Getting ahead

Prediction as a window into language, and language as a window into the predictive brain

Micha Heilbron

COLOFON

The work described in this thesis was carried out at the Donders Institute for Brain, Cognition, and Behaviour, Radboud University, with financial support from the Dutch Research Council (NWO), awarded to Micha Heilbron and to Floris de Lange.

ISBN

978-94-6284-275-5

Cover design

Studio FFF Amsterdam

Print

XXXXX

Getting ahead

Prediction as a window into language, and language as a window into the predictive brain

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 28 juni 2022
om 16.30 uur precies

door

Micha Heilbron
geboren op 11 september 1992
te Amsterdam

Promotoren

Prof. dr. F.P. de Lange

Prof. dr. P. Hagoort

Manuscriptcommissie

Prof. dr. J.M. McQueen

dr. E.G. Fedorenko (Massachusetts Institute of Technology, Verenigde Staten)

Prof. dr. J.B. Pierrehumbert (University of Oxford, Verenigd Koninkrijk)

Contents

1 Introduction **3**

2 Word contexts enhance the neural representation of individual letters in early visual cortex **23**

3 Tracking naturalistic linguistic predictions with deep neural language models **65**

4 A hierarchy of linguistic predictions during natural language comprehension **75**

5 Prior uncertainty modulates beta-band activity during the perception of natural speech **119**

6 Prediction and preview strongly affect reading times but not skipping during natural reading **147**

7 Discussion **181**

Nederlandse samenvatting **195**

Acknowledgements **199**

Publication list **203**

Curriculum vitae **205**

Donders Graduate School for Cognitive Neuroscience **207**

Research data management **209**

Bibliography **213**

Chapter 1

Introduction

Some passages in this chapter are based on:
de Lange FP., Heilbron M., and Kok, P. 2018. How do expectations shape perception? *Trends in Cognitive Sciences*. 22.9

Introduction

As you are reading these words, your eyes are scanning the page by making short erratic jumps, each time stopping just a fraction of a second – but long enough for your brain to transform little black marks into a conscious awareness of meaning. This transformation is perhaps even more striking for spoken language. Natural speech is a continuous stream, where words run into each other and often overlap. You hear it for what it is when hearing a foreign language: a seamless stream of sounds without any boundaries – *and it is all so fast.* Yet the brain can transform this stream into a chain of meaningful words and the relations between them, all virtually in real time.

How does the brain achieve this remarkable feat? It has been suggested that a key ingredient to the answer lies in the constant prediction of the incoming signal. This suggestion is what I have been investigating over the past four years of my life. It is a suggestion made by an emerging theoretical framework known as 'predictive processing' which describes the brain as essentially a prediction machine. The information processing schemes envisioned by the framework can account for a breathtaking range of phenomena – from psychology to physiology – and promise to offer unifying computational principles of brain function. Surprisingly, these very same processing schemes are found (in a rudimentary form) in the speech recognition and predictive text systems on our phones, and are the driving force behind some of the most spectacular breakthroughs in artificial intelligence of the past few years.

So far, most empirical work on predictive processing has focussed on perception – in particular vision[1]. But as I will explain in this introduction, I believe that predictive processing can also shine new light on language – and, conversely, that studying language can provide unique insights into the predictive brain. To do so, I first review the key ideas of the framework. Then I discuss how they relate to similar ideas about language processing that have been explored before, and outline what the new framework may offer the study of language. Ultimately, the picture that emerges from the work in this thesis is one of language processing as inherently predictive. A view in which the brain's ability to understand or 'follow' any piece of language depends on its ability to get ahead and predict it.

But I'm getting ahead of myself here. Let's start with the basics: what is predictive processing?

[1]Some notable exceptions notwithstanding (see e.g. Arnal, Wyart, and Giraud, 2011; Blank and Davis, 2016; Gagnepain, Henson, and Davis, 2012; Shain et al., 2020; Sohoglu and Davis, 2020; Sohoglu et al., 2012 for work on language.)

Predictive processing as a framework for studying the brain

Predictive processing is not a single model or even a single theory – instead it refers to a broad class of models that propose that information processing in the brain fundamentally relies on predictions. The framework combines, unifies, and in some cases rediscovers a range of ideas from psychology or neuroscience, machine learning and information theory. Many of these ideas have a long history, but only in the past decade or so it was recognised that they can be unified into a single overarching framework for studying the brain (Clark, 2013; Friston, 2010; Heeger, 2017; Hohwy, 2013; Huang and Rao, 2011). A key tenet of the framework is that the brain constructs *generative* models of the world. A model is called generative when it captures the full statistical structure of the data such that it can generate new data instances – such as images in the case of vision or sentences in the case of language. Predictive processing models typically focus on neocortex, which they cast as embodying a *hierarchical* generative model – a model that can generate patterns of activity 'from the top-down' that external stimuli would elicit 'from the bottom-up' (Friston, 2005; Lee and Mumford, 2003; Mumford, 1992; Rao and Ballard, 1999). Contrasting with traditional accounts of cortical processing as merely the bottom-up detection of increasingly abstract features (Marr, 1982; Riesenhuber and Poggio, 1999), the framework proposes a more active notion of the brain constantly trying to predict the incoming input and minimise the prediction error: the discrepancy between the prediction and the signal.

Arguably, the primary appeal of predictive processing is the broad scope of its explanatory potential. Supposedly, this single scheme of prediction and prediction-error minimisation can account for breathtaking range of phenomena, from (classical) tuning of low-level sensory neurons and the (extra-classical) modulations thereof (Huang and Rao, 2011; Rao and Ballard, 1999), via perceptual phenomena such as illusions and gestalt principles (Bar, 2007; Yuille and Kersten, 2006), up to imagination, sensorimotor control, action and motivation (Den Ouden, Kok, and De Lange, 2012; Friston et al., 2017). As such, it is proposed as a fundamental neurocomputational principle (Keller and Mrsic-Flogel, 2018) and has been even hailed by some a 'grand unified theory' of mind and brain (Clark, 2013; Friston, 2010; Hohwy, 2013). Since we are considering language processing (in the receptive sense, so excluding language production), we are primarily interested in three putative functions of prediction, which I detail below.

First, prediction can be used for *recognition* – or more generally, inference. This is motivated by the insight that most recognition problems the brain faces are *inverse problems*: problems that require inverting the arrow of causality. For instance, object recognition requires going from the activity patterns that make up a sensory

impression back to the external object that caused them. Similarly, parsing can be seen as inferring the hidden syntactic structure used to generate an incoming word sequence. Solving inverse problems is difficult because they are typically *ill-posed*, having multiple (often infinitely many) possible solutions. Finding a solution therefore requires imposing *constraints* that can rule out alternative solutions. Drawing on Bayesian and 'analysis-by-synthesis' approaches to perception (Kersten, Mamassian, and Yuille, 2004; Lee and Mumford, 2003; Neisser, 1967; Yuille and Kersten, 2006), predictive processing proposes that predictions derived from generative models offer such constraints.

Some of these predictions stem from low-level distributional priors (such as that light typically comes from above). However, recognition often benefits from more complex, contextual predictions – such as when a faint edge turns out to be a critical boundary only in the light of a higher-level interpretation of a scene (Fig 1.1). Such inferences are proposed to operate via an analysis-by-synthesis procedure, where low-level cues deliver bottom-up proposals which evoke higher-level hypotheses which are then tested by comparing the input to a top-down prediction of the signal. This top-down prediction can be seen as a knowledge-based reconstruction of the input – and the procedure is called analysis by synthesis since perceptual analysis is performed by comparing the raw incoming signal to the synthesised top-down signal. Although algorithmic details differ, models generally propose that this is implemented with a single operation, where each 'level' of the cortical hierarchy tries to predict the activity patterns at the level below (Friston, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999). Bad predictions result in prediction errors which are used to finesse the prediction at the higher level. This procedure is repeated, simultaneously throughout the hierarchy, until the error is minimised and perception is achieved. Empirically, this scheme can account for a wide range of findings in perception, and is more directly supported by studies reporting signatures of prediction error and the effects of top-down predictive feedback (see de Lange, Heilbron, and Kok, 2018; Heilbron and Chait, 2018 for reviews).

