchy is represented as a gradient along a temporo-parietal pathway. This rather coarse region of interest was further refined to only include regions implicated in speech processing. To this end, we used intersubject correlation (Nastase et al., 2019) as a measure of neural activity consistently evoked across participants listening to speech in noise. As we were primarily interested in shared responses to the speech stream, this approach allowed us to leverage the inconsistency of the noise stream across participants. The presentation of sound textures in different order likely evoked more heterogeneous neural responses, leading to a diminished shared representation of the noise stream. Therefore, we inferred that the shared neural responses we observed were largely driven by the speech stream, which was the same for all participants.

At the first level, hyperaligned functional time series of each participant (see Supplementary Text 2) were concatenated across experimental runs and correlated with every other participant on a vertex-by-vertex basis, resulting in pairwise maps of intersubject Product-moment correlations. A group map was created by calculating the median correlation coefficient across pairs of participants for each vertex. At the second level, we applied a bootstrap hypothesis test with 10,000 iterations. To create the null distribution, we iteratively resampled participants with replacement and derived median group maps from their pairwise correlation maps. When the same participant was sampled more than once in a bootstrap iteration, the pairwise correlation map of that participant with herself was not included in the computation of the group map. The actual median intersubject correlation was ranked against the normalized null distribution to obtain a p-value for each vertex. Intersubject correlations were computed with the Python package BrainIAK (Kumar et al., 2020).

Finally, we used a multi-modal parcellation (Glasser et al., 2016) to select those lateral temporal and parietal parcels of which at least 80 % of the vertices had a significant intersubject correlation ($p < 0.01$, adjusted for false discovery rate (FDR); Benjamini & Hochberg, 1995) in one hemisphere. The following parcels were included in the region of interest (ROI): early auditory cortex (EAC), auditory association cortex (AAC), lateral temporal cortex (LTC), temporo-parietal-occipital junction (TPOJ), inferior parietal cortex (IPC), and superior parietal cortex (SPC). As the temporal MT+ complex is thought to be mainly involved in visual processing, this region was not considered an appropriate candidate parcel. All further analyses including MRI data were limited to the temporo-parietal ROI, which was organized along the anterior-posterior (left: 124 mm, right: 167 mm) and inferior-superior axis (left: 234 mm, right: 212 mm).

Functional data analysis. The starting point of our analyses was the question whether the timescales of speech prediction organize along a temporo-parietal processing hierarchy. In a forward model, we encoded the predictiveness of timescales into univariate neural activity and fit a gradient along the peak locations sensitive to specific timescales of surprisal. Next, we compared the explanatory power of both language models in a backward model, which decoded surprisal at different timescales from multivariate patterns of neural activity in temporo-parietal parcels. Finally, we modelled functional connectivity between peak locations to test whether the timescales of surprisal gate the information flow along the gradient.

Encoding model. The encoding approach (similar to e.g., Santoro et al., 2017) allowed us to quantify for each temporo-parietal vertex, which features of predictiveness it preferentially represents. Two separate encoding models were estimated for each vertex in the ROI of single participants, one for each language model. Besides the features of predictiveness specific to language models, both models included the same linguistic and acoustic features as nuisance regressors. We modelled neural activity as a function of 30 HRF-convolved features characterizing speech and noise stimuli by:

$$a = Sw + \epsilon,$$

where $a^{samples \times 1}$ is the activity vector (or BOLD time course) corresponding to a vertex, $S^{samples \times features}$ is the stimulus matrix of features, $w^{features \times 1}$ is a vector of estimated model weights, and $\epsilon^{samples \times 1}$ is a vector of random noise.

All models were estimated using ridge regression with four-fold cross validation. We paired odd-numbered functional runs with their subsequent even-numbered run, resulting in four data splits per participant. Each of the four data splits was selected as a testing set once; all other data splits were used as a training set. Within each fold, generalized cross-validation (Golub et al., 1979) was carried out on the training set to find an optimal estimate of regularization parameter λ from the data, searching 100 values evenly spaced on a logarithmic scale from 10^{-5} to 10^{8}. Weights of predictiveness were extracted from the model fit with the optimal regularization parameter and averaged across cross-validation folds to obtain stable weights.

To evaluate the performance of encoding models and their ability to generalize to new data, we applied the weights estimated on the training set to the features of the held-out testing set in each cross-validation fold. The predicted BOLD time series was correlated with the actual BOLD time series. The resulting Product-moment correlation coefficient is the encoding accuracy, which was averaged across cross-validation folds and Fisher *z*-transformed.

Additionally, we created null distributions of weights and encoding accuracies by estimating forward models on scrambled features of predictiveness (similar to e.g., Musall et al., 2019). We set up 500 separate models, which included scrambled features of predictiveness but intact linguistic and acoustic features. Models were estimated largely following the cross-validation scheme outlined for observed data. However, we re-used optimal regularization parameters from non-scrambled models of corresponding folds. All ridge regression models were implemented using the RidgeCV function in the Python package scikit-learn (Pedregosa et al., 2011).

Peak selection. For both language models, we derived five temporo-parietal maps in the left and right hemisphere of single participants: one weight map for each timescale of word surprisal. Maps represented the sensitivity of brain regions to timescale surprisal; positive weights indicate increasing BOLD activity to more surprising words.

To illustrate the location and extent of brain regions modulated by timescale surprisal, we performed an analysis similar to cluster-based permutation tests in Fieldtrip (Maris & Oostenveld, 2007). For each timescale, vertex-wise weights observed across participants were tested against zero by means of a one-sample *t*-test. We combined a vertex into a cluster with its adjacent vertices if it was significant at an alpha level of 0.05 and had at least two significant neighbours. We clustered vertices with negative *t*-values separately from vertices with positive *t*-values. The summed *t*-value of an observed cluster served as the cluster-level statistic and was compared with a Monte Carlo approximated null distribution of summed *t*-values. This null distribution was created by performing clustering on scrambled partitions of timescale-specific weight maps and selecting the largest summed *t*-value for each partition. An observed cluster was considered significant if its summed *t*-value was exceeded by no more than 2.5 % of summed *t*-values from scrambled partitions.

Beyond this rather coarse mapping of temporo-parietal brain regions onto the timescales of surprisal, our main analysis focused on how timescale-specific peak locations distribute along the inferior-superior axis only. Of note, we hypothesized that a hierarchy of speech prediction evolves from temporal to parietal areas, which corresponds to the inferior-superior axis of our ROI. A window with a height of 2 mm was shifted along the inferior-superior axis of the temporo-parietal ROI in steps of 1 mm. All weights of a timescale falling into the window were averaged, thereby collapsing across the anterior-posterior axis. The resulting one-dimensional weight profile of a timescale spanned inferior to superior locations and was smoothed using robust linear regression over a window of 70 mm. For each unilateral weight profile of single participants, local maxima (i.e., a sample larger than its two neighbouring samples) were determined.

We applied two different approaches to select one peak location for each timescale from these local maxima. In the naïve approach of peak selection, the local maximum with the highest positive value was defined as a peak. As this approach makes it hard to find a consistent order of timescales when surprisal is not just processed along the dorsal but also the ventral processing stream, we also applied a pre-informed peak selection approach explicitly targeting the dorsal stream. Here, the peak of the first timescale had to be in the inferior half of the axis (i.e., temporal regions) and peaks of longer timescales had to be superior to the peak of the first timescale. Whenever no timescale peak could be defined, the largest positive value was selected. Both peak selection approaches yielded five timescale-specific coordinates on the inferior-superior axis for each participant, hemisphere and language model.

Additionally, we applied naïve peak selection to weight maps, which were rotated by −45° before collapsing across the first dimension. In the left hemisphere, an increase on that new axis indicated a shift to more superior and posterior regions, thereby simultaneously modelling effects on the inferior-superior and anterior-posterior axis. However, original right-hemispheric maps already had a strong rotation off the inferior-superior axis, thus rotating these maps rather brought them into alignment with the inferior-superior axis of non-rotated left-hemispheric maps.

Gradient fitting. We fit linear functions to coordinates of single participants across the timescales of surprisal. Models included an intercept term and the slope parameter was extracted from each fit. A positive slope indicates a gradient of timescale surprisal, where shorter timescales are represented in more inferior (anterior) temporo-parietal regions than longer timescale, which are represented in more superior (posterior) regions. We tested grand-average slope parameters against a null distribution of slopes with 10,000 partitions, which was created by randomly shuffling the coordinates of single participants across the timescales of surprisal and recalculating their slopes. As the first timescale was pre-set to have the most inferior coordinate in the pre-informed peak selection approach, this specific coordinate was not shuffled when calculating the null distribution for this approach. To compare slope parameters between language models, we performed a Monte Carlo approximated permutation test (n = 10,000), using the difference of means as a test statistic. As secondary analyses, gradients of predictive processing were also calculated for the timescales of word entropy and similarity. In a control analysis, a gradient was fit to timescale peaks following the same procedure described above but replacing features of predictiveness by contextual dissimilarity when estimating forward models.

To round off the encoding analysis, we compared temporo-parietal encoding accuracies between both language models. As we were interested in effects specific to the predictiveness of speech, encoding accuracies were z-scored to the null distribution of accuracies from scrambled features of predictiveness. A cluster-based permutation paired-sample t-test was calculated (n = 1,000, vertex-specific alpha level: 0.05, cluster-specific alpha level: 0.05). In comparison to the cluster test described above for the weight maps, we here created a null distribution of summed t-values by contrasting accuracies of language models whose labels had been randomly shuffled in single participants.

Decoding model. In our decoding approach (similar to e.g., Santoro et al., 2017), we quantified how much information multiple vertices jointly contain about a feature of predictiveness. For each language model, five separate backward models were estimated in each of six temporo-parietal parcels of single participants, one for each timescale of word surprisal. We modelled timescale-specific word surprisal as a function of neural activity in all vertices forming a parcel by:

$$s = Aw + \epsilon,$$

where $s^{samples \times 1}$ is the stimulus vector of a feature, $A^{samples \times vertices}$ is the activity matrix of BOLD time courses corresponding to the vertices of a parcels, $w^{vertices \times 1}$ is a vector of model weights, $\epsilon^{samples \times 1}$ is a vector of random noise.

The same cross-validation scheme as described for the encoding model was applied. However, instead of predicting BOLD activity, we here reconstructed surprisal at different timescales. By correlating the actual stimulus time series with the one predicted on the held-out testing set, we obtained the decoding accuracy of a parcel. Decoding accuracies were z-scored to the null distribution of accuracies determined for scrambled stimulus time series. We compared decoding accuracies between language models in each hemisphere, parcel and timescale by means of a Monte Carlo approximated permutation test (n = 10,000) on the difference of means. Resulting p-values were corrected for multiple comparisons using FDR correction.

Functional connectivity. To model the information flow between brain regions sensitive to the different timescales of word surprisal, we identified five unique seeds for both language models in each temporo-parietal hemisphere. On the inferior-superior axis, we re-used the grand-median coordinate of each timescale as localized in the pre-informed peak selection. The corresponding coordinate on the anterior-posterior axis was localized by shifting a moving average with a window centred on the inferior-superior coordinate along the anterior-posterior axis (width: 2 mm, height: 5 mm), and determining peak locations on smoothed weight profiles of single participants. Following, we placed a sphere with a radius of 5 mm on peak coordinates from both axes and averaged BOLD time courses of vertices falling within this sphere, yielding the timescale-specific neural activity of seeds.

We expected increased information flow between seeds of adjacent timescales when one timescale becomes uninformative for the prediction of upcoming speech. This modulatory influence of surprisal on connectivity was modelled along the lines of a psychophysiological interaction (PPI; Friston et al., 1997). In a standard PPI analysis, the neural time series of one brain region is regressed onto the pointwise product of an experimental stimulus and the neural time series of another brain region. Here, we extended this approach by creating timescale-specific interactions: BOLD time series of seeds were multiplied by their corresponding HRF-convolved surprisal time series but not any of the surprisal time series at another timescale.

Functional connectivity was calculated for both language models in each participant and hemisphere. We set up five regression models, with every seed being selected as a target once. The physiological (BOLD) time series of the target seed was mapped onto the physiological, psychological (timescale-specific surprisal), and psychophysiological time series from all other (predictor) seeds. Models were estimated within the same cross-validation scheme outlined for the encoding model. We extracted all four weights from psychophysiological interaction terms of each target seed and arranged weights in a 5-by-5 matrix, with target seeds on the main diagonal and predictor seeds off the diagonal. This matrix of observed psychophysiological interactions was compared to a matrix with hypothesized interaction weights: The diagonals below and above the main diagonal were set to 1 (indicating increased coupling when surprisal at a neighbouring timescale is high), all other items were set to -1. We calculated the Euclidean distance of single-participant matrices to this hypothesized matrix. The mean of observed Euclidean distances was compared to a null distribution of 10,000 mean Euclidean distances calculated on BOLD time series of target seeds randomly shifted in time by the number of samples in 1 to 7 functional runs. Euclidean distances were compared between language models in each hemisphere by means of a Monte Carlo approximated permutation test (n = 10,000) on the difference of means.

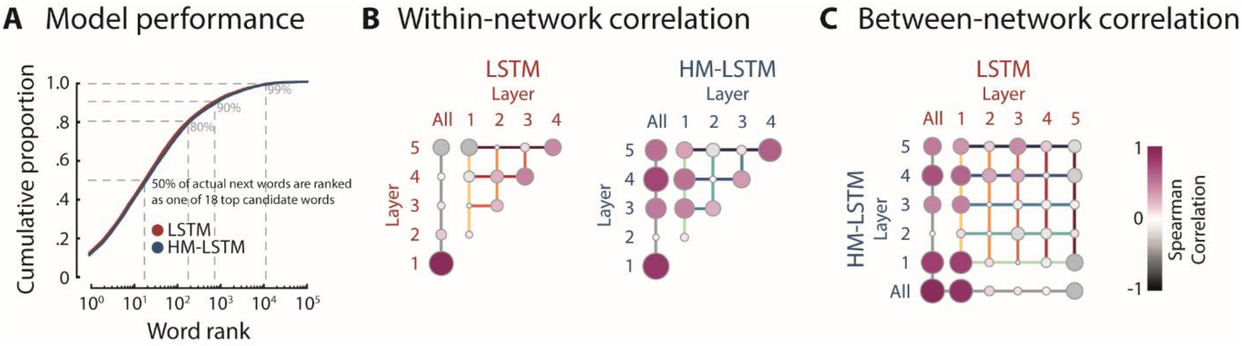
Behavioural data analysis. Reading times were used to test the behavioural relevance of the predictiveness determined by our language models. Trials with reading times shorter than 0.001 seconds or longer than 6 seconds were considered invalid and excluded. Further, we converted reading times to speed (number of words per 100 seconds) and excluded trials exceeding 3 standard deviations within a run and participant from all further analyses. On average, 1.31 % of trials (*SD* = 1.12, *Ra* = 0.32–6.15) were removed. Finally, reading speed was z-scored within runs.