The strong emphasis on prior knowledge raises an obvious question: how does the brain learn such powerful internal models? Intriguingly, learning can itself be driven by prediction. Predictive processing models typically propose that the very same error-correction algorithm that (over short timescales) drives inference also (over longer timescales) drives learning (Friston, 2005; Rao and Ballard, 1999). Here the framework integrates insights from predictive self-supervised learning, a form of learning where by predicting the input, a system can use prediction errors to perform error-driven learning without supervision. In AI, this form of predictive learning is today one of the most rapidly developing areas, resulting in highly influential unsupervised learning techniques like contrastive predictive coding (Henaff et al., 2019;

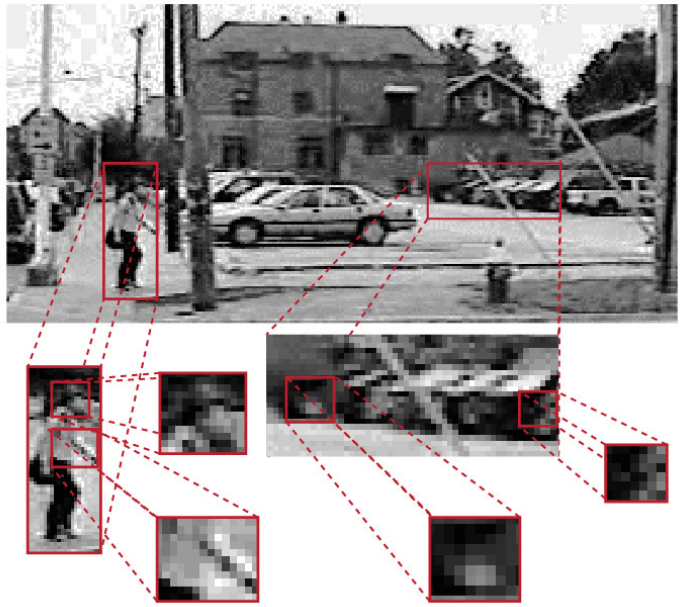

Figure 1.1. Ambiguity in low-level vision and its resolution by high-level context. When only analysing a small region of a natural image, even detecting simple features like edges can be difficult. In the image, distinguishing the pedestrian's shirt or hair from the background – or their arm from an edge in the sidewalk – is practically impossible based on a local analysis of low-level features. But at a high level the image is not ambiguous: we can easily see a pedestrian and a row of parked cars. Critically, this high-level information in turn allows us to disambiguate the low-level information: having identified a walking person we can recognise a subtle difference in luminance as the key boundary between the pedestrian and the sidewalk. Predictive processing models propose that this disambiguating effect of high-level context on low-level features is implemented by top-down connections from high-level cortical neurons (analysing abstract features at larger spatial scales) to low-level neurons (analysing simple features at smaller spatial scales). Image adapted from UIUC cars dataset.

LeCun, 2016; Liu et al., 2021; Oord, Li, and Vinyals, 2019). In neuroscience, such generative-model based predictive learning has been theoretically explored as an alternative to (Dayan et al., 1995; Hinton et al., 1995) or implementations of (Whittington and Bogacz, 2019) standard backpropagation, and can empirically account for a range of learning and plasticity effects observed in the brain (Bakhtiari et al., 2021; Gillon et al., 2021; Jehee et al., 2006; Rao and Ballard, 1999).

Finally, prediction can be used for *compression*. The term predictive coding was originally coined as the name of a signal compression technique. The key idea here is that if only the deviation or prediction error is encoded, the predictable component of the signal can be discarded, resulting in a more efficient code. Predictive coding

models implement this kind of compression strategy as an integral part of neural processing, by postulating that only the error term is transmitted to higher-order areas[2] (Friston, 2005; Rao and Ballard, 1999). Here the framework connects to and unifies key notions from efficient and sparse coding (Barlow, 1961; Chalk, Marre, and Tkacik, 2018; Olshausen and Field, 1996; Smith and Lewicki, 2006). Empirically, this form of predictive compression can account for many low-level physiological phenomena such as response properties of neurons, ranging from the retina (Hosoya, Baccus, and Meister, 2005; Srinivasan, Laughlin, and Dubs, 1982) to sensory cortex (Huang and Rao, 2011; Rubin et al., 2016).

Summing up, predictive processing proposes how prediction can drive recognition or inference, learning, and compression. The principles can account for, and are supported by, a range of findings from the perceptual system – suggesting that predictions may be fundamental to neural computation.

Predictive language processing: a pre-history

Language seems like an ideal domain for the predictive processing approach. After all, language is full of regularities that must be learned and can be used to predict (Chang, Dell, and Bock, 2006), as wel as redundancies that can be compressed (Jaeger, 2010; Shannon, 1951). At the same time, language is rife with ambiguities – at all levels of analysis – that require priors or contextual predictions to resolve (Hindle and Rooth, 1993; Piantadosi, Tily, and Gibson, 2012). It should perhaps come as no surprise, then, that most of the key ideas of predictive processing have long been explored within the realm of language – often before being applied to perception or cognition more broadly. To assess what the contemporary view of the predictive brain can offer the study of language processing, I will first briefly review this theoretical and empirical 'pre-history' of predictive processing of language.

Theoretical arguments for predictive language processing

For *recognition*, the idea that perceptual analysis cannot only consist of a passive detection of features but must involve some active, reconstructive (i.e. *generative*) component was recognised early on in the cognitive science of language. For instance, already in the late fifties and early sixties, Halle and Stevens proposed the *analysis by synthesis* model of speech perception (Halle and Stevens, 1962; Stevens, 1960). This model proposed that listeners generate hypotheses about the incoming words

[2] In hierarchical predictive coding models of cortex, the error coding scheme is in an important way different from purely-compression-based schemes (such as those in the retina), because the predictable information is not just discarded but is still represented and internally reconstructed from top-down (or lateral) connections.

Box 1.1 | What's in a prediction?

In psycholinguistics, the term prediction is typically used in a temporal sense only: it is reserved for anticipatory pre-activation of information. Here, I follow the predictive processing framework and use a more inclusive definition. In fact, while some predictive processing models deal with temporal prediction (e.g. Heeger, 2017; Lotter, Kreiman, and Cox, 2017) most models (all models of vision) do not even include a temporal dimension, and only generate so-called *nowcasts* – predictions of the present. While this may seem like a misnomer, it makes sense when we consider the statistical definition of the term. Statistically, a prediction is an extrapolation from a model to potential observations. Whereas a model is specified via parameters over latent variables, a prediction is specified in terms of observable data. In other words, given a model and a higher-level hypothesis, *a prediction tells us what observations to expect.* The essence of prediction lies in the absence of sufficient data. Whether this is because the prediction is about the future – or because it is about current but not yet (fully) observed events – is irrelevant. Essential is the inference from a latent to expected observations that goes beyond the data given.

This definition is useful because it shows how seemingly unrelated phenomena – from biases in orientation perception to anticipatory effects in recognition – may be examples of the same principle and can be explained in the same mathematical terms. In language, most of the interesting regularities unfold over time, such that the statistical and colloquial sense of prediction will often align. However, this is not always the case, such as when the brain can 'predict' a letter from its neighbouring letters (**Chapter 2**). For me, whether such an inference is a 'prediction' depends on the mechanism – whether it involves inferences about expected observations from a generative model – but not on whether it involves prediction over time.

and that recognition is achieved by comparing the incoming signal to internally synthesised top-down predictions of the most likely *potential* phonemes. Interestingly, while its direct impact on models of language understanding was limited (Bever and Poeppel, 2010), the analysis by synthesis framework became more widely known after of its extension into a framework for visual perception (Neisser, 1967). In its own way, the more influential (and contentious) motor theory of speech perception also casted speech perception as the matching of acoustic input to actively generated predictions derived from a generative (articulatory) model – and did so in the sixties

already (Liberman et al., 1967).