For each participant, we predicted reading speed in a forward model, adopting the same cross-validated ridge regression scheme used for the analysis of fMRI data. Our feature space included the predictiveness of words as well as word frequency, word length (number of letters), content vs. function words and trial number as nuisance regressors. As this was a high-pace task, some features might have unfolded their effect on reading speed only over the course of a few words. Therefore, we added time-lagged versions of features to the model, that is, shifting features by -2 to 5 word positions. There were no lagged versions of the predictor coding for trial number added to the model.

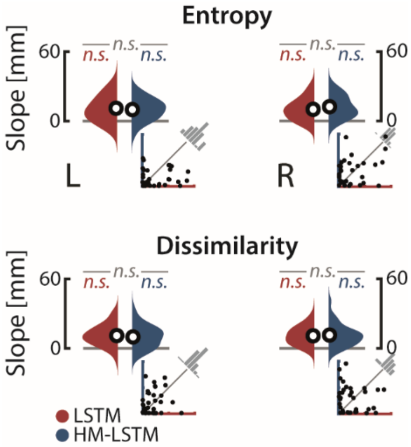
To investigate whether predictiveness had an effect on reading speed beyond the effect of nuisance regressors, we compared the predictive accuracy of forward models in single participants to a null distribution of accuracies from models with scrambled features of predictiveness. The performance of language models was compared by z-scoring observed encoding accuracies to the null distribution and running a Monte Carlo approximated permutation test (*n* = 10,000) on the difference of means. This analysis was also carried out for the timescales of contextual dissimilarity.

3.4.6 Supplementary materials

Supplementary figures



Supplementary Figure 3.8. Evaluating language models. (A) For each language model (LSTM: red, HM-LSTM: blue), we extracted the rank of the next word from the probability distribution of all candidate words. Language models ranked more than 50 % of words in the text as one of 18 top candidates words (out of more than 90,000 words). **(B)** Spearman correlations of word surprisal between single layers of language models, separately for LSTM (left) and HM-LSTM (right). In addition, correlations of single layers with full models are shown; color and size of circles scale to correlation coefficients. **(C)** Spearman correlations of word surprisal between language models.



Supplementary Figure 3.9. Encoding the timescales of entropy and dissimilarity. Along the dorsal stream, linear functions were fit to peak coordinates of entropy (top) and dissimilarity (bottom) across timescales. Resulting slope parameters were compared to null distributions drawn from scrambled coordinates and between language models (LSTM: red, HM-LSTM: blue), separately for the left (left column) and right hemisphere (right column). Black circles represent grand-average slope parameters; insets depict coefficients of determination for linear fits of single participants. *n.s.*: not significant.

Supplementary methods

Supplementary Text 1: Convolving features with the hemodynamic response function

For features of predictiveness and linguistics, we modelled (higher frequency, randomly spaced) information on the word level as hemodynamic responses sampled at the (lower frequency, equally spaced) TR of fMRI data. This was achieved by creating feature vectors of zeros corresponding to the length of a functional run with a sampling frequency of 1,000 Hz, which allowed word onsets and offsets to be represented with high temporal precision. For each word in a run, a boxcar function, which was scaled to the feature's value at that particular word, was placed on all samples falling in between word onset and offset. The resulting vector including feature values for all words in a run was convolved with SPM's canonical hemodynamic response function (HRF; Penny et al., 2006) and downsampled to the TR. Acoustic features, on the other hand, were already sampled to the TR and therefore directly convolved with the HRF.

Supplementary Text 2: Preprocessing structural and functional MRI

Structural MRI data preprocessing. MRI data were preprocessed with fMRIPrep 1.2.4 (Esteban et al., 2019), which is based on Nipype 1.1.6 (Gorgolewski et al., 2011) and employs Nilearn 0.5.0 (Abraham et al., 2014) in many internal operations. For each participant, the T1w image was corrected for intensity non-uniformity using N4BiasFieldCorrection (ANTs 2.1.0; Tustison et al., 2010) and then skull-stripped using the OASIS template in antsBrainExtraction.sh (ANTs 2.2.0). Individual brain surfaces were reconstructed from T1w and T2w reference images using recon-all (FreeSurfer 6.0.1; Dale et al., 1999). The T1w reference image was spatially normalized to the MNI152NLin2009cAsym template (Fonov et al., 2009) through nonlinear registration with antsRegistration (ANTs 2.2.0; Avants et al., 2008).

A brain mask was created by reconciling ANTs-derived and FreeSurfer-derived segmentations of the cortical grey matter according to a customized variation of the implementation in Mindboggle (Klein et al., 2017). Brain tissue segmentation of cerebrospinal fluid, white matter and grey matter was performed on the T1w reference image using FAST (FSL 5.0.9; Zhang et al., 2001).

Functional MRI data preprocessing. For each functional run, BOLD time series were motion corrected using mcflirt (FSL 5.0.9; Jenkinson et al., 2002) and slice time corrected using 3dTshift (AFNI 20160207; Cox, 1996). After unwarping BOLD images based on the susceptibility distortion estimated from field maps, the BOLD reference image was aligned to the native T1w reference image using boundary-based registration with six degrees of freedom (Greve & Fischl, 2009) as implemented in bbregister (FreeSurfer). BOLD images were resampled to standard space. To correct for head motion, non-aggressive Automatic Removal of Motion Artifacts using Independent Component Analysis (ICA-AROMA; Pruim et al., 2015) was performed on the resampled and smoothed (6 mm FWHM Gaussian kernel) BOLD images. On average, 50.54 % of the maximal 200 components per functional run (*Ra* = 32.93–71.78, *SD* = 9.21) were classified as motion-related artefacts. Additional confounding noise time series like the average signal within cerebrospinal fluid and white matter as well as framewise displacement were calculated in Nipype following the definitions by Power and colleagues (2014). A Discrete Cosine Transform (DCT) basis set of six functions with a cut-off at 0.008 Hz was generated for temporal high-pass filtering.

After running fMRIPrep, we regressed out high-pass filters as well as cerebrospinal fluid and white matter signals from the BOLD time series using 3dTproject (AFNI 19.2.24). To avoid reintroducing previously removed artefacts into the functional data (Lindquist et al., 2019), we projected the ICA-AROMA artefact components onto the additional nuisance covariates and used the residuals as predictors orthogonal to prior predictors. The denoised BOLD images were resampled to the fsaverage5 template in surface space by averaging across the cortical ribbon in 5 equally spaced steps at each vertex using trilinear interpolation. All resamplings can be performed with a single interpolation step: volumetric resamplings were performed using antsApplyTransforms (ANTs 2.1.0) with Lanczos interpolation; surface resamplings were performed using mri_vol2surf (FreeSurfer). In each functional run, the first 10 baseline volumes as well as the last volume were removed from time series and all further analysis were carried out on z-scored single-vertex BOLD time series.

Functional alignment to a common space. To account for small spatial variations in intersubject response tuning, functional time series were projected into a common space using searchlight hyperalignment across the whole cortex as described by Guntupalli and colleagues (2016). Here, we centred a sphere (or searchlight) with a radius of 20 mm on each vertex and determined the optimal rotation of response vectors (or functional time series) within each searchlight in three iterations using Procrustes transformation. An intermediate common space was initialized by rotating one participant’s response vectors to best match the responses of a randomly chosen reference participant. All other participants were successively brought into alignment, with the average of all previously rotated response vectors as a reference. In a second iteration, all original response vectors were aligned to the intermediate common space and the average of resulting rotated response vectors became the final common space. In the third iteration, hyperalignment parameters mapping single participant's original response vectors to the final common space were calculated. Parameters corresponding to vertices of overlapping searchlights were averaged. We ran hyperalignment on four independent data splits (i.e., pairing up every fourth of eight functional runs) and averaged transformation matrices across data splits to derive final parameters for each participant. Hyperalignment was performed in PyMVPA (2.6.6; Hanke et al., 2009).

4 General discussion

The present thesis set out to identify the neural mechanisms that enable humans to make use of contextual constraints in predicting speech, ultimately fostering comprehension under adverse listening conditions. While detailed discussions of experimental results can be found in the discussion sections of respective studies, the aim of the following chapter is to discuss three topics central to this thesis in more detail. I will start out by revisiting the suitability of artificial neural networks as models of language prediction. Next, I will review the implications of different predictability metrics for understanding the neural underpinnings of language prediction. Finally, I will put forward a unifying framework that characterizes the interactions of biological neural networks in language prediction.

4.1 Summary of experimental results

Study 1 investigated the interplay of domain-specific and domain-general biological neural networks in speech comprehension. In detail, fMRI was recorded while participants repeated spoken sentences varying in acoustic intelligibility and semantic predictability. For sentences with a predictable final word, activation in regions of the semantic network was stronger with increasing intelligibility. Activity in an extended semantic network including frontal and ventromedial regions scaled to the individual comprehension gain from high predictability at intermediate intelligibility. In contrast, the cingulo-opercular network showed stronger activation at intermediate intelligibility of sentences with low predictability. Within this domain-general network, the inhibitory influence from pre-SMA to left insula increased when task demands were either highest or lowest, with the behavioural predictability gain being stronger under states of greater inhibition. This study demonstrates that successful speech comprehension under adverse conditions relies on the adaptive down-regulation of domain-general regions when semantic information is available.

Study 2 aimed to examine how the complex constraints of context in natural speech shape predictive processing. In a first step, we adapted a metric of similarity between words and their preceding context: Multidimensional embeddings representing the meaning of words in a text were correlated with the average embedding of all their preceding context words. For sentences used in Study 1, final words with high predictability were more similar to their preceding context than final words with low predictability, validating the applicability of semantic similarity as a proxy of word-by-word predictability. In a next step, fMRI was recorded while participants listened to a one-hour story embedded in a stream of resynthesized natural sounds. For each word in the story, the semantic similarity to its preceding context was determined at five timescales of increasing context length. We expected more dorsal regions in temporo-parietal cortex to be sensitive to the similarity of words at longer timescales, thereby organizing along a hierarchy of timescales consistent with accounts of predictive coding. However, no such hierarchy was observed.

Study 3 aimed to overcome the shortcomings of the similarity metric used in Study 2. We trained two artificial neural networks on a large corpus of text to predict upcoming words by their preceding context. The artificial neural networks operated on five layers drawing their predictions from increasingly abstract representations of context. Critically, one artificial neural network continuously updated new incoming words into context representations at all timescales, whereas the other network represented context in the form of distinct events that were updated at event boundaries. To determine how predictive context at different timescales is for upcoming words, we extracted probability distributions from single layers of artificial neural networks for each word in the story presented in Study 2. These distributions determine the probability of being the next word for each word in a large vocabulary of candidate words.

Study 4 tested whether predictive processing is hierarchically organized in temporo-parietal cortex. We used surprisal evoked by the actual next word at five timescales as derived from Study 3. Speed in a self-paced reading task increased for less surprising words, demonstrating the behavioural relevance of the machine-derived surprisal metric. Using fMRI data acquired in Study 2, we found word surprisal at longer timescales to evoke increased BOLD activity in more parietal regions, giving rise to a temporo-parietal surprisal hierarchy. Along this hierarchy, surprisal gated connectivity between temporal receptive windows of immediately shorter and longer timescales. Crucially, this hierarchy was only found for sparsely but not continuously updated surprisal. This demonstrates that the sparse updates inherent to the event-based representation of context pose an efficient coding scheme for predictive processing in speech comprehension.

4.2 Are artificial neural networks a valid model of language prediction in the human brain?

In Study 4, we found an event-based "surprisal hierarchy" to evolve along the temporo-parietal pathway, with word surprisal at longer timescales represented in more parietal cortex. We derived timescale-specific surprisal from artificial neural networks trained to predict the next word in a story. However, such artificial neural networks have been criticized as "black boxes", whose computations are neither transparent nor similar to human processing. Is it legitimate to treat the predictions from our machine learning routine as an adequate approximation of human language processing?

A comparative approach to how we can study language prediction with artificial neural networks

In recent years, there has been an increasing number of studies using artificial neural networks to study the neural underpinnings of language comprehension. While most artificial neural networks are trained on the objective to predict the next word, not all studies employing machine-derived word representations are suitable to investigate language prediction. What are the requirements for a study to successfully investigate speech prediction with artificial neural networks?

A first line of research extracts word-specific vector representations from *one* select layer of *one* artificial neural network. For example, such studies use vectors representing semantic features (e.g., Huth et al., 2016) or next-word probabilities (e.g., Heilbron et al., 2020), that is, metrics that are otherwise tedious to acquire for natural speech. The aim of such studies is to identify brain regions sensitive to semantic representations (e.g., Huth et al., 2016; Frank & Willems, 2017) or neural dynamics sensitive to interactions between semantics and lower-level phonetics (e.g., Broderick & Lalor, 2020; Heilbron et al., 2020). These studies are indifferent to the specific computations that produced word vector representations.

A second line of research extracts word-specific vector representations from *one* select layer of *more than one* artificial neural network. For example, such studies show that vectors derived from state-of-the-art transformer models (Radford et al., 2019) are more accurate in predicting neural responses to speech than word embeddings (Pennington et al., 2014), both in terms of spatial (Schrimpf et al., 2020) and temporal electrocorticographical responses (Goldstein et al., 2020). A critical difference between the compared metrics is that transformer models yield vector representations tuned to the context of a word, whereas embeddings are pre-trained, static vector representations insensitive to word context. The sheer number of architectural distinctions between the two artificial neural networks, however, makes it impossible to give a more nuanced evaluation of the computations accounting for observed differences in prediction accuracies. Therefore, these studies can be considered mere proof-of-concept studies confirming that the brain is sensitive to context-specific representations.

While these two lines of research highlight the general applicability of artificial neural networks as models of biological neural processing, they do not capitalize on the computational versatility of artificial neural networks as a window into the precise computations underlying biological neural processing.

Taking a step towards leveraging the computations performed by artificial neural networks, Chien and Honey (2020) presented two groups of participants with a story that was either preceded by the same or different contexts. Following, they used two computational models to simulate the temporal alignment of intersubject neural responses in receptive temporal windows of increasing length. Simulations showed that an artificial neural network with continuous, surprisal-gated updates to context representations matched biological response patterns best. While such simulation studies are an elegant means to test the neural fit of computational models against each other, they oftentimes rely on contrasting different experimental conditions.