Regarding processing architecture, the very mechanism of top-down recognition itself – at least its computationally explicit form – was pioneered in early connectionist models of language: the interactive activation model of visual word recognition (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) and its extension to speech, TRACE (McClelland and Elman, 1986; McClelland, Rumelhart, and Group, 1986). These models explained context effects via top-down connections that allowed information from the higher-order (e.g. word) level to directly guide recognition at the lower-order (e.g. letter) level. Because these top-down connections encode which lower-level features to expect given some higher-level expectation, they formally embody a top-down generative model. While the authors did not yet use this term (nor the term 'prior' or 'prediction'), these interactive models of word recognition are precursors of later hierarchical bayesian models of perception more broadly (Lee and Mumford, 2003; McClelland, 2013; Yuille and Kersten, 2006). [3]

Meanwhile, and largely isolated from developments in cognitive science, engineers working on language converged on their own kind of predictive processing solution, with the statistical approach to automatic speech recognition (ASR; Bahl, Jelinek, and Mercer, 1983; Baker, 1975; Jelinek, Bahl, and Mercer, 1975). A signature characteristic of the approach is that recognition systems constitute not just an acoustic model but also a *language model* – a generative model computing the probability of a word given the preceding words. In other words, these systems recognise speech not just as a function of the incoming acoustics but also as a function of its internal predictions of which word is coming next (Jelinek, 1998). Current ASR systems use end-to-end neural networks instead of explicit probabilistic models (Baevski et al., 2020; Bahdanau et al., 2016; Chan et al., 2016; Graves, Mohamed, and Hinton, 2013); however, they still use language models and fundamentally rely on the same predictive strategy (Jurafsky and Martin, 2021; Toshniwal et al., 2018). While the success of this engineering solution does not necessarily tell us anything about the brain, it is a striking fact that in all these decades, every successful recognition system – for spoken and written language alike – has relied on linguistic predictions from a generative model (Jelinek, 1998; Jurafsky and Martin, 2021; Schroeder, 2004).

The statistical approach also came to dominate other domains of computational linguistics, notably models of grammar and parsing (Bod, Scha, and Sima'an, 2003; Charniak, 1997; Manning and Schutze, 1999). Psycholinguists built on these developments by combining probabilistic context-free grammars (PCFG) with top-down probabilistic parsers to formulate theories of syntactic comprehension (Hale, 2001;

[3] While top-down connections are an elegant implementation of priors – especially in a hierarchical scheme (Lee and Mumford, 2003) – they should not be equivocated: it is perfectly possible to construct a bottom-up Bayesian model.

Levy, 2008). By casting comprehension as an expectation-based process in which the brain is constantly predicting all potential full-sentence analyses consistent with the input so far, these theories could explain a wide range of syntactic processing phenomena, and became considerably influential.

Beyond recognition or inference, other functions of prediction proposed by the predictive processing framework have also been explored in the domain of language. Notably, the idea that prediction can drive *learning* was explored by Elman in his pioneering work on recurrent neural networks (RNNs; Elman, 1990, 1991). Elman showed that by predicting the next word given the previous words, a neural network could perform error-driven learning without supervision. Strikingly, the representations learned by these networks captured abstract distinctions (such as between nouns and verbs) and multiple levels of subcategories within them (such as animate vs. inanimate objects; Elman, 1990). In other words, by simply training the networks to predict, the models *learned to 'understand'*. While Elman worked with highly simplified 'toy languages', these findings are the foundation for recent breakthroughs in *natural* language processing. Recently, deep learning based language processing systems have dramatically improved and are now deployed in many applications – mostly by leveraging the predictive learning explored by Elman. These models are trained simply to predict words in a context, but then learn about language much more broadly, and can be applied to practically any language processing task (Brown et al., 2020; Devlin et al., 2019; Peters et al., 2018; Radford et al., 2018, 2019; Ruder et al., 2019); and develop representations that can be used to predict brain response patterns to language (Caucheteux and King, 2020; Schrimpf et al., 2020). Careful analysis of these networks has revealed that they are not just practically useful, but 'understand' a striking amount of linguistic structure – all discovered by simply predicting language, without any supervision (Linzen and Baroni, 2021; Manning et al., 2020).

Finally, the idea that prediction can be used for *compression* also has long been applied to language. For instance, linear predictive coding (LPC; Elias, 1955) has been used for speech compression for decades (Atal and Schroeder, 1970) and is still being applied, for instance in the speech codec of *Skype* (Gray, 2010; Koen, Skak, and Vandborg, 2010). In the study of perception, predictive compression principles have been used to explain various neural/cognitive phenomena, such as the response characteristics of sensory neurons (Bialek, Nemenman, and Tishby, 2001; Chalk, Marre, and Tkacik, 2018; Gill et al., 2008; Huang and Rao, 2011; Rubin et al., 2016). For the neuroscience of language, I am not aware of such usages of compression principles – it seems that prediction-for-compression has so far remained a feat of engineering only.

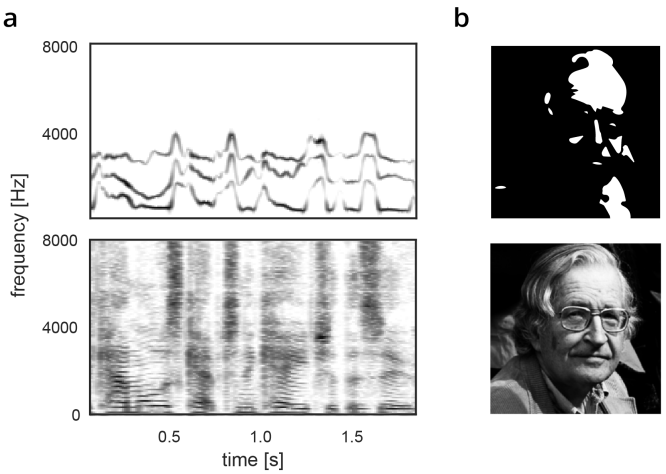

Figure 1.2. Top-down perceptual insight in ambiguous stimuli. a) A spectrogram of a spoken sentence (lower row) and its heavily reduced sine-wave speech representation (top row). The sine wave speech lacks spectral detail: traditional acoustic cues like distinctive features are completely absent and the sound is often not even perceived as speech initially. However, appropriate context can introduce dramatic gestalt switch such that the sine waves are perceived as a fully intelligible sentence. For a demonstration, click here to listen to the sine wave speech recording and here for the clear recording.[4] **b)** Two-tone or Mooney image. The upper panel may appear as a random arrangement of black and white shapes – but to people who have taken a linguistics class or have otherwise seen the lower panel before, the image can be recognised as the contours of a famous thinker. Image: Wikimedia commons.

Empirical support for predictive language processing

Empirically, there have also long been findings well in line with predictive processing.

For instance, there is a wealth of behavioural work showing strong effects of context on word recognition. In fact, one of the earliest phenomena described in empirical psychology is such an effect: the fact that letters are more easily recognised when embedded in a word (Cattell, 1886), also known the word superiority effect (Reicher, 1969; Wheeler, 1970). Contextual effects are perhaps even stronger in speech, where linguistic context can greatly enhance recognition of degraded speech (Miller, Heise, and Lichten, 1951). One of the most dramatic example of such enhancement is found in sine wave speech (Remez et al., 1981). This is a technique that reduces the full speech spectrogram to only three or four sinusoids, resulting in something like the acoustic version of a Mooney image (Mooney, 1957; see Figure 1.2). As with Mooney images, higher-level context can induce dramatic gestalt switches, trans-

[4] Or type tinyurl.com/swsent or tinyurl.com/normsent in your browser, respectively. Sounds are from the personal website of Dr Matt Davis at the University of Cambridge.

forming the perception from some strange beeps and whistles into a fully intelligible spoken sentence – often considered a hallmark of top-down effects in speech (Davis and Johnsrude, 2007). Context effects in speech also exist at the lexical level. In the phonemic restoration effect, one segment of a word is replaced by noise (e.g. el_phant) but subsequently restored or 'filled-in' during perception. This effect is so strong that participants often report having heard the missing segment (Warren, 1970). Even sub-lexical knowledge can such effects, as illustrated by illusory vowels. When subjects are presented with consonant clusters that are phonotactically illegal in their native language, their brain 'restores' the stimulus by perceptually inserting a vowel such that the percept confirms to their language. Native speakers of Japanese, for instance, would perceive the nonword /ebzo/ as /ebuzo/ (Dupoux et al., 1999).

Another set of suggestive clues comes from electrophysiology, which has long shown that the brain responds differently to words that violate linguistic expectations or regularities. In fact, this goes back to one of the first reported and most discussed neural signatures of language: the N400 (Kutas and Hillyard, 1980; Kutas and Hillyard, 1984). The amplitude of this negative deflection peaking at 400 ms post word onset is stronger for unpredictable or anomalous words – and this effect is a graded, continuous function of the degree of unexpectedness (Kutas and Hillyard, 1984). Interestingly, different responses have been shown to be sensitive to violations of linguistic regularities at different levels, such as the P600 to syntactic violations (Hagoort, Brown, and Groothusen, 1993; Osterhout and Holcomb, 1992), or the N200 to phonological violations (Brink, Brown, and Hagoort, 2001). More recent computational approaches have shown that the amplitude modulation of the N400 by predictability can be well-captured in terms information-theoretic surprisal (the negative log probability; Frank et al., 2015). Although alternative interpretations exist (see below), this modulation by predictability fits neatly with a phenomenon found in neuroscience more widely: that response strength is proportional to a stimulus' unexpectedness – and specifically its negative log probability (Friston, 2005).

Eye movements provide another rich source of clues. For instance, words that are unexpected in context are read more slowly, as measured by fixation durations in eye tracking or response times in self-paced reading (Ehrlich and Rayner, 1981; Staub, 2015). The effect is analogous to the predictability effect on the N400, strongly suggesting that predictable words are processed more efficiently. Strikingly, computational modelling of large datasets of reading times has shown that this effect exists not just for highly predictable 'target' words, but was found to scale logarithmically with a word's contextual probability (i.e. linear in surprisal) up to 6 orders of magnitude (Smith and Levy, 2013). In other words, the effect is not just present when comparing words of probability of 0.9 vs 0.09, but is equally strong for seemingly sub-

tler differences such as between $p = 0.009$ vs $p = 0.0009$. While the predictability effect has been known since the 1980's (Ehrlich and Rayner, 1981), a more recent and highly influential line of work comes form the visual world paradigm. Here, participants view visual scenes while being presented with spoken sentences (Allopenna, Magnuson, and Tanenhaus, 1998). This revealed that the eyes move spontaneously to the object that the unfolding sentence *could* be referring to – often in clear anticipation of presentation of the predicted target word.

Together, the empirical literature work has long shown that that language processing is highly contextual and incremental, and that unpredictable words are processed less efficiently.

Controversies and open questions

And yet – despite the long history of appealing theoretical arguments and empirical clues – the idea of predictive language processing has for decades been rejected in psycholinguistics and remains contentious until today. While in the sixties psycholinguists still proposed that comprehenders constantly generate hypotheses about upcoming words (Goodman, 1967; Miller and Isard, 1963) this idea was then discarded and remained almost taboo until at least the early 2000s. This state of affairs is well characterised in an inventory by Van Berkum et al. (2005), who note that (shy of one exception) *"the authors of recent psycholinguistics textbooks (e.g., Harley, 2001; Jay, 2003; Whitney, 1998) make no reference to the possibility that people might predict upcoming language"*; and that *"prediction has also been notably absent in authoritative monographs and survey chapters on language comprehension (e.g., Cutler & Clifton, 1999; Frazier, 1999; Kintsch, 1998; Perfetti, 1999; Pinker, 1994)."*

A major reason for this enduring skepticism is the the influence of two theoretical frameworks. First, the modularity of mind in cognitive science, with its informational encapsulation and bottom-up processing (Fodor, 1983; Forster, 1981; Forster, 1989); and second generative grammar in linguistics which, since Chomsky's famous critiques of Markov models (Chomsky, 1957), treated statistical approaches to language with deep suspicion (Pereira, 2000). The idea was that prediction would be unduly costly and hopelessly inefficient since the open-ended, generative nature of language makes it fundamentally *unpredictable*. This argument is well illustrated by an often cited footnote by Jackendoff (2003, p. 59) where he explains why it is not *"useful to conceive of understanding the sentence in terms of predicting what word will come next"*: *"One might well predict that what comes after little in 'The big star's behind a little ...' is likely to be a noun (though it might be "blue" or even "very old") but that still leaves open some tens of thousands of other choices."* (Note how this casually glosses over the fact that even inferring that the next word will likely be a noun

can represent a considerable reduction in uncertainty – a testament to the historical aversion to think about language in statistical, information-theoretic terms.)

Apart from these powerful theoretical inclinations, a key empirical motivation for rejecting the role of prediction was the perceived lack of misprediction costs. For instance, already from the pioneering work by Marta Kutas (Kutas and Hillyard, 1984), the N400 seemed primarily sensitive to the semantic relatedness between a word and the preceding context, but little or not to whether it matched the specifically predicted 'target' word (see Van Petten and Luka, 2012 for a comprehensive discussion). As such, the well-known predictability effects on N400 amplitude and reading times were explained as a kind of confound. Because words that are more predictable in a context *also* tend to be more semantically related to that context, they could therefore – by virtue of mere 'intra-lexical priming' (Van Berkum et al., 2005) – be *easier to integrate* (see, e.g. Brown and Hagoort, 1993). Context effects in recognition were similarly explained in a bottom-up fashion, as reflecting easier integration and reduced thresholds at a downstream, post-perceptual decision stage (Norris, 1994; Norris, McQueen, and Cutler, 2000).

But in the past decade or so, the tides have been turning. Due to the rise of probabilistic/Bayesian models in psycholinguistics (Chater and Manning, 2006; Jurafsky, 2003) and psychology and neuroscience (Knill and Pouget, 2004), and a range of ingenious experimental designs (Allopenna, Magnuson, and Tanenhaus, 1998; DeLong, Urbach, and Kutas, 2005; Van Berkum et al., 2005) the idea of top-down and predictive language processing steadily gained traction (Altmann and Mirkovic, 2009; Davis and Johnsrude, 2007; Dell and Chang, 2014; Pickering and Garrod, 2013). However, the precise role of prediction remains hotly debated, especially after one of the most influential studies supporting a strong form of prediction (DeLong, Urbach, and Kutas, 2005) recently failed to replicate (Nieuwland et al., 2018; see also Nieuwland, Arkhipova, and Rodriguez-Gomez, 2020).

The literature on predictive processes in language spans multiple subproblems – from top-down effects in word recognition (Balota, Yap, and Cortese, 2006; Davis and Johnsrude, 2007; McClelland, Mirman, and Holt, 2006; Norris, McQueen, and Cutler, 2000) to anticipation in sentence processing (Kutas, DeLong, and Smith, 2011) – and subfields – such as the ERP literature (Federmeier, 2007; Nieuwland, 2019) and the eye movement literature (Huettig, Rommers, and Meyer, 2011; Staub, 2015). Therefore, a comprehensive overview of the empirical debate is beyond the scope of this introduction. However, one way to roughly characterise the discussion at large, is to notice that much disagreement centers around two key questions in particular (see also Huettig, 2015; Ryskin, Levy, and Fedorenko, 2020 for a similar characterisation). Namely, *when* does language processing involve prediction, and *what* is being predicted? I will briefly elaborate on each question in turn.

When does language processing involve prediction? Whether it is from first-person impressions – such as the temptation to finish someone's sentences when talking to a person with a stutter – or from ingenious experimental designs – such as shadowing (Marslen-Wilson, 1973) or visual world paradigms (Allopenna, Magnuson, and Tanenhaus, 1998) – it is clear that language comprehension *can*, at least sometimes, invoke predictions. One of the key disagreements, however, regards the conditions under which this occurs. While some suggest that prediction occurs constantly as an integral part of language processing (Federmeier, 2007; Kuperberg and Jaeger, 2016; Kutas, DeLong, and Smith, 2011) others suggest that comprehenders might only predict in (relatively rare) highly constraining contexts, *"and otherwise adopt a laissez-faire "wait and see" strategy"* (Van Petten and Luka, 2012).