Another line of research extracts word-specific vector representations from *each* layer of *one* artificial neural network. For example, context-specific layer activations (Jain & Huth, 2018) and layer units (Jain et al., 2020) derived from an LSTM encode temporo-parietal and frontal neural responses, with auditory cortex being more sensitive to short timescales represented at lower layers than more parietal and frontal regions. While these studies successfully replicate results from classic scrambling experiments, they lack the potency to trace back timescale effects to specific computations that could elucidate the organizational principle of context in cerebral cortex.

In study 4, we aimed to capture predictive processing at multiple timescales in natural speech and simultaneously specify the underlying computations. This called for an approach not established in computational language modelling yet (for an example in vision see e.g., Yamins et al., 2014). We extracted word-specific vector representations from *each* layer of *two* artificial neural networks. More specifically, we chose two multi-layered artificial neural networks that differed in one single computational feature (i.e., the boundary detector) but were otherwise identical (e.g., number of layers and units, objective and optimization function). First, this allowed us to attribute potential differences between effects of artificial neural networks on neural responses to this specific computational feature. Second, we set up models in a way that allowed us to derive a probability distribution for the next word from each single layer. This approach underscores that artificial neural networks are not a black box but provide an accessible set of parameters and weights that becomes interpretable by contrasting effects of different network architectures (Cichy & Kaiser, 2019).

Does the flexibility in training artificial neural networks come at the cost of generalizable results?

The number of “researcher degrees of freedom” in psychology (Silberzahn et al., 2018) and neuroscience (Botvinik-Nezer, 2020) is notoriously high, making it challenging to obtain results that generalize to new data and alternative analysis workflows. The variability in analytic choices is even greater when using outputs from artificial neural networks to model biological neural responses. For a start, researchers have to make many arbitrary choices on the architecture and parameter settings of the artificial neural networks they wish to train. For another, it is feasible to set up and train many different artificial neural networks until one network fits biological data sufficiently well. However, such flexibility in data analysis generally increases the risk of reporting spurious (or false positive) effects (Simmons et al., 2011).

The iterative process of tuning artificial neural networks is common practice in machine learning and an inevitable prerequisite for good model performance. In Study 3, I describe how we tested different combinations of layer and unit quantities in artificial neural networks to arrive at a satisfactory accuracy of predicting words from their context. This procedure parallels the construction of a questionnaire in psychology, where items are repeatedly revised for quality criteria before they are finally deployed in an experiment. The decisive point in such a procedure is that parameter tuning is carried out *before* the machine-based metric is used to model any biological data. This entails that the choice of parameters is not biased toward its fit to biological data and therefore does not increase the risk of finding false positive results in following analyses.

Even if the selection of an artificial neural network is made independently of biological data, the artificial neural network selected may in itself be unreliable. A standard approach in machine learning also used in Study 3 is to select the final artificial neural network after testing many different architectures by means of cross-validation (e.g., see Arlot & Celisse, 2010). The rationale behind cross-validation is that we can reduce the risk of selecting a network that has “overlearned” the regularities (and thereby noise components) of the data it was trained on by testing its ability to generalize to unseen data. We applied this logic not just when selecting artificial neural networks in Study 3 but also when estimating models on biological neural data in Study 4 (Varoquaux et al., 2017). This two-stage cross-validation scheme suggests that our surprisal metric as well as the established link between this metric and biological data is robust against testing temporo-parietal predictive processes on new text stimuli and additional participants.

To strengthen the credibility of our results, we limited our analyses to temporo-parietal regions of interest and formulated the concrete hypothesis of a neural hierarchy coding for surprisal at multiple timescales (Kerr, 1998). While this reduced the overall number of analyses possible (Bishop, 2020), the attentive reader might object that we had carried out an analysis with a negative result on the same dataset already in Study 2. This concern is at least partially invalidated by the fact that the similarity metric used in Study 2 and the surprisal metric used in Study 4 are conceptually fundamentally different from one another. Nevertheless, we could have further strengthened the credibility of our results by pre-registering a detailed statistical analysis plan and blinding the data analyst to the type of artificial neural network (e.g., LSTM vs. HM-LSTM) until the end of analysis (Munafò et al., 2017).

Together, this demonstrates that training artificial neural networks without knowledge of biological data, cross-validating the generalizability of estimated models, and limiting the number of analyses by stating clear hypotheses are powerful strategies to overcome the challenges posed by the flexibility in constructing artificial neural networks. The virtue of these methodological decisions is not least illustrated by the fact that the event-based “surprisal hierarchy”, an effect most central to our hypothesis in Study 4, proved meaningful on various benchmarks. In particular, statistical post-hoc power was well above 99% (one sample t-test, d = 0.72, alpha level = 0.05, n = 34; Faul et al., 2007), the p-value was small enough to easily reach a more stringent significance threshold of 0.005 (Benjamin et al., 2018) and the effect was significant in more than half of participants (Smith & Little, 2018).

Are artificial neural networks a proxy of human behaviour emerging from language prediction?

When projecting neural responses onto machine-derived outputs, the underlying assumption is that biological computations manifest in behaviour and that artificial computations producing outputs that resemble this behaviour are more likely to comply also with the underlying biological computations. The potency to describe behaviour is considered a touchstone for the suitability of artificial neural networks for modelling human processing (Saxe et al., 2021).

The artificial neural networks trained in Study 3 determined the probability of being the next word for each word in a large vocabulary of candidate words. In Study 4, we extracted the probability of the actual next word and transformed it to surprisal (Hale, 2016), an index of the amount of information conveyed by a word that cannot be explained away by its preceding context. We validated model-based surprisal by testing whether its effect on behavioural responses accords with effects observed for standard measures of word surprisal. In line with studies deriving word surprisal from Markov models (Smith & Levy, 2008), cloze probability tests (Lowder et al., 2018), recurrent neural networks (Monsalve et al., 2012) and LSTM networks (Goodkind & Bicknell, 2018), we found participants to slow down when presented with a surprising word in a self-paced reading task. Consistently, we also found the pupil to dilate when reading more surprising words (unpublished data by Schmitt et al.; for similar results see Frank & Thompson, 2012).

The effect of machine-based surprisal on reading speed yielded statistically robust results differentiable from potentially confounding effects like word length. However, this analysis tested the similarity between machine-based and human-based surprisal only indirectly by inferring the relationship between the two from their concurring influence on another variable, namely reading speed or pupil dilation. A more direct measure would be to pit machine-based and human-based predictions against each other in the form of a group statistic (Ma & Peters, 2020). For example, one could perform a cloze task where participants fill in the blanks in a text with the most probable continuation and compare these human-derived predictions to machine-derived predictions (see Goldstein et al., 2020).

As the cloze procedure provides one predictability value accumulated over all context available, it is not suitable when interested in the predictability of a word at a specific timescale. An approach that is straightforward to implement and often used to study timescale-specific effects is to scramble speech at different linguistic units (e.g., sentence vs. paragraph; see Lerner et al., 2011). In theory, one could collect cloze probabilities on such scrambled texts to determine how similar they are to the timescale-specific predictions of artificial neural networks. However, a drawback to such an approach is that one would have to know how to scramble texts so that they match the actual timescales in humans—which was exactly the question we were aiming to answer in the first place. An alternative approach to specifically probe the validity of the event-based HM-LSTM would be to compare machine-based event boundaries to human event segments by instructing participants to identify event boundaries at different levels of granularity.

Are artificial neural networks a proxy of the biological neural networks underlying language prediction?

The computations carried out by artificial neural networks are complex, as they entail a large number of subroutines and depend on a large number of parameters. At what level of detail can we interpret these computations in terms of biological neural processes and, more specifically, what can we learn from these computations about language prediction in temporo-parietal cortex regions?

A first important computational component of an LSTM (in this case also subsuming the HM-LSTM) is the formation of memory. According to the jargon of machine learning, each layer of an LSTM comprises short-term and long-term memory representations of context. While these terms may sound familiar to cognitive neuroscientists, they mean something fundamentally different in machine learning.

The long-term memory of an LSTM exists in the form of a vector, which codes for context that might become relevant at *some* point. We provided LSTMs with a maximum number of 500 context words only, yet long-term memory in cognitive neuroscience refers to the storage of knowledge and events acquired over a lifetime (Norris, 2017). As long-term memory representations of artificial neural networks are subject to frequent updates, the nature of these representations is more similar to what psychologists mean by short-term memory: keeping a limited amount of information temporarily accessible (Cowan, 2008).

The short-term memory of an LSTM represents information that is expected to be predictive of the next word. As these short-term memory representations are used to derive probability distributions for the prediction of the next word, they can rather be referred to as working memory in neuroscientific terms, which is often used interchangeably with the term short-term memory but emphasizes the components of short-term memory specifically used to plan and carry out behaviour (Cowan, 2008). Is the conception of LSTM cells as working memory further supported by our empirical results in Study 4?

Previous studies have shown that LSTMs operate on increasingly long timescales, with lower layers coding for local (syntactical) information and higher layers coding for long range, context-dependent relationships (Peters et al., 2018). We replicated this hierarchy of context length and complexity for the sparsely updating HM-LSTM by showing that word prediction accuracy for lower layers is less affected by scrambling words at longer timescales than higher layers. Additionally, we showed that the timescales of word surprisal derived from the event-based HM-LSTM unfold along the temporo-parietal pathway. This result converges with accounts on working memory proposing that short-term information is stored along a posterior-to-frontal axis, with detailed information represented in sensory regions and abstract information represented in prefrontal regions (Christophel et al., 2017).

Beyond the maintenance of information for a limited period, a key function of working memory is that it groups information into smaller chunks (Cowan, 2008). In Study 4, we used two different artificial neural network architectures to test how context must be chunked in working memory to support speech prediction and found that next word prediction relies on the event-based organization of these working memory representations in the HM-LSTM. The events were defined by a boundary detector, yet we lack understanding of the criterion used by this detector to define these boundaries. Critically, the granularity of HM-LSTM segmentation has been shown to be susceptible to the manipulation of model parameters (Kádár et al., 2018), questioning that machine-derived boundaries match states of increased neural activity observed at the end of human-derived boundaries (Zacks et al., 2001; Ditman et al., 2008; Whitney et al., 2009). This highlights that with our five timescales we do not necessarily model all timescales also present in neural processing in detail but that we rather model a basic computational principle in general.

Assuming that HM-LSTM layers represent hierarchical working memory representations, what is the essence of prediction processes in these networks? One current opinion is that artificial neural networks do not model the generative structure of the world but instead learn an over-parameterized direct fit to data (Hasson et al., 2020). This view holds that artificial neural networks are limited to making correct predictions for test data falling in the range of examples presented during training, thereby being unable to generalize to data outside the “interpolation zone". Importantly, Hasson and colleagues (2020) do not interpret this as a shortcoming of artificial neural networks but rather draw an analogy between computations of artificial neural networks and the general principle of evolution. Indeed, such model-free (stochastic) processes have been observed in humans and are considered a computationally less demanding processing route (Niv, 2009).

A complementary processing route for more demanding prediction processes might entail more complex generative models. Important features of such a route might be the ability to learn and relearn generative models (but see Saxe et al., 2019), to shape predictive processes by top-down attention, and to integrate long-term knowledge as well as cognitive control into predictive processes. A major effort in machine learning is directed towards the development of artificial neural networks that extract such complex, human-like generative models (Lake et al., 2017; Barrett et al., 2019).

In general, most of the similarities between artificial and biological neural networks proposed here work by analogy and dismiss the details of computations carried out (Marr, 1982). However, our results demonstrate that such analogies are a powerful means to investigate the neural computations underlying predictive processing, in particular with respect to the formation of working memory representations laying the foundation for making predictions. This view is largely in line with current accounts proposing to focus on the general principles or parameters of artificial neural networks in modelling brain function (e.g., architecture, learning rate, cost function; Saxe et al., 2021) instead of aiming to recreate the brain with its billions of neurons.

4.3 Revisiting the metrics of language prediction

Throughout this thesis, I have targeted prediction processes in humans in many different ways. I presented participants with controlled vs. natural language, I derived predictions from humans vs. machines, and I characterized the predictability of words in terms of cloze probability, surprisal, and similarity. What can we learn from the differential effects these factors have on neural responses about the mechanisms underlying prediction in language comprehension?

Cloze probability and surprisal

The central finding in Study 4 was the "surprisal hierarchy" in temporo-parietal cortex. We found that brain regions from middle temporal gyrus to angular gyrus are sensitive to word surprisal at increasingly coarse events. A similar spatial pattern was found in Study 1, where activity in lateral temporo-parietal regions like posterior middle temporal gyrus and angular gyrus increased for sentences with a highly predictable final word (unpublished data; Rysop et al., 2021), especially when speech was more intelligible. Are the controlled and natural listening studies tapping into the same temporo-parietal network?

One difference between the observed spatial patterns is that BOLD activity in Study 1 increased in a temporal and another parietal cluster but intermediate regions at the border of temporal and parietal cortex showed no effect of predictability. At first glance, this suggests no implication of a processing hierarchy extending from temporal to parietal regions.

An explanation for this might be that we applied a strict statistical threshold (family-wise error correction for multiple comparisons) to whole-brain statistical maps on the group level, thereby capitalizing on large statistical effects. Indeed, also clusters of surprisal at intermediate timescales in Study 4, which are located at the border between temporal and parietal cortex, were smaller and weaker than those for shortest timescales in lower temporal and longest timescales in inferior parietal regions. Another potential explanation for the absence of an effect at intermediate regions is that participants listened to isolated sentences devoid of complex contextual dependencies. When context operates on a single timescale (or event), it might not be possible (or necessary) to recruit the full processing hierarchy with its multiple timescales. Instead, the temporo-parietal pathway may flexibly adapt to the structure of context.

A crucial difference between temporo-parietal activations in the controlled and natural listening study is that activity in Study 4 increased for more surprising words (i.e., low predictability), whereas activity in Study 1 increased for sentences with more constraining context (i.e., high predictability). How can we resolve this apparent contradiction of increased vs. decreased activity to more predictable language in the same brain region?

As outlined in the discussion section of Study 4, we interpreted the increase of activation in response to more surprising words as an increase of prediction errors and prediction updating in the face of inaccurate predictions. When the constraint of context is high but the prediction is violated, surprisal is high. However, when the constraint of context is low, the number of plausible candidate words is high and these candidates share an equally low probability of being the next word. This is reflected in a low precision of the prediction, which can be calculate as the entropy of a probability distribution. Critically, word surprisal is thought to be weighted by precision (Kwisthout et al., 2017), so that the prediction error for imprecise predictions is diminished. For instance, Weissbart and colleagues (2020) show in EEG that precision-weighted surprisal modulates central and occipital responses to words during natural speech comprehension.

While we would expect that surprisal in natural speech is down-weighted by precision in some cases (i.e., no clear prediction) but not in others (i.e., clear prediction), surprisal in Study 1 should be suppressed in all sentences with low predictability. For example, the sentence "We are very pleased with the new" poses no constraints on the final word "sheets" (i.e., low precision), which should result in strong suppression of the prediction error, and therefore lower overall activity in temporo-parietal regions involved in predictive processing. In contrast, all sentences with high contextual constraint end with the most probable continuation. As language prediction is thought to be probabilistic (Kuperberg & Jaeger, 2016), though, the brain likely activates, at least to a certain degree, additional candidate continuations that are also probable. Therefore, even the most probable continuation will elicit a small prediction error. In a scenario where we contrast sentences of low constraint (and suppressed prediction error) with sentences of high constraint (and small prediction error), BOLD activity is higher for sentences with a final word of high compared to low predictability. This suggests that the activation patterns seen in Study 1 and Study 4 are compatible but have likely been blurred by effects of precision weighting.

In the discussion section of Study 1, we interpreted the recruitment of temporo-parietal regions as up-regulation of the language network when informative semantics are available that can guide speech comprehension. In light of the results of Study 4, I suggest that the activity increase for highly predictable sentences might not just reflect language processing in general but predictive processing more specifically.

This hypothesis is challenged by the fact that predictability in Study 1 is confounded by semantic richness. In particular, the number of "pointer" words in a sentence, which semantically link to the sentence-final word (Kalikow et al., 1977), manipulates predictability. These pointer words are concrete words like "bed" in sentences with high constraint and contrast with more abstract words like "pleased" in sentences with low constraint. As concrete vs. abstract words are known to have disparate behavioural (Recchia & Jones, 2012) and neural effects (Fiebach & Friederici, 2004), the increase of activity in temporo-parietal regions for sentences with high predictability might also be driven by more general effects of semantic processing.

In sum, our results demonstrate that activity in temporo-parietal regions increases when the pre-assigned precision of a prediction is high but the prediction is violated as indexed by increased surprisal when finally encountering the actual next word. This highlights the importance of also taking into account precision when modelling predictive processing in the brain, a factor we have not explicitly considered in our experiments.

Semantic similarity

In an initial attempt to model the neural dynamics of predictive processing in Study 2, we selected the semantic similarity of a word to its preceding context as a metric. Similarity is calculated by correlating the embedding of a target word with the average embedding of the content words preceding it (adapted from Frank & Willems, 2017; Broderick et al., 2018).

In previous studies, the similarity between a word and its context has been shown to modulate the N400 component in EEG (Camblin et al., 2007; Metusalem et al., 2012; Paczynski & Kuperberg, 2012), typically considered a brain signature sensitive to sentence constraint (Kutas & Federmeier, 2011). This suggests that word predictability hinges on semantic similarity, and indeed we found the similarity of final words to their preceding context to be higher when final words were highly predictable from context.

When calculating semantic similarity at different timescales, (i.e., by using different lengths of preceding context), we found lower temporal regions to be more sensitive to similarity at shorter timescales and higher parietal regions to be more sensitive to longer timescales. However, we did not find a gradient of timescales from lower to higher-order temporo-parietal regions that would indicate consistent ordering of timescales. The brain regions modulated by similarity are the same regions found by Frank and Willems (2017) when modelling semantic similarity only at one (short) timescale, thereby raising doubts that similarity is relevant to the computations of hierarchical language prediction.

In a recent study, Broderick and Lalor (2020) used semantic similarity to study human speech prediction in natural listening scenarios. In an EEG experiment, they showed that lower-level envelope and phonetic features of speech were encoded more strongly when higher-order words were more surprising but also when they were more similar to their preceding context. This result was interpreted in terms of predictive coding, where top-down effects of surprisal on phonetic encoding would represent the prediction error and top-down effects of similarity would represent the prediction. An important implication of this interpretation is that the internal model used by the brain to make a prediction on upcoming speech is simply to average across semantic representations of preceding context.

However, language prediction is thought to go beyond such a mere effect of lexical association (Camblin et al., 2007), which is nicely reflected by the fact that syntactic predictions cannot be explained by association but must be inferred from more complex dependencies (van Berkum et al., 2005). To illustrate this point, in a scenario where our prediction is the average semantic representation of context, we would only achieve perfect predictability when speech is as redundant as that of a talking parrot.

An explanation for the effects of similarity observed by Broderick and Lalor (2020) might be that they represent a less controlled account of language prediction, where concepts related to context are activated in terms of spreading activation like in priming (Lau et al., 2013). Such a strategy might work particularly well when no explicit prediction can be made and might be a complementary route in tuning expectations of upcoming speech in lower processing stages.

4.4 A unifying framework for the biological neural networks of language prediction

The present thesis characterized the biological neural networks implicated in processing the predictability of speech. While Study 1 focused on differential contributions of the domain-general cingulo-opercular network and the domain-specific language network, Study 4 focused on a sub-network within the language network, that is, the temporo-parietal part of the auditory dorsal pathway. Here, I aim to integrate four biological neural networks into a broader framework of language prediction: (1) auditory dorsal and (2) ventral pathway of the language network, (3) inferior frontal gyrus, and (4) hippocampus (Figure 4.1).

The auditory dorsal pathway as a network of event-based language prediction

The starting point for the development of the language prediction framework is the temporo-parietal surprisal hierarchy we found in Study 4.

To recapitulate, the brain segregates context into multiple events of increasing length and stores these events along the auditory dorsal pathway in working memory, with longer and temporally more distant events represented in more parietal regions (Baldassano et al., 2017). Computationally, such timescale-specific event representations have to be updated only at event boundaries (Chung et al., 2016). More specifically, when an event at a shorter timescale ends, it is recombined with an event represented at a longer timescale.

The main finding of Study 4 was that the brain uses these temporo-parietal event representations as context for making predictions. The observed surprisal hierarchy indicates that the event represented at a distinct timescale is used to inform a timescale-specific prediction about the next word. As we have shown that connectivity to neighbouring timescales increases when a timescale-specific prediction is not matched by the actual next word, timescales seem to hierarchically interact in line with accounts of predictive coding (Rao & Ballard, 1999; Friston, 2005;). In particular, this suggests that the prediction error is fed forward to an immediately longer timescale, where the prediction is updated and fed back to an immediately shorter timescale (Chao et al., 2018). Following from the probabilistic nature of word prediction (Kuperberg & Jaeger, 2016), the prediction made at a timescale should be transferred to the immediately lower processing stage in the form of a probability distribution across words, which is then combined with the probabilistic prediction at this lower processing stage.

The dorsal auditory prediction hierarchy allows to flexibly draw on rich semantic context represented in the form of multiple events at different levels of abstraction. I put forward that such event-based predictions are functionally advantageous compared to predictions derived from context representations continuously tuned to current word input (Hochreiter & Schmidhuber, 1997; Chien & Honey, 2020), as event-based prediction processes can fall back on other, independent events when some events are currently uninformative for the prediction of upcoming words.

Finally, I argue that the temporo-parietal surprisal hierarchy extends to lower-level phonetic processing stages to provide a prediction at a scale matching the incoming sensory signal (Gwilliams et al., 2018, 2020). At these lower processing stages, also the uncertainty about the final prediction might be encoded to adjust for the degree of effort needed to accurately sample the incoming signal (Gwilliams & Davis, in press) and to weight the prediction error (Kwisthout et al., 2017).

The auditory ventral pathway as a network of time-invariant language prediction

While the auditory dorsal pathway is thought to inform predictions by event-based context, not all facets of context are best described by such an event structure. For example, the event-based organization of context does not allow linking two related events interrupted by an unrelated event in working memory. The auditory ventral pathway has been proposed to fulfil exactly such a function by chunking speech features into increasingly abstract concepts, irrespective of the temporal order they were presented in (Bornkessel-Schlesewsky & Schlesewsky, 2013).

The auditory ventral pathway, mirror-symmetric to the dorsal pathway, emanates from early auditory cortex and extends into lateral anterior temporal cortex (Hickok & Poeppel, 2004, 2007; Friederici & Gierhan, 2013). This region has been shown to be activated in language processing in general (Wilson et al., 2008; Binder et al., 2011) and predictive language processing in particular (Willems et al., 2016; Davis & Hasson, 2018), with more complex speech features represented in more anterior temporal regions (DeWitt & Rauschecker, 2012; Sheng et al., 2018) Additionally, anterior temporal lobe has been shown to be sensitive to parsing strategies that allow to efficiently process sentence structures in which phrases are embedded in other phrases (e.g., “The plumber [who the contractor likes] visited the home”; Brennan & Pylkkänen, 2017). This is in support of my hypothesis that the ventral stream may contribute to hierarchical speech prediction by exploiting another, more “nested” or “branched” facet of context.

In Study 1 and 4, we found no significant effects in anterior temporal cortex. However, this does not *per se* preclude an implication of the ventral pathway in predictive language comprehension. In general, it is technically difficult to capture activation in this brain region using fMRI (Visser et al., 2010). In Study 1, we likely observed no anterior temporal effects because sentence stimuli neither had a nested semantic nor syntactical structure (Humphries et al., 2001, 2006). In Study 4, we likely observed no anterior temporal effects because we modelled only event-based predictive processes, which I firmly assume not to be represented in anterior temporal regions.

If there is indeed a second, auditory ventral hierarchy of time-invariant language prediction in play, this raises the question how this hierarchy might interact with the auditory dorsal prediction hierarchy. The two hierarchies might either interact only at shorter timescales to combine their independent predictions into one final prediction, or they might interact also at longer timescales. Previous studies do not give a clear picture of the connectivity between higher-order ventral (i.e., anterior temporal cortex) and dorsal brain regions (i.e., angular gyrus; e.g., Binney et al., 2012; Jackson et al., 2016). From a theoretical point of view, though, it is doubtful that predictions are combined already at longer timescales because these predictions are based on very different aspects of context at potentially incompatible levels of abstraction.

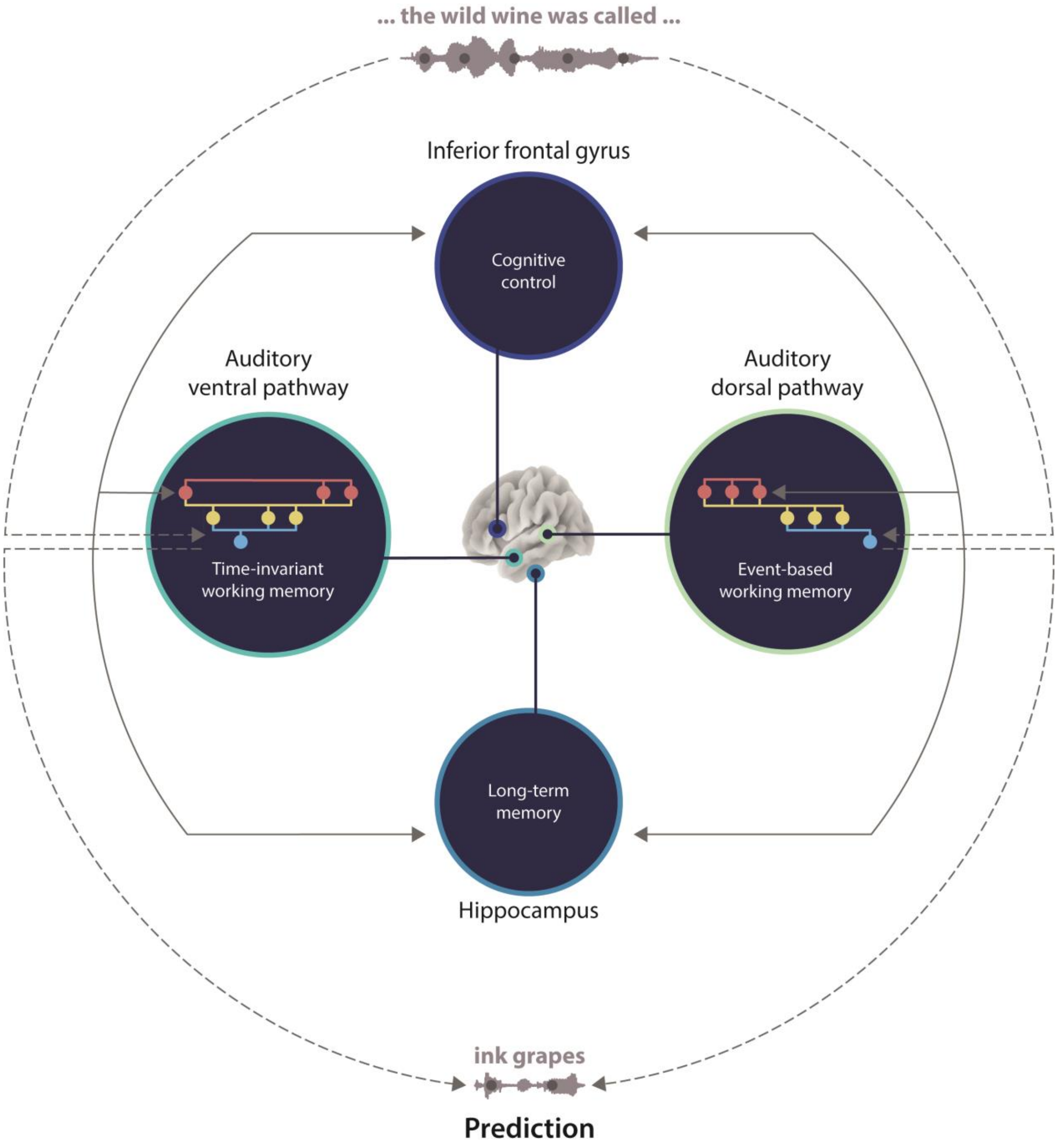


Figure 4.1. A unifying framework for the biological neural networks of language prediction. The incoming speech signal (sound wave at the top; dots representative of words) is segmented into multi-timescale working-memory representations by two complementary pathways. The auditory dorsal pathway (right blue circle) segregates context into increasingly long and temporally more distant events (red dots), with longer timescales represented in more parietal cortex regions. The auditory ventral pathway (left blue circle) segregates context into increasingly abstract representations irrespective of the temporal order they were presented in, with longer timescales represented in more anterior temporal cortex regions. The context representations in working memory are used to make predictions on upcoming words by feeding the prediction error forward to an immediately longer timescale, where the prediction is updated and fed back to the immediately shorter timescale (not depicted). The formation of predictions is shaped by long-term memory representations via hippocampus (bottom blue circle) and by cognitive executive control via inferior frontal gyrus (IFG; top blue circle), both interfacing with the auditory dorsal and ventral prediction hierarchy at long timescales. Finally, the prediction of the next word (sound wave at the bottom) is read out from the shortest timescale of the ventral and dorsal pathway and combined into a final prediction.