A complicating factor here, noted by various authors (Huettig and Mani, 2016; Kutas, DeLong, and Smith, 2011; Van Petten and Luka, 2012), is that many studies on prediction focus on precisely such highly constraining contexts. Other aspects of popular experimental designs, such as the presence of the predicted target in visual world paradigms, or the use of slow, word-by-word visual presentation of sentence materials in ERP studies (but see Brink and Hagoort, 2004; Hagoort and Brown, 2000), have also been criticised as being 'prediction-encouraging' (Huettig and Mani, 2016; Van Petten and Luka, 2012). Therefore, even when studies can rule out alternative explanations like integration difficulty, it often remains unclear whether the observed prediction effect is representative of language processing in general (see Mantegna et al., 2019; Nieuwland et al., 2020 for recent examples of this conundrum). In similar vein, even the well-established effects of lexical and sentential context on word recognition – which supposedly reflect the active nature of the process – have been shown to be variable and can for instance depend on the ambiguity of the input (Burton, Baum, and Blumstein, 1989; McQueen, 1991). In other words, perhaps word recognition is strongly driven by prior knowledge only when the input is noisy or ambiguous, such as in many studies – sine wave speech here being the extreme case.

What is being predicted? Aside from when the brain relies on predictions, a second key question regards the representational content and nature of such predictions. Arguably the least controversial proposal is that the brain is engaged in prediction at a highly abstract, semantic level (Federmeier, 2007). Other proposals like *surprisal theory*, primarily describe forward-looking, predictive effects at the level of syntax (Hale, 2001; Levy, 2008). Beyond these more abstract levels of representations – and much more controversially – some have suggested, both based on empirical results (Van Berkum et al., 2005) and theoretical considerations (Smith and Levy, 2008) that predictions are made at the lexical level. The strongest proposals take this notion even further and propose that prediction occurs at all representational levels simultaneously, down to visual and auditory word forms (Dikker et al., 2010; Kuperberg

and Jaeger, 2016). The discussion on top-down effects in word recognition can be understood as revolving around this same question, as it boils down to whether contextual information (anticipatory or not) is propagated down to the earliest levels or whether it is confined to later processing stages.

One complicating factor in the issue of processing levels is that an unexpected word is almost always unexpected in multiple ways. For instance, when replacing a highly predicted noun with a verb, the word not only has a different syntactic category but also (at least slightly) a different meaning, phonological form, etc. Analogously, a single effect can often be explained at multiple levels. For instance, effects of word unexpectedness (e.g. on reading tines or the N400) can be explained predictively either as reflecting prediction at the lexical level directly (Frank et al., 2015; Smith and Levy, 2008; Szewczyk and Schriefers, 2018); or, in the case of the N400, as reflecting prediction error at the *semantic* level (Rabovsky, Hansen, and McClelland, 2018). However, effects of lexical unexpectedness could also reflect expectation-based *parsing*, where a word's predictability determines the size of the update (in KL divergence) of each potential whole-sentence syntactic interpretation (Hale, 2001; Levy, 2008).

Independent of the issue of processing level is the question whether predictions are probabilistic. To many in psycholinguistics, the notion of prediction was by definition reserved for the all-or-none pre-activation of specific lexical items (Luke and Christianson, 2016; Van Berkum et al., 2005; Van Petten and Luka, 2012). A conceptual (or at least terminological) difficulty thus arrises that what one person might consider e.g. a probabilistic semantic prediction (Federmeier, 2007; Rabovsky, Hansen, and McClelland, 2018) would to others not count as a prediction at all (Van Petten and Luka, 2012). Some results might seem to suggest a probabilistic view, such as correlations between the extent of a word's expectedness and the N400 amplitude it evokes. However, because data is analysed in the aggregate, such a correlation *on average* could both reflect trial-by-trial sensitivity to the *amount* of predictability, or the trial-by-trial (or participant-by-participant) *probability* that the word was either (categorically) predicted or not (Van Petten and Luka, 2012).

Language in a predictive processing framework

Given this long history of work on and debate about predictive processes in language, what can an extremely general framework like predictive processing still offer to the study of language? And conversely, what can studying *language* – of all cognitive faculties – teach us about the predictive brain?

Why predictive processing?

Theoretically, placing language in a predictive processing perspective offers theoretical unification. Many of the key notions (generative models, predictive recognition, predictive learning, top-down effects) are not technically new to the study of language but have been studied somewhat independently across psycholinguistics, computational linguistics and artificial intelligence. Predictive processing offers to unify these ideas and connect them to neuroscience and psychology more broadly by understanding them in terms of universal principles of neural computation.

Empirically, the framework offers principled and testable – and rather bold – answers to the two key questions surrounding the role of prediction in language processing outlined above. In response to the first question (i.e. the *when* question), predictive processing suggests that language processing should *always* involve a degree of prediction, and not be limited to specific conditions or tasks. This question is addressed in **Chapter 4** and more indirectly in chapter **Chapter 3, 5, and 6**. Regarding the representational nature of predictions (i.e. the *what* question), the framework proposes that prediction should in principle occur at *all levels of analysis* [5] and that predictions should be probabilistic. Moreover, following hierarchical inference, processing at higher levels should extensively inform processing at lower levels. This implies that context effect in recognition should (at least in part) involve top-down, interactive processing, and that high-level predictions should inform low-level ones. I test these hypotheses in **Chapter 2, 4, and 6**. Finally, predictive processing (or at least dominant incarnations thereof) inspire neural hypotheses linking specific aspects of prediction to top-down and bottom-up signalling, which have been linked to neural oscillations in specific frequency bands. I test this in chapter **Chapter 5**.

Why language?

But we should also ask the opposite question: what can studying language tell us about the predictive brain? Why not study something 'simpler' (like visual perception) where the bottom-up processing stream is better characterised? Indeed, trying to test fundamental principes of neural information processing by studying something as complex as language may even seem absurd. However, I believe language is an appealing test ground, because it has least two properties which offer unique opportunities for studying predictive processing.

First, language is governed by complex and yet relatively transparent regularities. More technically, linguistic regularities are – compared to for instance visual regu-

[5]Note that this is a key difference with predictive recognition as currently used in e.g. speech recognition systems, where prediction is only used during decoding and only at a single representational level, typically the lexicon.

larities – relatively low-dimensional. This makes it comparatively straightforward to develop powerful, broad-coverage generative models which can approximate hypothetical linguistic priors or predictions in the brain for arbitrary linguistic input. In a high-dimensional domain like vision or audition, by contrast, priors/predictions are – beyond highly constrained, simplified problems – often rather elusive. Instead of trying to quantify predictions or priors, studies on predictive processing in perception have therefore largely resorted to experimentally-imposed regularities, which are typically extremely simple (such as arbitrary associations between stimuli). When studying language, by contrast, one can simply use natural language and approximate the linguistic expectations that ostensibly arise spontaneously using a generative model (see, e.g. Frank et al., 2015; Smith and Levy, 2013; Willems et al., 2016). Language, in other words, allows for studying predictive processing *in the wild*. This is important because predictive processing supposedly applies to *all neural processing*, not just to the specific case of actively engaged subjects perceiving extremely simple regularities. In my doctoral work, I have extensively leveraged this opportunity by studying predictions during natural language comprehension, see **Chapter 3, 4, 5, 6**.

Second, because language consists of discrete symbols that combine in a strictly compositional fashion, language also has a uniquely transparent set of *hierarchical levels*. This is important because multi-level and hierarchical prediction is central to the framework (see Friston, 2008; Lee and Mumford, 2003 for more technical discussions). Vision is also hierarchical and arguably compositional, but while the lowest levels are well-characterised (e.g. Gabor filterbanks) the higher levels get very mysterious very quickly. At a high level, one could say that scenes are composed of objects, but for most objects we have no idea what they would be composed of. Sentences, by contrast, consist of words, which in turn consist of phonemes (or letters). This allows to manipulate and quantify the predictions at different levels and scales, such as local predictions (of letters or phonemes within words) and global predictions (of words within sentences). In perception, this has been done before, but typically using experimentally-imposed and slightly contrived processing levels (e.g. tones-within-stmuli vs stimuli-within-block, see Chao et al., 2018; Wacongne et al., 2011). In language, by contrast, we can probe predictions at processing levels that are intrinsic to language and directly relate to the processing hierarchy itself. Apart from timescales, language also has distinct processing levels at the same scale (e.g. syntax and semantics) which can also be explicitly dissociated relatively easily. Throughout this thesis, I have greatly made use of this property, by manipulating and modelling predictions at different processing levels (**Chapter 2, 4**), and by testing which levels of contextual information are taken into account (**Chapter 4, 6**).