Hippocampus as a hub of long-term memory based language prediction[6]

Another reservoir of rich context suitable to inform predictions is long-term memory. For example, one might recall talking to someone about the very same topic a couple of weeks ago, so this context becomes relevant to predicting what the interlocutor might be about to say in the current situation. How are such long-term memories integrated with short-term working memory representations in language prediction?

Indeed, short-term and long-term memory systems can be couched into the larger framework of the dual reference frame system (Bottini & Doeller, 2020), where flexible sensory knowledge in parietal cortex interacts with stable conceptual knowledge in hippocampus. As parietal cortex has been shown to functionally interface with the broader medial temporal lobe regions (Uddin et al., 2010; Sestieri et al., 2017), and more specifically with hippocampus at event boundaries of longer timescales during movie watching (Baldassano et al., 2017), I speculate that the auditory dorsal hierarchy of speech prediction might functionally extend from receptive windows in parietal cortex to hippocampus.

Consistent with the key characteristics of the speech prediction hierarchy, hippocampus codes for boundaries in the environment (Spiers et al., 2015; Brunec et al., 2018), hierarchically organizes memories (Alexander & Nitz, 2017) and engages in predictive coding (Johnson & Redish, 2007; Stachenfeld et al., 2017). In Study 1, we found that an extended language network including ventromedial regions and specifically anterior parahippocampal gyrus scaled in activation with the individual behavioural predictability gain under adverse listening conditions, thereby underscoring the relevance of long-term memories in successful predictive language comprehension.

Inferior frontal gyrus as a hub of cognitive control in language prediction

The inferior frontal gyrus (IFG), alongside premotor cortex, is deemed the apex of the auditory ventral and dorsal pathway (Hickok & Poeppel, 2007). Building up on the proposal that the temporo-parietal hierarchy of working memory extends to frontal regions (Christophel et al., 2017), how might these regions support language prediction?

[6] This section was partly adapted from Schmitt et al. (2020).

In Study 1, we found that the behavioural predictability gain at intermediate intelligibility activated more ventral portions of bilateral IFG. These ventral regions have been associated with controlled semantic retrieval (Davey et al., 2016) and semantic tasks that require activating additional semantic knowledge (Badre et al., 2005). This implies that the recruitment of ventral regions facilitates language prediction when context is rich but difficult to interpret, possibly by providing complex internal models from which to infer upcoming speech.

In contrast, we found higher activation of dorsal IFG in the left hemisphere under conditions of low semantic predictability and increasing acoustic intelligibility. These dorsal regions have been associated with top-down executive cognitive control in challenging tasks (Davey et al., 2016) and semantic tasks that require selecting one candidate over competing candidates (Badre et al., 2005). This implies that the recruitment of dorsal regions facilitates language prediction when context is ambiguous and there are many candidate words.

The ventral (semantic control) and dorsal (multiple demand) regions of IFG have been functionally linked in previous studies showing a posterior-anterior gradient of cognitive control along ventrolateral prefrontal cortex (Koechlin et al., 2003; Badre & D'Esposito, 2007), with more controlled processes represented in more anterior (and dorsal) regions (Bahlmann et al., 2015; Jeon & Friederici, 2015). Transferring this hierarchical organization in IFG to language prediction, I argue that ventral IFG is recruited more strongly when prediction is challenged by complex contextual dependencies, whereas dorsal IFG is additionally recruited when contextual constraints are low.

How does IFG connect to the temporo-parietal prediction hierarchy? Previous studies showed that activity in IFG relies on longer timescales of speech being intact (Wilson et al., 2008; Lerner et al., 2011) and that expectations drive connectivity between IFG and superior temporal gyrus (Phillips et al., 2016; Garrido et al., 2018;). This suggests that IFG predominantly interacts with the longer timescales of the auditory dorsal (and by extension possibly ventral) pathway to enrich working memory based predictions by higher-order control processes under challenging conditions.

4.5 Conclusion

To return to the initial question of this thesis: How is the prediction of language supported by the dynamic interplay within and between biological neural networks?

First, stronger activation of domain-specific language and long-term memory networks as well as inhibition within the domain-general cingulo-opercular network facilitated the behavioural predictability gain under adverse listening conditions. This indicates that the cingulo-opercular network deploys executive control processes that are adaptively recruited when contextual constraint is low. Second, the similarity of words to their preceding context is not an appropriate model of predictive processing in temporo-parietal regions. At most, similarity explains less elaborate effects of spreading activation. Third, event-based artificial neural networks proved suitable for modelling word-by-word predictions at multiple timescales. In particular, comparing differential effects of multiple artificial neural network architectures on biological neural responses is a powerful means to identify specific computations relevant to processing in the brain. Fourth, the temporo-parietal surprisal hierarchy found in response to natural speech indicates that event-based context representations are used to make timescale-specific predictions about the next word. I suggest that the event-based organization of context provides a semantically rich and computationally efficient network architecture for predictive processing in general.

Finally, I put forward that the dynamic interplay between biological neural networks of working memory, long-term memory and cognitive control shapes language prediction. More specifically, I propose that context representations along event-based dorsal and time-invariant ventral hierarchies in auditory regions subserve the formation of predictions at multiple timescales. Predictive processing at longer timescales might interface with long-term memories via hippocampus. IFG might exert cognitive control over longer timescales in auditory regions, with ventral IFG providing complex internal models for making predictions and dorsal IFG supporting predictions when context is ambiguous.

The results demonstrate that language prediction relies on a set of elaborate computations—multi-layered internal models, event-based context representations, and hierarchical prediction passing—as well as the adaptive recruitment of cognitive control and long-term memory networks. Together, this demonstrates a pivotal role of predictive processing in successful speech comprehension.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs.DC].*

Aborn, M., Rubenstein, H., & Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology*, *57*(3), 171–180.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*(14), 1–10.

Abrams, D. A., Ryali, S., Chen, T., Balaban, E., Levitin, D. J., & Menon, V. (2013). Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. *Cerebral Cortex*, *23*(7), 1703–1714.

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. *Brain and Language*, *122*(1), 42–54.

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Alain, C., Du, Y., Bernstein, L. J., Barten, T., & Banai, K. (2018). Listening under difficult conditions: An activation likelihood estimation meta-analysis. *Human Brain Mapping*, *39*(7), 2695–2709.

Alexander, A. S., & Nitz, D. A. (2017). Spatially Periodic Activation Patterns of Retrosplenial Cortex Encode Route Sub-spaces and Distance Traveled. *Current Biology*, *27*(11), 1551-1560.e4.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801.

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *arXiv:1607.06450 [stat.ML].*

Badre, D., & D'Esposito, M. (2005). Dissociable Controlled Retrieval and Generalized Selection Mechanisms in Ventrolateral Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *19*(12), 2082–2099.

Badre, D., & D'Esposito, M. (2007). Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. *Neuron*, 47, 907–918.

Bahlmann, J., Blumenfeld, R. S., & D'Esposito, M. (2015). The Rostro-Caudal Axis of Frontal Cortex Is Sensitive to the Domain of Stimulus Information. *Cerebral Cortex*, *25*(7), 1815–1826.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709-721.e5.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*(3), 364–390.

Bao, P., Purington, C. J., & Tjan, B. S. (2015). Using an achiasmic human visual system to quantify the relationship between the fMRI BOLD signal and neural response. *ELife*, *4*, e09600.

Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. *11th Language Resources and Evaluation Conference*, 792–797.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology*, *55*, 55–64.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron*, *85*(2), 390–401.

Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J., & Ungerleider, L. G. (2016). Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. *Current Biology*, *26*(17), 2280–2290.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*(1), 289–300.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2nd ed.). *Springer*.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a Test of Speech Perception in Noise. *Journal of Speech, Language, and Hearing Research*, *27*(1), 32–48.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., Grabowski, T. J., Langfitt, J. T., Loring, D. W., Lowe, M. J., Koenig, K., Morgan, P. S., Ojemann, J. G., Rorden, C., Szaflarski, J. P., Tivarus, M. E., & Weaver, K. E. (2011). Mapping anterior temporal lobe language areas with fMRI: A multicenter normative study. *NeuroImage*, *54*(2), 1465–1475.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, *7*(3), 295–301.

Binney, R. J., Parker, G. J. M., & Lambon Ralph, M. A. (2012). Convergent Connectivity and Graded Specialization in the Rostral Human Temporal Lobe as Revealed by Diffusion-Weighted Imaging Probabilistic Tractography. *Journal of Cognitive Neuroscience*, *24*(10), 1998–2014.

Bishop, D. (2020). How scientists can stop fooling themselves. *Nature*, *584*, 9.

Blank, H., & Davis, M. H. (2016). Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. *PLOS Biology*, *14*(11), e1002577.

Boldt, R., Malinen, S., Seppä, M., Tikka, P., Savolainen, P., Hari, R., & Carlson, S. (2013). Listening to an Audio Drama Activates Two Processing Networks, One for All Sounds, Another Exclusively for Speech. *PLoS ONE*, *8*(5), e64489.

Bonnici, H. M., Richter, F. R., Yazar, Y., & Simons, J. S. (2016). Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *The Journal of Neuroscience*, *36*(20), 5462–5471.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Reconciling time, space and function: A new dorsal–ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60–76.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in Cognitive Sciences*, *19*(3), 142–150.

Bottini, R., & Doeller, C. F. (2020). Knowledge Across Reference Frames: Cognitive Maps and Image Spaces. *Trends in Cognitive Sciences*, *24*(8), 606–619.

Botvinik-Nezer, R. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*, 84–88.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brennan, J. R., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41*(S6), 1515–1531.

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, *14*(6), 277–290.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803-809.e3.

Broderick, M., & Lalor, E. (2020). Co-existence of prediction and error signals in electrophysiological responses to natural speech. *bioRxiv*. https://doi.org/10.1101/2020.11.20.391227

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 1–14.

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44.

Brunec, I. K., Moscovitch, M., & Barense, M. D. (2018). Boundaries Shape Cognitive Representations of Spaces and Events. *Trends in Cognitive Sciences*, *22*(7), 637–650.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*, *58*(5), 412–424.

Buračas, G. T., Zador, A. M., DeWeese, M. R., & Albright, T. D. (1998). Efficient Discrimination of Temporal Patterns by Motion-Sensitive Neurons in Primate Visual Cortex. *Neuron*, *20*(5), 959–969.

Burt, J. B., Demirtaş, M., Eckner, W. J., Navejar, N. M., Ji, J. L., Martin, W. J., Bernacchia, A., Anticevic, A., & Murray, J. D. (2018). Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nature Neuroscience*, *21*(9), 1251–1259.

Bzdok, D., Hartwigsen, G., Reid, A., Laird, A. R., Fox, P. T., & Eickhoff, S. B. (2016). Left inferior parietal lobe engagement in social cognition and language. *Neuroscience and Biobehavioral Reviews*, *68*, 319–334.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128.

Cameron, C. A., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*(2), 329–342.

Camilleri, J. A., Müller, V. I., Fox, P., Laird, A. R., Hoffstaedter, F., Kalenscher, T., & Eickhoff, S. B. (2018). Definition and characterization of an extended multiple-demand network. *NeuroImage*, *165*, 138–147.

Caspers, S., Eickhoff, S. B., Geyer, S., Scheperjans, F., Mohlberg, H., Zilles, K., & Amunts, K. (2008). The human inferior parietal lobule in stereotaxic space. *Brain Structure and Function*, *212*(6), 481–495.

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., & Zilles, K. (2006). The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. *NeuroImage*, *33*(2), 430–448.

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., & Dehaene, S. (2018). Large-Scale Cortical Networks for Hierarchical Prediction and Prediction Error in the Primate Brain. *Neuron*, *100*(5), 1252-1266.e3.

Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H., & Wang, X.-J. (2015). A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron*, *88*(2), 419–431.

Cheung, V. K. M., Meyer, L., Friederici, A. D., & Koelsch, S. (2018). The right inferior frontal gyrus processes nested non-local dependencies in music. *Scientific Reports*, *8*, 3822.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Chien, H.-Y. S., & Honey, C. J. (2020). Constructing and Forgetting Temporal Context in the Human Cerebral Cortex. *Neuron*, *106*(4), 675-686.e11.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124.

Chung, J., Ahn, S., & Bengio, Y. (2016). Hierarchical Multiscale Recurrent Neural Networks. a*rXiv:1609.01704 [cs.LG]*.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Clark, A. (2013a). The many faces of precision (Replies to commentaries on “Whatever next? Neural prediction, situated agents, and the future of cognitive science”). *Frontiers in Psychology*, *4*, 1–9.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clos, M., Langner, R., Meyer, M., Oechslin, M. S., Zilles, K., & Eickhoff, S. B. (2014). Effects of prior information on decoding degraded speech: An fMRI study. *Human Brain Mapping*, *35*(1), 61–74.

Cocchi, L., Sale, M. V., L Gollo, L., Bell, P. T., Nguyen, V. T., Zalesky, A., Breakspear, M., & Mattingley, J. B. (2016). A hierarchy of timescales explains distinct effects of local inhibition of primary visual cortex and frontal eye fields. *ELife*, *5*, e15252.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *10*(1), 37–46.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, *20*(3), 304–313.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in Brain Research* (pp. 323–338). Elsevier.

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*, 179–194.

Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., Krieger-Redwood, K., Bernhardt, B. C., Smallwood, J., & Jefferies, E. (2016). Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. *NeuroImage*, *137*, 165–177.

Davis, B., & Hasson, U. (2018). Predictability of what or where reduces brain activity, but a bottleneck occurs when both are predictable. *NeuroImage*, *167*, 224–236.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical Processing in Spoken Language Comprehension. *The Journal of Neuroscience*, *23*(8), 3423–3431.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, *37*(27), 6539–6557.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Concerns with Nieuwland et al. (2017). http://www.kutaslab.ucsd.edu/FinalDUK17Comment9LabStudy.pdf

Demirtaş, M., Burt, J. B., Helmer, M., Ji, J. L., Adkinson, B. D., Glasser, M. F., Van Essen, D. C., Sotiropoulos, S. N., Anticevic, A., & Murray, J. D. (2019). Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics. *Neuron*, *101*(6), 1181-1194.e13.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505–E514.

Dietrich, S., Hertrich, I., Müller-Dahlhaus, F., Ackermann, H., Belardinelli, P., Desideri, D., Seibold, V. C., & Ziemann, U. (2018). Reduced performance during a sentence repetition task by continuous theta-burst magnetic stimulation of the pre-supplementary motor area. *Frontiers in Neuroscience*, *12*, 1–13.

Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2008). Time travel through language: Temporal shifts rapidly decrease information accessibility during reading. *Psychonomic Bulletin & Review*, *15*(4), 750–756.

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e9.

Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in Cognitive Sciences*, *12*(3), 99–105.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483.

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*(8), 2530–2541.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335.

Ekstrom, A. (2010). How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation. *Brain Research Reviews*, *62*(2), 233–244.

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia*, *50*(9), 2154–2164.

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The Brain Dynamics of Rapid Perceptual Adaptation to Adverse Listening Conditions. *The Journal of Neuroscience*, *33*(26), 10688–10697.

Erb, J., & Obleser, J. (2013). Upregulation of cognitive control networks in older adults' speech comprehension. *Frontiers in Systems Neuroscience*, *7*, 1–13.

Erb, J., Schmitt, L.-M., & Obleser, J. (2020). Temporal selectivity declines in the aging human auditory cortex. *ELife*, *9*, e55300.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203–12208.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495.

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Miller, K. L., Ugurbil, K., & Yacoub, E. (2010). Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging. *PLoS ONE*, *5*(12), e15710.

Fiebach, C. J., & Friederici, A. D. (2004). Processing concrete words: FMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia*, *42*(1), 62–70.

Fitzhugh, M. C., Hemesath, A., Schaefer, S. Y., & Baxter, L. C. (2019). Functional Connectivity of Heschl's Gyrus Associated With Age-Related Hearing Loss: A Resting-State fMRI Study. *Frontiers in Psychology*, *10*, 2485.

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102.

Forster, K. I. (1981). Priming and the Effects of Sentence and Lexical Contexts on Naming Time: Evidence for Autonomous Lexical Processing. *The Quarterly Journal of Experimental Psychology*

Frank, S. L., & Thompson, R. L. (2012). Early effects of word surprisal on pupil size during reading. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, *Japan*, *34*, 1554–1559.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203.

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254.

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and Modulatory Interactions in Neuroimaging. *NeuroImage*, *6*(3), 218–229.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1–3), 70–87.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234.

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neuroscience & Biobehavioral Reviews*, *118*, 42–64.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220.

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6), 1–19.

Garrido, M. I., Rowe, E. G., Halász, V., & Mattingley, J. B. (2018). Bayesian Mapping Reveals That Attention Boosts Neural Responses to Predicted and Unpredicted Stimuli. *Cerebral Cortex*, *28*(5), 1771–1782.

Geranmayeh, F., Brownsett, S. L. E., & Wise, R. J. S. (2014). Task-induced brain activity in aphasic stroke patients: What is driving recovery? *Brain*, *137*(10), 2632–2648.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. *Neuroinformatics*, *7*(1), 73–82.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*, 171–178.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Flinker, A., Devore, S., Doyle, W., Friedman, D., … Hasson, U. (2020). Thinking ahead: Prediction in context as a keystone of language in humans and machines. *bioRxiv*. https://doi.org/10.1101/2020.12.02.403477

Golestani, N., Hervais-Adelman, A., Obleser, J., & Scott, S. K. (2013). Semantic versus perceptual interactions in neural processing of speech-in-noise. *NeuroImage*, *79*, 52–61.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, *21*(2), 215–223.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, *USA*, 10–18.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*, 13.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72.

Guediche, S., Blumstein, S. E., Fiez, J. A., & Holt, L. L. (2014). Speech perception under adverse conditions: Insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, *7*, 126.

Guediche, S., Reilly, M., Santiago, C., Laurent, P., & Blumstein, S. E. (2016). An fMRI study investigating effects of conceptually related sentences on the perception of degraded speech. *Cortex*, *79*, 57–74.

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, *26*(6), 2919–2934.

Gwilliams, L., & Davis, M. (in press). Extracting language content from speech sounds: An information theoretic approach. In L.L. Holt, J.E. Peelle, A.B. Coffin, A.N. Popper, & R.R. Fay (Eds.), *Speech Perception. Springer Handbook of Auditory Research* (pp. 113–139). Springer.

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*. https://doi.org/10.1101/2020.04.04.025684

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *The Journal of Neuroscience*, *38*(35), 7585–7599.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 819–836). MIT Press.

Halai, A. D., Welbourne, S. R., Embleton, K., & Parkes, L. M. (2014). A comparison of dual gradient-echo and spin-echo fMRI of the inferior temporal lobe. *Human Brain Mapping*, *35*(8), 4118–4128.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, USA*, 159–166.

Hale, J. (2016). Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M., & Bowtell, R. W. (1999). „Sparse" Temporal Sampling in Auditory fMRI. *Human Brain Mapping*, *7*(3), 213–223.

Hallam, G. P., Whitney, C., Hymers, M., Gouws, A. D., & Jefferies, E. (2016). Charting the effects of TMS with fMRI: Modulation of cortical recruitment within the distributed network supporting semantic control. *Neuropsychologia*, *93*, 40–52.

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics*, *7*(1), 37–53.

Hartwigsen, G. (2018). Flexible Redistribution in Cognitive Networks. *Trends in Cognitive Sciences*, *22*(8), 687–698.

Hartwigsen, G., Golombek, T., & Obleser, J. (2015). Repetitive transcranial magnetic stimulation over left angular gyrus modulates the predictability gain in degraded speech comprehension. *Cortex*, *68*, 100–110.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. *The Journal of Neuroscience*, *28*(10), 2539–2550.

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313.

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

Heinzerling, B., & Strube, M. (2017). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. *arXiv:1710.02187 [Cs.CL]*.

Henson, R. N. (2015). Analysis of Variance (ANOVA). *In A. W. Toga (Ed.), Brain Mapping: An Encyclopedic Reference* (pp. 477–481). Elsevier.

Hermes, D., Nguyen, M., & Winawer, J. (2017). Neuronal synchrony and the relation between the blood-oxygen-level dependent response and the local field potential. *PLOS Biology*, *15*(7), e2001461.

Hertrich, I., Dietrich, S., & Ackermann, H. (2016). The role of the supplementary motor area for speech and language processing. *Neuroscience and Biobehavioral Reviews*, *68*, 602–610.

Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7–8), 1145–1166.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1–2), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing: An analysis using event-related brain potentials. *Psychobiology*, *19*(4), 286–300.

Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *International Journal of Audiology*, *57*(7), 483–492.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, *76*(2), 423–434.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679.

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fMRI study. *Neuroreport*, *12*(8), 1749–1752.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*), 453–458.

Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2016). The Semantic Network at Work and Rest: Differential Connectivity of Anterior Temporal Lobe Subregions. *The Journal of Neuroscience*, *36*(5), 1490–1501.

Jain, S., & Huth, A. (2018). Incorporating Context into Language Encoding Models for fMRI. b*ioRxiv*. https://doi.org/10.1101/327601

Jain, S., Vo, V. A., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *bioRxiv*. https://doi.org/10.1101/2020.10.02.324392

Jefferies, E. (2013). The neural basis of semantic cognition: Converging evidence from neuropsychology , neuroimaging and TMS. *Cortex*, *49*(3), 611–625.

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford University Press.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841.

Jeon, H.-A., & Friederici, A. D. (2015). Degree of automaticity and the prefrontal cortex. *Trends in Cognitive Sciences*, *19*(5), 7.

Johnson, A., & Redish, A. D. (2007). Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *The Journal of Neuroscience*, *27*(45), 12176–12189.

Jung, J. Y., Cloutman, L. L., Binney, R. J., & Lambon Ralph, M. A. (2017). The structural connectivity of higher order association cortices reflects human functional brain networks. *Cortex*, *97*, 221–239.

Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science*, *20*(2), 137–194.

Jurafsky, D., & Martin, J. H. (in preparation). *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Kádár, Á., Côté, M.-A., Chrupała, G., & Alishahi, A. (2018). Revisiting the Hierarchical Multiscale LSTM. a*rXiv:1807.03595 [cs.CL]*.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351.

Kandylaki, K. D., Nagels, A., Tune, S., Kircher, T., Wiese, R., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2016). Predicting „When" in Discourse Engages the Human Dorsal Auditory Stream: An fMRI Study Using Naturalistic Stories. *The Journal of Neuroscience*, *36*(48), 12180–12191.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Keller, T. A. (2001). The Neural Bases of Sentence Comprehension: A fMRI Examination of Syntactic and Lexical Processing. *Cerebral Cortex*, *11*(3), 223–237.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, *4*(11), e1000209.

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. a*rXiv:1412.6980 [cs.LG]*.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347.

Kleiman, G. M. (1980). Sentence frame contexts and lexical decisions: Sentence-acceptability and word-relatedness effects. *Memory & Cognition*, *8*(4), 336–344.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, *302*(5648), 1181–1185.

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270.

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America*, *83*(5), 1852–1862.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.

Kumar, M., Ellis, C. T., Lu, Q., Zhang, H., Capotă, M., Willke, T. L., Ramadge, P. J., Turk-Browne, N. B., & Norman, K. A. (2020). BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis. *PLOS Computational Biology*, *16*(1), e1007549.

Kumar, S., Kaposvari, P., & Vogels, R. (2017). Encoding of Predictable and Unpredictable Stimuli by Inferior Temporal Cortical Neurons. *Journal of Cognitive Neuroscience*, *29*(8), 1445–1454.

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

Kurby, C. A., & Zacks, J. M. (2012). Starting from scratch and building brick by brick in comprehension. *Memory & Cognition*, *40*(5), 812–826.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

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*(1), 621–647.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Kwisthout, J., Bekkering, H., & van Rooij, I. (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, *112*, 84–91.

La Camera, G., Rauch, A., Thurbon, D., Lüscher, H.-R., Senn, W., & Fusi, S. (2006). Multiple Time Scales of Temporal Response in Pyramidal and Fast Spiking Cortical Neurons. *Journal of Neurophysiology*, *96*(6), 3448–3464.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *Journal of Neurophysiology*, *94*(3), 1904–1911.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *The Journal of Neuroscience*, *31*(8), 2906–2915.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, D. A., Bayley, P. J., & Squire, L. R. (2004). The anatomy of semantic knowledge: Medial vs. Lateral temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6710–6715.

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., & Friston, K. J. (2013). A Neurocomputational Model of the Mismatch Negativity. *PLoS Computational Biology*, *9*(11), e1003288.

Lindquist, M. A., Geuter, S., Wager, T. D., & Caffo, B. S. (2019). Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *Human Brain Mapping*, *40*(8), 2358–2376.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869–878.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD Signal. *Annual Review of Physiology*, *66*(1), 735–769.

Lohmann, G., Stelzer, J., Lacosse, E., Kumar, V. J., Mueller, K., Kuehn, E., Grodd, W., & Scheffler, K. (2018). LISA improves statistical analysis for fMRI. *Nature Communications*, *9*(1), 1–9.

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, *42*(S4), 1166–1183.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. a*rXiv:2005.02181 [cs.AI]*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *The Journal of Neuroscience Methods*, *164*(1), 177–190.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Marreiros, A. C., Kiebel, S. J., & Friston, K. J. (2008). Dynamic causal modelling for fMRI: A two-state model. *NeuroImage*, *39*(1), 269–278.

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, *58*(1), 25–45.

Mattar, M. G., Kahn, D. A., Thompson-Schill, S. L., & Aguirre, G. K. (2016). Varying Timescales of Stimulus Integration Unite Neural Adaptation and Prototype Formation. *Current Biology*, *26*(13), 1669–1676.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978.

Mattys, S. L., & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. *Perception & Psychophysics*, *70*(7), 1235–1242.

McCallum, W. C., Farmer, S. F., & Pocock, P. V. (1984). The effects of physical and semantic incongruities on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology*, *59*(6), 477–488.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron*, *71*(5), 926–940.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia*, *50*(5), 762–776.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567.

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. a*rXiv:1301.3781 [cs.CL].*

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

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., & Malach, R. (2005). Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. *Science*, *309*(5736), 951–954.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, *17*(12), 1661–1663.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667–685.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 38.

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: Evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of Cognitive Neuroscience*, *25*(11), 1824–1850.

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin*, *143*(9), 992–1009.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Obleser, J. (2014). Putting the Listening Brain in Context: Listening in Context. *Language and Linguistics Compass*, *8*(12), 646–658.

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, *20*(3), 633–640.

Obleser, J., Wise, R. J. S., Dresner, A. M., & Scott, S. K. (2007). Functional Integration across Brain Regions Improves Speech Perception under Adverse Listening Conditions. *The Journal of Neuroscience*, *27*(9), 2283–2289.

Oh, A., Duerden, E. G., & Pang, E. W. (2014). The role of the insula in speech and language processing. *Brain and Language*, *135*, 96–103.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*, 71–133.

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*(4), 426–448.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peelle, J. E. (2014). Methodological challenges and solutions in auditory functional magnetic resonance imaging. *Frontiers in Neuroscience*, *8*, 1–13.

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204–214.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, *Qatar*, 1532–1543.

Penny, W., Friston, K., Ashburner, J., Kiebel, S., & Nichols, T. (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier.

Penny, W., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, *6*(3), e1000709.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, USA*, 2227–2237.

Petkov, C. I., Kayser, C., Augath, M., & Logothetis, N. K. (2009). Optimizing the imaging of the monkey auditory cortex: Sparse vs. continuous fMRI. *Magnetic Resonance Imaging*, *27*(8), 1065–1073.

Phillips, H. N., Blenkmann, A., Hughes, L. E., Kochen, S., Bekinschtein, T. A., Cam-CAN, & Rowe, J. B. (2016). Convergent evidence for hierarchical prediction networks from human electrocorticography and magnetoencephalography. *Cortex*, *82*, 192–205.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Powell, H. W. R., Guye, M., Parker, G. J. M., Symms, M. R., Boulby, P., Koepp, M. J., Barker, G. J., & Duncan, J. (2004). Noninvasive in vivo demonstration of the connections of the human parahippocampal gyrus. *NeuroImage*, *22*(2), 740–747.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional Network Organization of the Human Brain. *Neuron*, *72*(4), 665–678.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341.