Outline of this thesis

The overarching goal of this thesis is to evaluate the predictive processing framework using language processing as a testbed. Given the scope of the framework, however, this is an admittedly daunting endeavour. In none of the chapters I therefore attempt (or pretend) to address it directly. Instead, I take an indirect route. Each chapter focusses on a specific question that touches on predictive processing more broadly – and can be construed as a testcase for the framework – while being specific enough to address. In the spirit of this broad framework, I take a wide-ranging approach, studying written and spoken language processing, in experiments ranging from participants viewing single letters to reading an entire novel.

In **Chapter 2** I take on the classic Word Superiority Effect – the phenomenon that letters are more easily recognised when embedded in a word. Top-down models of visual word recognition (and predictive processing more broadly) propose that this effect is at least in part perceptual in nature, with linguistic knowledge enhancing letter perception from the top-down. I tested this prediction in a tightly controlled fMRI experiment, and found strong evidence for representational enhancement, which was functionally coupled to the activation level in key areas of the reading network. These results are the first neural evidence of top-down representational enhancement in letter perception, and demonstrate how lexical context can modulate perceptual processing already at the earliest visual regions.

Chapter 3 is a short proof-of-principle chapter. Inspired by the publication of GPT-2 – a neural network that constituted a giant leap in the quality of generative language models – I explore how this model can be combined with deconvolution techniques to study word predictability effects on EEG in naturalistic conditions (audiobook listening). Using public domain EEG data, I show that the unexpectedness estimates from the network correlate with the brain response – revealing a modulation that exactly reproduces the N400 – and provide a better fit than previously used trigram language models. The results demonstrate that predictability effects are not a side-effect of 'prediction encouraging' designs, and highlight the potential of recent advances in AI for the cognitive neuroscience of language.

In **Chapter 4**, I build on and extend this work by more explicitly testing both the ubiquity and representational status of linguistic predictions – i.e. the *when* and *what* questions. I also include another dataset of high-quality MEG recordings that allow for high-precision source localisation. Prediction effects were clearly found over and above those of non-predictive confounds (such as acoustics and semantic integration) and were best explained by a model of casting prediction as probabilistic and ubiquitous. Next, by mathematically disentangling the lexical predictions from GPT-2 into distinct linguistic dimensions, I find dissociable signatures of syntactic,

phonemic and semantic predictions. Finally, I show that high-level (word-in-context) predictions inform low-level (phoneme-in-word) predictions, supporting hierarchical prediction. Together, the results demonstrate that language processing is inherently predictive, showing that the brain spontaneously predicts upcoming language at multiple levels of abstraction – even when passively listening to something as complex as a novel.

In **Chapter 5** I use the same datasets and modelling framework to test a popular neurophysiological hypothesis: that top-down probabilistic predictions are signalled via oscillations in the beta (12-30 Hz) and alpha (8-12 Hz) range. I test this by modelling contextual predictions about the incoming words on a phoneme-by-phoneme basis. In line with this idea, I find that prior confidence in the prediction about the incoming word is related to beta band amplitude. However, and more preliminarily, the effect seems opposite to what I expected: pre-stimulus beta was *weaker* when prior predictions were stronger. I discuss how this relates to the empirical literature (potentially highlighting a weakness in the original hypothesis), and outline ways in which the results in this chapter could be further strengthened.

In **Chapter 6**, I return to reading, this time focussing on eye movements in reading. The literature suggests that how long a word is looked at, and whether it is fixated at all (skipping), depends on both the extent to which a word could be *predicted* from context and discerned from a *parafoveal preview*. In this chapter, I estimate the relative importance of these two sources of information. I address this question in natural reading, combining deep neural network and Bayesian ideal observer modelling to quantify prediction and preview from moment to moment. Surprisingly, the most interesting dissociation was not between prediction and preview, but between skipping and reading times. For skipping, neither prediction nor preview was important - the vast majority of skipping was explained by a simple oculomotor model. For reading times, by contrast, we found clear and roughly equal (but independent) effects of prediction and preview. Together, the results challenge dominant models of eye movements in reading, by showing that skipping is driven by low-level factors. However, they also reveal a limit on hierarchical prediction by showing that predictions based on linguistic context do not inform parafoveal preview – highlighting a difference between reading and speech perception (cf. **Chapter 4**).

Finally, in **Chapter 7**, I discuss and synthesise the empirical findings in this thesis, highlight the remaining open questions, and discuss avenues for future research.

Chapter 2

Word contexts enhance the neural representation of individual letters in early visual cortex

Abstract

Visual context facilitates perception, but how this is neurally implemented remains unclear. One example of contextual facilitation is found in reading, where letters are more easily identified when embedded in a word. Bottom-up models explain this word advantage as a post-perceptual decision bias, while top-down models propose that word contexts enhance perception itself. Here, we arbitrate between these accounts by presenting words and nonwords and probing the representational fidelity of individual letters using functional magnetic resonance imaging. In line with top-down models, we find that word contexts enhance letter representations in early visual cortex. Moreover, we observe increased coupling between letter information in visual cortex and brain activity in key areas of the reading network, suggesting these areas may be the source of the enhancement. Our results provide evidence for top-down representational enhancement in word recognition, demonstrating that word contexts can modulate perceptual processing already at the earliest visual regions.

This chapter is based on:
Heilbron M, Richter, D., Ekman, M., Hagoort P, de Lange FP. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications 11(1)*, 1-11.

Introduction

Context-based expectations can strongly facilitate perception, but how this is neurally implemented remains a topic of debate (Bar et al., 2006; de Lange, Heilbron, and Kok, 2018). One famous and striking example of contextual facilitation is found in reading, where letters are more easily identified when embedded in a linguistic context such as a word or name (e.g., a road sign) than in a random string (e.g., a license plate; Cattell, 1886).

Historically, two opposing accounts have been proposed to explain this so called 'word superiority effect'. Under the guessing-based account, letter identification occurs in a bottom-up fashion and the advantage offered by words constitutes only a post-perceptual advantage in 'guessing' the correct letter (Paap et al., 1982; Thompson and Massaro, 1973). Alternatively, the perceptual account explains word superiority as a top-down effect, proposing that higher-order linguistic knowledge can enhance perceptual processing of the individual letters (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982). A rich behavioural literature, dating back several decades(Reicher, 1969; Wheeler, 1970) has documented that even when the ability to guess the correct letter is experimentally controlled, the word advantage persists (Balota, Yap, and Cortese, 2006). This has been interpreted as evidence that the effect must (at least in part) reflect top-down perceptual enhancement – a view that remains dominant until today (Dehaene, 2009).

However, some lingering doubts have persisted. For instance, ideal observer analysis has shown that the efficiency of letter recognition is much lower than that of a fully holistic (word-based) observer, and lies within the theoretical limits of a strictly letter-based (feedforward) observer – even when considering word superiority (Pelli, Farell, and Moore, 2003). Moreover, advances in deep learning have shown that letters and other complex objects can be accurately recognized in context by bottom-up architectures, further questioning the need to invoke top-down explanations (LeCun, Bengio, and Hinton, 2015). Beyond these theoretical arguments, neural evidence for the perceptual locus of this supposedly top-down effect is lacking. This is remarkable, since the top-down interpretation of word superiority makes a clear neural prediction: if the behavioural word advantage is due to a perceptual enhancement of letter stimuli, then it should be accompanied by an enhancement of sensory information in the early visual areas that process the individual letters already.