Pruim, R. H. R., Mennes, M., Buitelaar, J. K., & Beckmann, C. F. (2015). Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *NeuroImage*, *112*, 278–287.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*. https://openai.com/blog/better-language-models/

Radvansky, G. A. (2012). Across the Event Horizon. *Current Directions in Psychological Science*, *21*(4), 269–272.

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718.

Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6*, 1–16.

Regev, M., Simony, E., Lee, K., Tan, K. M., Chen, J., & Hasson, U. (2018). Propagation of Information Along the Cortical Hierarchy as a Function of Attention While Reading and Listening to Stories. *Cerebral Cortex*, *29*(10), 4017–4034.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Richmond, L. L., & Zacks, J. M. (2017). Constructing Experience: Event Models from Perception to Action. *Trends in Cognitive Sciences*, *21*(12), 962–980.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The Neural Mechanisms of Speech Comprehension: FMRI studies of Semantic Ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269.

Rogalsky, C., Matchin, W., & Hickok, G. (2008). Broca's area, sentence comprehension, and working memory: An fMRI study. *Frontiers in Human Neuroscience*, *2*, 1–14.

Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, *118*(3), 339–358.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A New Statistic. *Psychological Science*, *5*(6), 329–334.

Runyan, C. A., Piasini, E., Panzeri, S., & Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature*, *548*(7665), 92–96.

Rysop, A. U., Schmitt, L.-M., Obleser, J., & Hartwigsen, G. (2021). Neural modelling of the semantic predictability gain under challenging listening conditions. *Human Brain Mapping*, *42*(1), 110–127.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences*, *114*(18), 4799–4804.

Sainburg, T., Theilman, B., Thielk, M., & Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nature Communications*, *10*(1), 3636.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67.

Scharinger, M., Bendixen, A., Herrmann, B., Henry, M. J., Mildner, T., & Obleser, J. (2016). Predictions interact with missing sensory evidence in semantic processing areas: Predictions and Missing Sensory Evidence. *Human Brain Mapping*, *37*(2), 704–716.

Schild, H. H. (1990). *MRI made easy (...well almost)*. Schering.

Schmälzle, R., Häcker, F. E. K., Honey, C. J., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*, *10*(8), 1137–1143.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schmidt, C. F., Zaehle, T., Meyer, M., Geiser, E., Boesiger, P., & Jancke, L. (2008). Silent and Continuous fMRI Scanning Differentially Modulate Activation in an Auditory Language Comprehension Task. *Human Brain Mapping*, 29, 46–56.

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., & Obleser, J. (2019). Reading times and temporo-parietal BOLD activity encode the semantic hierarchy of language prediction. *2019 Conference on Cognitive Computational Neuroscience, Germany*. https://doi.org/10.32470/CCN.2019.1333-0

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., & Obleser, J. (2020). Predicting speech from a cortical hierarchy of event-based timescales. *bioRxiv*. https://doi.org/10.1101/2020.12.19.423616

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5–35.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N. G., Tenenbaum, J. B., & Fedorenko, E. (2020). Artificial Neural Networks Accurately Predict Language Processing in the Brain. *bioRxiv*. https://doi.org/10.1101/2020.06.26.174482

Schuberth, R. E., & Eimas, P. D. (1977). Effects of Context on the Classification of Words and Nonwords. *Human Perception and Performance*, *3*(1), 27–36.

Schuberth, R. E., Spoehr, K. T., & Lane, D. M. (1981). Effects of stimulus and contextual information on the lexical decision process. *Memory & Cognition*, *9*(1), 68–77.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Schwiedrzik, C. M., & Freiwald, W. A. (2017). High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron*, *96*(1), 89-97.e4.

Seghier, M. L. (2013). The Angular Gyrus. *The Neuroscientist*, *19*(1), 43–61.

Seghier, M. L., Fagan, E., & Price, C. J. (2010). Functional Subdivisions in the Left Angular Gyrus Where the Semantic System Meets and Diverges from the Default Network. *The Journal of Neuroscience*, *30*(50), 16809–16817.

Sestieri, C., Shulman, G. L., & Corbetta, M. (2017). The contribution of the human posterior parietal cortex to episodic memory. *Nature Reviews Neuroscience*, *18*(3), 183–192.

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

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, *30*, 50–64.

Shannon, R., Fu, Q.-J., & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Oto-Laryngologica Supplementum*, *522,* 50–54.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304.

Sheng, J., Zheng, L., Lyu, B., Cen, Z., Qin, L., Tan, L. H., Huang, M.-X., Ding, N., & Gao, J.-H. (2018). The Cortical Maps of Hierarchical Linguistic Structures during Speech Perception. *Cerebral Cortex*, *29*(8), 3232–3240.

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, *35*(5), 1981–1996.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., … Vianello, M. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Smith, N. J., & Levy, R. (2008). Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society, USA, 30.* https://escholarship.org/uc/item/3mr8m3rf

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101.

Spiers, H. J., Hayman, R. M. A., Jovalekic, A., Marozzi, E., & Jeffery, K. J. (2015). Place Field Repetition and Purely Local Remapping in a Multicompartment Environment. *Cerebral Cortex*, *25*(1), 10–25.

Sporns, O., & Betzel, R. F. (2016). Modular Brain Networks. *Annual Review of Psychology*, *67*(1), 613–640.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653.

Stephan, K. E., Penny, W., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Stephens, G. J., Honey, C. J., & Hasson, U. (2013). A place for time: The spatiotemporal structure of neural dynamics during natural audition. *Journal of Neurophysiology*, *110*(9), 2019–2026.

Strange, B. A., Otten, L. J., Josephs, O., Rugg, M. D., & Dolan, R. J. (2002). Dissociable Human Perirhinal, Hippocampal, and Parahippocampal Roles during Verbal Encoding. *The Journal of Neuroscience*, *22*(2), 523–528.

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Bulletin*, *30*(4), 415–433.

Teipel, S. J., Bokde, A. L. W., Meindl, T., Amaro, E., Soldner, J., Reiser, M. F., Herpertz, S. C., Möller, H.-J., & Hampel, H. (2010). White matter microstructure underlying default mode network connectivity in the human brain. *NeuroImage*, *49*(3), 2021–2032.

Todorovic, A., & de Lange, F. P. (2012). Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. *The Journal of Neuroscience*, *32*(39), 13389–13395.

Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, *66*(4), 319–327.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320.

Uddin, L. Q., Supekar, K., Amin, H., Rykhlevskaia, E., Nguyen, D. A., Greicius, M. D., & Menon, V. (2010). Dissociable Connectivity within Human Angular Gyrus and Intraparietal Sulcus: Evidence from Functional and Structural Connectivity. *Cerebral Cortex*, *20*(11), 2636–2646.

Vaden, K. I., Kuchinsky, S. E., Ahlstrom, J., Dubno, J. R., & Eckert, M. A. (2015). Cortical Activity Predicts Which Older Adults Recognize Speech in Noise and When. *The Journal of Neuroscience*, *35*(9), 3929–3937.

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J., Dubno, J. R., & Eckert, M. A. (2013). The Cingulo-Opercular Network Provides Word-Recognition Benefit. *The Journal of Neuroscience*, *33*(48), 18979–18986.

Vaden, K. I., Teubner-Rhodes, S., Ahlstrom, J., Dubno, J. R., & Eckert, M. A. (2017). Cingulo-opercular activity affects incidental memory encoding for speech in noise. *NeuroImage*, *157*, 381–387.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394–417.

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

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179.

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2010). Semantic Processing in the Anterior Temporal Lobes: A Meta-analysis of the Functional Neuroimaging Literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094.

Vlaardingerbroek, M. T., & den Boer, J. A. (2003). *Magnetic Resonance Imaging. Theory and Practice*. Springer.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 20754–20759.

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *ELife*, *7*, e39061.

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166.

Whitney, C., Huber, W., Klann, J., Weis, S., Krach, S., & Kircher, T. (2009). Neural correlates of narrative shifts during auditory story comprehension. *NeuroImage*, *47*(1), 360–366.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Wilson, S. M., Molnar-Szakacs, I., & Iacoboni, M. (2008). Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension. *Cerebral Cortex*, *18*(1), 230–242.

Winn, M. (2016). Rapid Release from Listening Effort Resulting from Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, *20*, 2331216516669723.

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2017). How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cerebral Cortex*, *27*(10), 4988–5000.

Zempleni, M.-Z., Renken, R., Hoeks, J. C. J., Hoogduin, J. M., & Stowe, L. A. (2007). Semantic ambiguity processing in sentence context: Evidence from event-related fMRI. *NeuroImage*, *34*(3), 1270–1279.

Zhang, W., & Yartsev, M. M. (2019). Correlated Neural Activity across the Brains of Socially Interacting Bats. *Cell*, *178*(2), 413–428.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE* Transactions on Medical Imaging, 20(1), 45–57.

List of figures

List of tables

List of abbreviations

A	
AAC	auditory association cortex
AG	angular gyrus
aIns	anterior insula
B	
BOLD	blood-oxygen-level dependent
C	
CO	cingulo-opercular network
D	
DCM	dynamic causal modelling
E	
EAC	early auditory cortex
EVC	early visual cortex
F	
FDR	false discovery rate
fMRI	functional magnetic resonance imaging
FP	frontal pole
FWE	family-wise error rate
G	
GLM	general linear model
H	
HM-LSTM	hierarchical multiscale LSTM
HRF	hemodynamic response function
I	
IFG	inferior frontal gyrus
IPC	inferior parietal cortex
L	
LSTM	long short-term memory
LTC	lateral temporal cortex
M	
MC	motor cortex
MCC	middle cingulate cortex
MOG	middle occipital gyrus
P	
PCC	posterior cingulate cortex
PCG	paracingulate gyrus
PGa	anterior angular gyrus
PGp	posterior angular gyrus
PHG	parahippocampal gyrus
pMTG	posterior middle temporal gyrus
PPI	psychophysiological interaction
prec	precuneus
pre-SMA	pre-supplementary motor area
R	
RNN	recurrent neural network
ROI	region of interest

S	
SCC	subcallosal cortex
SFG	superior frontal gyrus
SMG	supramarginal gyrus
SNR	signal-to-noise ratio
SPC	superior parietal cortex
SPIN	speech in noise
SRT	speech reception threshold
SSC	somatosensory cortex
STG	superior temporal gyrus
T	
TMS	transcranial magnetic stimulation
TPOJ	temporo-parieto-occipital junction
V	
V1	primary auditory cortex

Summary

Introduction

Language without context is meaningless. In fact, context is often so meaningful that we can use it to predict upcoming words. When context is more constraining, speed and accuracy of responses to language increase (Tulving & Gold, 1963; Schuberth et al., 1981). At the same time, it does not harm performance when a prediction is violated (Frisson et al., 2017). This suggests that predictive processing is a beneficial strategy in comprehension. While we resort to such predictions especially when language is not clearly understood (Kalikow et al., 1977), predictive processing is thought to be a fundamental strategy that we adopt naturally and constantly (Kuperberg & Jaeger, 2016). However, it remains unknown, which functions are fulfilled by single brain regions and how these brain regions interact with one another to enable predictive speech comprehension.

Two networks implicated in speech processing are the domain-general cingulo-opercular and the domain-specific language network. The domain-specific network is recruited when there are semantic cues available to inform the prediction of upcoming words, while the domain-general network is recruited under adverse listening conditions when no semantic cues are available. In the first study, we asked how interactions between domain-general and domain-specific networks adapt to the demands posed by acoustic intelligibility and semantic predictability of speech. More specifically, we aimed to characterize those network interactions that foster successful speech comprehension in an fMRI sentence-repetition experiment.

In natural speech, there is rich context available, which is potentially relevant to making accurate predictions. When listening to speech, context is represented along a temporo-parietal hierarchy operating on increasingly long timescales, with inferior parietal regions coding for long timescales like paragraphs in comparison to short timescales like words (Lerner et al., 2011). Theories of predictive coding propose that predictions are formed along a hierarchy, where predictions are fed back from higher to lower timescales, whereas the prediction error is fed forward (Rao & Ballard, 1999; Friston, 2005). We performed an fMRI natural listening experiment to bridge the gap between the hierarchical organization of context during speech comprehension and the hierarchical organization of predictive processing in general.

A major challenge in using natural speech to study predictive processing is to determine the predictability for each word in the story. In recent years, artificial neural networks are increasingly used to model processing in the brain (Cichy & Kaiser, 2019). An artificial architecture designed to retain context in memory for making predictions on upcoming events is the long short-term memory cell (LSTM; Hochreiter & Schmidhuber, 1997), which operates on different levels of abstraction when stacked. We derived word predictability from two LSTMs that differed with respect to one computational mechanism (i.e., context updating rule), which allows to study the role of this specific computation for predictive processing in the brain.

The overarching goal of this thesis was to shed light on the biological neural network dynamics that underlie the prediction of speech when poor acoustics challenge comprehension. The present thesis set out to answer three questions: (1) How do intelligibility and predictability shape the interplay between domain-general and domain-specific networks in speech comprehension? (2) How can we model predictive processing in the brain with artificial neural networks? (3) How does the brain make predictions on the next word when confronted with the multitude of timescales in natural speech?

Results

In the first study, we recorded fMRI while participants performed a repetition task on sentences with varying semantic predictability and acoustic intelligibility. With better intelligibility, activation in the language network increased for sentences of high predictability, whereas activation in the cingulo-opercular network increased for sentences of low predictability. This speaks for the adaptive recruitment of the cingulo-opercular network when no domain-specific semantic cues are available to the language network. The behavioural predictability gain at intermediate intelligibility was associated with stronger activation of the language network and concurrent inhibition of the cingulo-opercular network.

In the next study, we dissociated from highly controlled stimuli and recorded fMRI in a natural listening task. Our central hypothesis was that context is resolved into its timescales to inform predictions on upcoming words at multiple levels of abstraction, with a timescale hierarchy evolving along temporo-parietal cortex. We operationalized timescale-specific word predictability as the similarity between the semantic vector representation of a word and the average vector across context of specific length. However, we found no temporo-parietal hierarchy for the timescales of similarity, possibly because the simple nature of the underlying generative model is reflective of lexical association.