Here, we test this prediction using a simple paradigm involving streams of words and nonwords. We use neural network simulations of the paradigm to confirm that top-down models would uniquely predict the enhancement of letter representations by word contexts. When we then perform the same experiment in human observers while recording brain responses using functional magnetic resonance imag-

ing (fMRI), we find that word contexts robustly enhance letter representations in early visual cortex. Moreover, compared to nonwords, words are associated with increased information-activation coupling between letter information in early visual cortex on the one hand, and blood-oxygen-level-dependent (BOLD) activity in key areas of the reading network on the other. These results suggest that word superiority is (at least in part) a perceptual effect, supporting prominent top-down models of word-recognition.

Results

Word contexts facilitate orthographic decisions

Participants (n=34) were presented with streams of words or nonwords consisting of five letters (see Fig. 1a), while maintaining fixation. We used a blocked design in which word and nonword (i.e. unpronounceable letter string) stimuli were presented in long trials of 10 items of which the middle letter (U or N) was kept fixed while the outer letters varied, creating a word or nonword context (each 10s trial containing only stimuli of one condition). To make reading visually challenging, stimuli were embedded in Gaussian noise (see *Methods*). To keep participants engaged, they performed an spelling discrimination task on specific target stimuli that occurred occasionally (1-2 times) per trial. Target stimuli were learned during a prior training session. Targets were presented either in their regular form or with one letter permuted, and participants had to categorize targets as 'spelled' correctly or incorrectly (i.e. presented in the learned form or permuted).

Participants were faster (median RT difference: -29.2 ms; Wilcoxon signed rank, $T_{34} = 40, p = 1.07 \times 10^{-5}, r = 0.87$) but not significantly more accurate (mean accuracy difference: 1.62%; t-test, $t_34 = 1.70, p = 0.098, d = 0.29$) for word compared to nonword targets. This observation is in line with the word superiority effect, but from the behaviour alone it is unclear whether the word advantage was perceptual or post-perceptual.

Representational enhancement is a hallmark of top-down models

Because our paradigm is different from the traditional paradigms in the (behavioural) word superiority literature, we performed simulations of our experiment to confirm that the top-down account indeed predicts the representational enhancement we set out to detect. We used a predictive coding implementation (Spratling, 2016) of the influential Interactive Activation architecture proposed by McClelland and Rumelhart (1981) (see *Methods*).

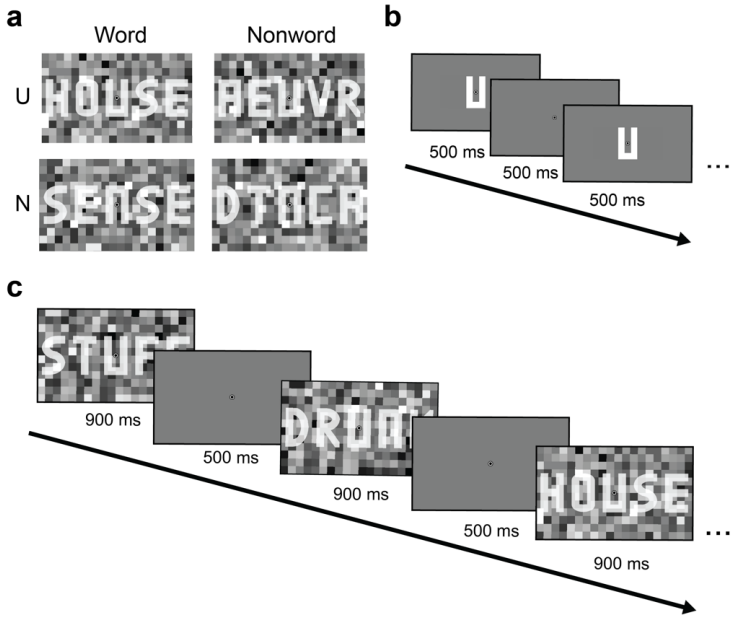

Figure 2.1. Experimental paradigm. (a) Example stimuli for each condition. Participants observed words or nonwords (i.e. orthographically illegal, unpronounceable strings) with a U or N as middle letter, resulting in four conditions. **(b)** Functional localiser. During the functional localiser, the key letters (U and N) were presented in isolation and without visual noise, while participants performed an irrelevant task at fixation. **(c)** Trial structure. We used a blocked design, in which each 14 s trial consisted of 10 words or nonwords with a fixed middle letter. Participants performed an orthographic discrimination task on specific, prelearned targets that occurred once or occasionally twice per trial. Participants were trained in a separate session to perform the task while maintaining fixation at the centre of the screen.

In the simulation, we ran artificial 'runs' in which we presented sets of word and nonword stimuli used in the experiment to the network (Figure 2.2a). To simulate experimental viewing conditions, we added Gaussian noise and ran the network until convergence so as to mimic long stimulus duration (see *Methods*) resulting in stimuli that were presented well-above recognition threshold (Figure S2.2). Representational strength was quantified by dividing the activity level for the correct letter unit by the sum of activity levels of all letters – a fraction that asymptotically goes to 1 as representational strength increases. After running 34 simulated runs with the top-down model, the relative evidence for the middle letter was confirmed to be much higher in words than nonwords (paired t-test, $t_34 = 50.5$, $p = 7.72 \times 10^{-33}$), despite the signal-to-noise ratio of the simulated stimuli being identical (Fig 2.3a). Importantly, when the same stimuli were presented to a network lacking word-to-letter feedback connections, no such difference was found (paired t-test, $t_34 = -0.24$,

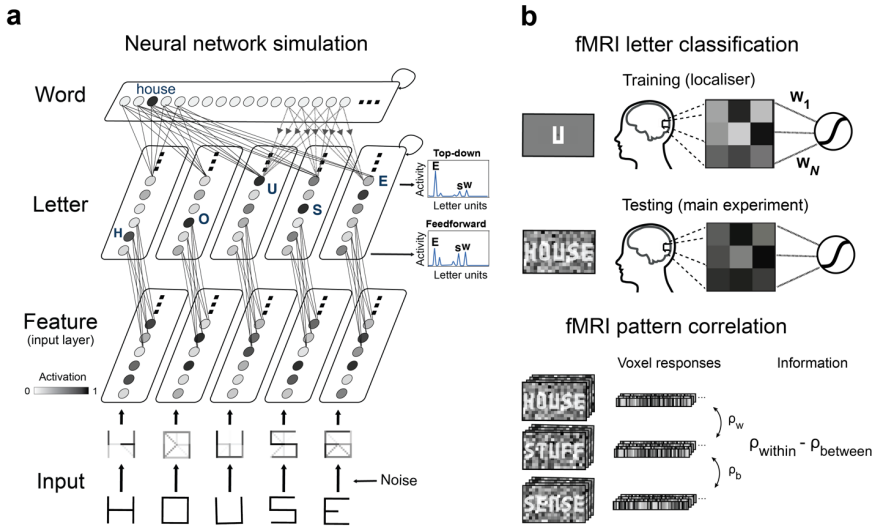

Figure 2.2. Probing representational enhancement in neural network models and the brain. **(a)** Modelling representational enhancement in a hierarchical neural network model (McClelland and Rumelhart, 1981; Spratling, 2016). Stimuli used in the experiment were encoded into vectors of visual features and overlaid with Gaussian noise (bottom rows). Inputs were presented to a network with or without word-to-letter feedback connections. For both networks, representational strength was quantified from the distribution of activity levels of letter units for the third position (principle illustrated for the fifth letter, E). Solid circles indicate units (representing features, letters or words); lines indicate feedforward connections, and dotted lines with arrows indicate feedback connections. Note that we used a predictive coding formulation of the network (Spratling, 2016) but for simplicity only the state estimator (prediction) units are shown in the schematic (see Methods for details) **(b)** Quantifying representational letter enhancement using multivariate pattern analysis. To probe letter representations in the brain, we used two multivariate pattern analysis (MVPA) techniques: classification (upper panel) and pattern correlation (lower panel).

$p = 0.81$), resulting in a significant interaction (two-sample t-test $t_{34} = 31.1$, $p = 3.5 \times 10^{-}41$). This confirmed that despite the differences between our and the classic paradigm, representational enhancement of letters by word contexts is a hallmark of top-down models of letter perception.