Under the assumption that we found no hierarchy because of similarity being too simple of an internal model for word prediction, we developed a new metric. We trained two artificial neural networks with five timescales to predict the next word in a story by its preceding context and read out timescale-specific word surprisal. Critically, timescale-specific context representations were updated either with each new incoming word (LSTM) or just at the boundary of events (HM-LSTM; Chung et al., 2016).

In a self-paced reading task, higher model-based surprisal was associated with longer reading times, thereby confirming the validity of this metric. In fMRI, event-based surprisal evolved along a temporo-parietal hierarchy, with coarser events represented in more parietal regions. Along the event-based hierarchy, surprisal gated bottom-up and top-down connectivity to temporal receptive windows of neighbouring timescales. In contrast, we found no hierarchy for continuously updated surprisal. This suggests that the sparse updates inherent to the event-based representation of context pose an efficient coding scheme for predictive processes in speech comprehension.

Discussion

Our findings demonstrate that (1) successful comprehension of predictable speech in adverse conditions relies on the activation of the language network but inhibition within the cingulo-opercular network, (2) semantic similarity is not an appropriate measure of hierarchical predictive speech processing in the brain, (3) artificial neural networks capture behavioural and neural signatures of predictive processing, (4) event-based timescale surprisal is represented along a temporo-parietal hierarchy.

An important implication of our results is that the temporo-parietal surprisal hierarchy observed in the natural listening task can be attributed to event-based context updates, which give rise to event segmentation, hierarchical working memory and hierarchical predictive language processing. This highlights that using artificial neural networks for model comparison is a viable account to adjudicate between different computational principles implemented in the human brain.

Together, speech prediction under challenging listening conditions recruited a broad language network facilitating comprehension. This network included temporo-parietal regions implicated in event-based predictive processing as well as IFG typically implicated in higher-order control processes and parahippocampal regions typically implicated in memory processes. Additionally, the cingulo-opercular network acted as a compensatory executive control function when no semantic cues for predictive processing were available.

Finally, I put forward that the dynamic interplay between biological neural networks of working memory, long-term memory and cognitive control shapes language prediction. The framework holds that an event-based dorsal and time-invariant ventral pathway in auditory regions simultaneously segment incoming speech into hierarchical timescales. These working memory representations might subserve the formation of predictions at multiple timescales, with longer timescales informing predictions at shorter timescales and bottom-up prediction errors initiating updates to predictions. Predictive processing at longer timescales might interface with long-term memories via hippocampus. Finally, IFG might exert cognitive control over longer timescales in auditory regions, with ventral IFG providing complex internal models for prediction and dorsal IFG supporting predictions when context is ambiguous. This framework suggests that hierarchical predictive processing extends beyond the auditory dorsal pathway, engaging also cognitive control and long-term memory.

Zusammenfassung

Einleitung

Sprache ist ohne Kontext bedeutungslos. Der Kontext eines Wortes ist sogar oftmals so bedeutsam, dass wir ihn nutzen können, um nachfolgende Wörter vorherzusagen. Allgemein erlaubt es uns ein informativer Kontext, schneller und genauer auf Sprache zu reagieren (Tulving & Gold, 1963; Schuberth et al., 1981). Gleichzeitig schadet es der Leistung aber auch nicht, wenn sich eine Vorhersage nicht erfüllt (Frisson et al., 2017). Dies legt nahe, dass die prädiktive Verarbeitung vorteilhaft für das Sprachverstehen ist. Während wir vor allem dann auf Vorhersagen zurückgreifen, wenn Sprache akustisch schwer verständlich ist (Kalikow et al., 1977), gilt die Prädiktion als grundlegende Verarbeitungsstrategie, die wir automatisch regelmäßig anwenden (Kuperberg & Jaeger, 2016). Unklar bleibt jedoch, welche Funktionen einzelne Hirnregionen erfüllen und wie diese Hirnregionen miteinander interagieren, um prädiktives Sprachverstehen zu ermöglichen.

An der Sprachverarbeitung sind das domänenübergreifende cingulo-operculäre und das domänenspezifische Sprachnetzwerk beteiligt. Das Sprachnetzwerk wird rekrutiert, wenn Kontext verfügbar ist, der die Vorhersage nachfolgender Wörter erlaubt. Wenn kein informativer Kontext verfügbar ist, wird unter ungünstigen Hörbedingungen hingegen das cingulo-operculäre Netzwerk rekrutiert. In einer ersten Studie untersuchten wir in einem fMRT-Experiment, wie sich das Zusammenspiel von domänenübergreifenden und -spezifischen Netzwerken an die akustische Verständlichkeit und semantische Vorhersagbarkeit von Sprache anpasst, während Probanden gesprochene Sätze wiederholten. Dabei war es unser Ziel, solche Netzwerkinteraktionen zu charakterisieren, die ein erfolgreiches Sprachverstehen ermöglichen.

Natürliche Sprache ist reich an Kontext, der für die Vorhersage von Sprache genutzt werden kann. Beim Hören von Sprache wird der Kontext entlang einer temporo-parietalen Hierarchie repräsentiert. Diese Hierarchie arbeitet auf zunehmend längeren Zeitskalen, wobei der inferiore Parietallappen Sprache auf langen Zeitskalen wie Absätzen verarbeitet (Lerner et al., 2011). Theorien zur prädiktiven Kodierung schlagen vor, dass auch Vorhersagen entlang einer solchen Hierarchie getroffen werden: Dabei sollen Vorhersagen von längeren zu kürzeren Zeitskalen rückgekoppelt werden, während der Vorhersagefehler andersherum von kürzeren zu längeren Zeitskalen weitergeleitet wird (Rao & Ballard, 1999; Friston, 2005). In einer zweiten Studie führten wir ein fMRT-Experiment durch, in dem Probanden einer Geschichte zuhörten, um die empirische Lücke zwischen der hierarchischen Organisation von Kontext während des Sprachverstehens und der hierarchischen Organisation von prädiktiver Kodierung im Allgemeinen zu schließen.

Eine Herausforderung bei der Verwendung von natürlicher Sprache als Stimulusmaterial ist es, die Vorhersagbarkeit für jedes einzelne Wort zu bestimmen. In den letzten Jahren werden zunehmend künstliche neuronale Netze verwendet, um die Verarbeitung im Gehirn zu modellieren (Cichy & Kaiser, 2019). Eine künstliche Netzwerkarchitektur, die Kontext auf verschiedenen Abstraktionsebenen im Gedächtnis behält, um diesen dann für die Vorhersage von Ereignissen zu nutzen, ist das *Long short-term memory* (LSTM; Hochreiter & Schmidhuber, 1997). Wir bestimmten Wortvorhersagen für die im Experiment präsentierte Geschichte mithilfe zweier LSTMs, die sich in Bezug auf den Mechanismus unterscheiden, der Repräsentationen von Kontext aktualisiert. Ein solcher Ansatz erlaubt es, die Rolle dieses Mechanismus für die Vorhersageprozesse im Gehirn zu untersuchen.

Das übergeordnete Ziel dieser Arbeit war es, die Dynamiken biologischer neuronaler Netzwerke zu beleuchten, die der Vorhersage von Sprache zugrunde liegen, wenn schlechte Akustik das Sprachverstehen erschwert. Dabei stellten sich drei Fragen: (1) Wie beeinflussen Verständlichkeit und Vorhersagbarkeit das Zusammenspiel zwischen domänenübergreifenden und domänenspezifischen Netzwerken beim Sprachverstehen? (2) Wie können wir die prädiktive Verarbeitung im Gehirn mit künstlichen neuronalen Netzen modellieren? (3) Wie macht das Gehirn Vorhersagen über das nächste Wort, wenn wir in der natürlichen Sprache mit einer Vielzahl von Zeitskalen konfrontiert sind?

Ergebnisse

In einer fMRT-Studie wiederholten Probanden gesprochene Sätze mit unterschiedlicher semantischer Vorhersagbarkeit und akustischer Verständlichkeit. Bei besserer Verständlichkeit stieg die Aktivierung im Sprachnetzwerk für Sätze mit hoher Vorhersagbarkeit, während die Aktivierung im cingulo-operculären Netzwerk bei Sätzen mit geringer Vorhersagbarkeit zunahm. Dies spricht für die adaptive Rekrutierung des cingulo-operculären Netzwerks, wenn dem Sprachnetzwerk keine domänenspezifischen semantischen Hinweise zur Verfügung stehen. Ein durch die Vorhersagbarkeit vermittelter Anstieg im Sprachverstehen bei mittlerer Verständlichkeit war mit einer stärkeren Aktivierung des Sprachnetzwerks und gleichzeitiger Hemmung des cingulo-operculären Netzwerks verbunden.

In der zweiten fMRT-Studie hörten Probanden eine Geschichte. Unsere Hypothese war, dass Kontext in seine Zeitskalen zerlegt wird, um auf verschiedenen Abstraktionsebenen Vorhersagen zu treffen. Dabei erwarteten wir, dass sich eine Hierarchie von Zeitskalen entlang des temporo-parietalen Kortex herausbildet. Wir operationalisierten die zeitskalenspezifische Vorhersagbarkeit eines Wortes als die Ähnlichkeit zwischen der Vektorrepräsentation eines Wortes und dem mittleren Vektor eines vorangegangenen Kontextausschnitts bestimmter Länge. Diese Zeitskalen der semantischen Ähnlichkeit organisierten sich allerdings nicht entlang einer temporo-parietalen Hierarchie, möglicherweise weil die semantische Ähnlichkeit weniger Vorhersage als lexikalische Assoziation widerspiegelt.

Unter der Annahme, dass die semantische Ähnlichkeit kein hinreichendes Modell für die Wortvorhersage ist, entwickelten wir einen neuen Ansatz. Wir trainierten zwei künstliche neuronale Netze mit jeweils fünf Zeitskalen darauf, das nächste Wort in einer Geschichte anhand des vorangegangenen Kontexts vorherzusagen. Dabei wurden die zeitskalenspezifischen Kontextrepräsentationen entweder mit jedem gesprochenen Wort (LSTM) oder nur am Ende von Ereignissen aktualisiert (HM-LSTM; Chung et al., 2016). Aus den Wortvorhersagen leiteten wir den *Surprisal* (oder Vorhersagefehler) ab.

In einer selbstgesteuerten Leseaufgabe war ein größerer *Surprisal* mit längeren Lesezeiten verbunden, was die Validität dieser Metrik bestätigt. Im fMRT zeigte sich eine ereignisbasierte temporo-parietale *Surprisal*-Hierarchie, wobei weiter parietal gelegene Regionen eine stärkere Aktivierung zeigten, wenn längere Ereignisse weniger informativ für die Vorhersage waren. Entlang dieser Hierarchie regulierte *Surprisal* die Bottom-up- und Top-down-Konnektivität zwischen Hirnregionen, die Vorhersagen auf benachbarten Zeitskalen verarbeiten. Im Gegensatz dazu fanden wir keine Hierarchie für Vorhersagen, die auf kontinuierlich aktualisierter Zeitskalen beruhen. Dies deutet darauf hin, dass die seltenen Aktualisierungen, die der ereignisbasierten Repräsentation von Kontext zugrunde liegen, ein effizientes Kodierungsschema für prädiktive Prozesse beim Sprachverstehen darstellen.

Diskussion

Die Ergebnisse zeigen, dass (1) besseres Sprachverstehen von vorhersagbarer Sprache bei mittlerer Verständlichkeit mit der Aktivierung des Sprachnetzwerks, aber der Hemmung des cingulo-operulären Netzwerks einhergeht, (2) semantische Ähnlichkeit kein geeignetes Maß für die hierarchisch-prädiktive Sprachverarbeitung im Gehirn ist, (3) künstliche neuronale Netze prädiktive Prozesse im Menschen widerspiegeln, (4) Sprachvorhersagen auf der Zerlegung von Kontext in Ereignisse entlang einer temporo-parietalen Hierarchie beruhen.

Eine wichtige Implikation dieser Ergebnisse ist, dass die temporo-parietale *Surprisal*-Hierarchie auf ereignisbasierten Updates von Kontextrepräsentationen basiert, die den Grundstein für die Ereignissegmentierung, das hierarchische Arbeitsgedächtnis und die hierarchisch-prädiktive Sprachverarbeitung legen. Gleichzeitig unterstreichen diese Ergebnisse auch, dass uns die Verwendung von künstlichen neuronalen Netzen erlaubt, verschiedene Modelle für die Algorithmen im menschlichen Gehirn miteinander zu vergleichen.

Zusammengenommen rekrutierte das Gehirn für die Wortvorhersage unter anspruchsvollen Hörbedingungen ein umfassendes Netzwerk von Spracharealen, um das Verstehen zu erleichtert. Dieses Netzwerk umfasste temporo-parietale Regionen, die an der ereignisbasiert-prädiktiven Sprachverarbeitung beteiligt sind; den IFG, der typischerweise an Kontrollprozessen beteiligt ist; sowie parahippocampale Regionen, die typischerweise an Gedächtnisprozessen beteiligt sind. Zusätzlich erfüllte das cingulo-operculäre Netzwerk eine kompensatorische exekutive Kontrollfunktion, wenn kein semantischer Kontext für die prädiktive Sprachverarbeitung verfügbar war.

Abschließend schlage ich ein Modell vor, in dem das dynamische Zusammenspiel von Netzwerken des Arbeitsgedächtnisses, des Langzeitgedächtnisses und der kognitiven Kontrolle die Sprachvorhersage prägt. Das Modell besagt, dass ein ereignisbasierter dorsaler und ein zeitinvarianter ventraler Pfad in auditiven Regionen Sprache parallel in hierarchische Zeitskalen segmentieren. Diese Arbeitsgedächtnisrepräsentationen könnten die Bildung von Vorhersagen auf mehreren Zeitskalen unterstützen, wobei längere Zeitskalen Vorhersagen auf kürzeren Zeitskalen beeinflussen und Vorhersagefehler die Überarbeitung von Vorhersagen auf längeren Zeitskalen ermöglichen. Die Vorhersageprozesse auf längeren Zeitskalen könnten über den Hippocampus mit dem Langzeitgedächtnis verbunden sein. Schließlich könnte der IFG kognitive Kontrolle über Vorhersagen auf längeren Zeitskalen in auditiven Regionen ausüben, wobei der ventrale IFG komplexe interne Modelle für Vorhersagen bereitstellt und der dorsale IFG Vorhersagen unterstützt, wenn der Kontext keine eindeutige Vorhersage zulässt. Dieses Modell legt nahe, dass die hierarchisch-prädiktive Verarbeitung von Sprache über auditive Regionen hinausgeht und auch kognitive Kontrolle und Langzeitgedächtnis einbezieht.