Word contexts enhance letter representations

Next, we tested whether we could find a similar enhancement effect in early visual cortex in our participants. To do so, we first trained a classifier for each participant on an independent dataset from a functional localiser run, during which the two middle letters (U or N) were presented in isolation and without Gaussian noise (see Figure 2.2a). We then tested the classifier's ability to identify the middle letter of the words

and non-words presented in the main experiment, in a trial-based fashion (each trial lasting 14 s and consisting of 10 stimuli). We reasoned that if word context enhances the *sensory* representations of letters (e.g. enhancing the letter features in noise), this should be apparent in early visual areas, which we defined as the union of V1 and V2 (see *Methods).* To focus on voxels sensitive to the relevant part of the visual field, we selected the 200 voxels (the same number we used in a previous study(Richter et al., 2018)) most responsive during the localiser run. We were able to classify letter identity well above chance level (one sample t-test, $t_34 = 18.84$, $p = 3.13 \times 10^-19$, d = 3.23) reaching a mean overall decoding accuracy of 81.4% averaged over both conditions (see Figure 2.3).

Having established that letter identity can be extracted with high fidelity from early visual cortex, we went on to test if representational content was enhanced by word context. Strikingly, we found that classification accuracy was indeed higher for words compared to nonwords (Wilcoxon sign rank test, $T_34 = 141.5$, $p = 7.55 \times 10^{-3}$, r = 0.52; Fig 2.3b). To further examine this enhancement effect, we quantified representational content using an (arguably simpler) supplementary multi-voxel pattern analysis (MVPA) technique: pattern correlation analysis – the difference in voxel response pattern correlation that could be attributed to letter identity ('Pearson ρ within-letter' minus 'Pearson ρ between-letter'; see *Methods*). Reassuringly, the results aligned with those of the classification analysis: the correlation difference score being significantly higher for words than nonwords (Wilcoxon sign rank, T_34 = 103, $p = 8.83 \times 10^{-4}$, r = 0.67).

To confirm that the differences revealed by the classification and pattern correlation analysis were related to differences in representations of stimulus information and not to unrelated confounding factors, we performed a number of controls. First, we tested the stability of the results over different ROI definitions. Since both representational analyses used the 200 voxels that were most responsive during an independent functional localiser, we wished to ensure that the results were not unique to this a priori specified (but arbitrary) number. We therefore re-ran the same analyses for ROIs ranging from 50 to 1000 voxels with steps of 10. This revealed that the same pattern of effects was found over practically the entire range of ROI sizes **(Fig S2.3).

Another possibility is that the increased estimates of representational content could be explained by a simple difference in signal amplitude, potentially related to participants being more attentive to words than nonwords. To address this, we quantified BOLD amplitude per condition using a standard GLM-based approach (see *Methods*) but found no significant difference between conditions in the amplitude estimates for the corresponding voxels (paired t-test, $t_34 = -0.57$, $p = 0.57$, $d = 0.10$; Bayesian paired t-test, $BF_10 = 0.21$; see Figure S4). Importantly, we found no significant differences in eye-movement deviation from fixation between words

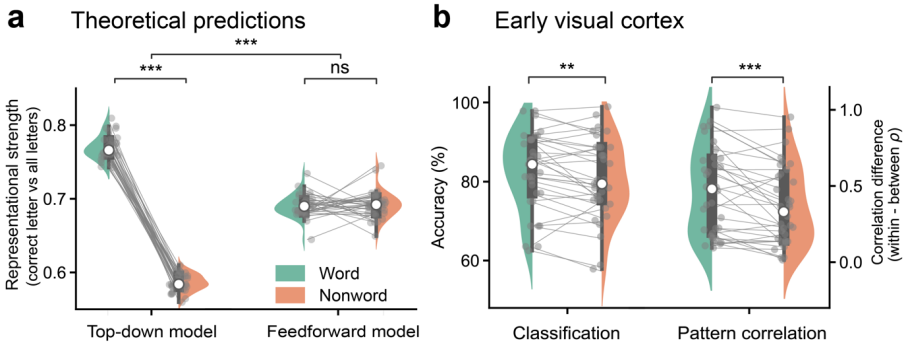

Figure 2.3. Word contexts enhance letter representations. **(a)** Theoretical predictions. We simulated 34 artificial 'runs' in which we exposed a network with (top-down model) and without (feedforward model) word-to-letter feedback connections to the experimental stimuli, and computed the average representational strength of the middle letter in word and nonword contexts. Note that the strong dissociation was observed despite the fact that the middle letter was well-above threshold in all conditions for both models (S2.2). **(b)** Letter representation in early visual cortex of 34 human observers. Two multivariate pattern analysis methods (see Methods) revealed that neural representations of letters were enhanced in word compared to nonword contexts, supporting the top-down model. In both panels, grey dots represent individual simulated 'runs' (a) or individual participants (b). Lines represent paired differences. White dots, boxes and whiskers represent between-subject medians, quartiles and 1.5 interquartile ranges, respectively. Significance levels correspond to $p < 0.01$ (**) or $p < 0.001$ (***) in a paired, 2-tailed Student's t or Wilcoxon sign rank test.

and nonwords (Wilcoxon T 32 = 197, $p = 0.21, r = 0.25$; Bayesian paired t-test, $BF_10 = 0.48$; see Figure S6 and *Methods*), confirming participants' ability to maintain fixation during the task did not differ significantly between conditions.

As a final control analysis, we wanted to confirm that the MVPA results relied on retinotopically specific information. This would be an important indication that both the letter information extracted from visual cortex, and its enhancement by word contexts, indeed originate from sensory representations. To this end, we performed a searchlight variant of the classification and pattern correlation analysis (see Supplement for details). This revealed (see Figures S2.7 and S2.8) that letter identity information was only visible in neural activity patterns in visual cortex, ruling out that decoding relied on a brain-wide signal. We further tested for retinotopic specificity within visual cortex by comparing the functionally defined central ROI (described above), to a functionally defined peripheral ROI (see *Methods* and *Supplementary Note 1* for more details). This revealed (Figure S2.9) that overall letter decoding was greatly reduced for the peripheral ROI compared to the central ROI, both for classification (paired t-test, $t_{34} = 15.59$, $p = 8.86 \times 10^{-17}$, $d = 2.67$) and pattern correlation analysis (paired t-test, $t_34 = 8.06$, $p = 2.65 \times 10^{-9}$, $d = 1.38$).

Importantly, we found a similar reduction in the peripheral ROI for the enhancement effect (the difference in decoding between conditions), again both for the classification (paired t-test, $t_{34} = 2.56, p = 0.015, d = 0.44$) and pattern correlation analysis (paired t-test, $t_{34} = 2.92, p = 6.31 \times 10^{-3}$, $d = 0.50$).

In sum, these analyses show that sensory letter information in early visual cortex, as estimated by classification and pattern correlation analysis, was increased in words compared to nonwords. This enhancement was present over a range of ROI definitions, but was reduced for peripheral compared to central ROIs, and could not be explained by confounding factors such as BOLD amplitude or eye-movements.

Representational enhancement across the visual hierarchy

Having established a perceptual enhancement effect by word context in early visual cortex, we then asked how this enhancement effect was distributed among specific visual areas. To this end, we further investigated 5 ROIs, four of which were defined anatomically (V1-V4) and one (VWFA) functionally; in each ROI, voxels were selected using the procedure described earlier (see *Methods* for details).

The results show consistent evidence for word enhancement in V1, V2 and V4 (all p's< 0.025; see Figure S2.10 for details), with both analyses. In contrast, V3 and VWFA showed no consistent evidence for word enhancement (see Figure S2.10). However, in these regions the overall classification accuracy and pattern information scores were also close to chance, making the absence of differences between conditions difficult to interpret. For regions V1-V4, we also tested for univariate amplitude differences between word and nonword conditions. Interestingly, in all four regions the sign of the univariate difference was negative (indicating weaker amplitude of responses to word stimuli), but note that only in V4 this difference was marginally significant (paired t-test, $t_{34} = 2.11$, $p = 0.04$,$d = -0.36$, uncorrected; Figure S2.5). In sum, we observed word enhancement across multiple regions in the visual hierarchy. Critically, none of the regions showed BOLD amplitude differences, ruling out the possibility that word enhancement was confounded by low-level attentional differences between conditions.

Information-activation coupling reveals putative neural sources

Having observed a hallmark of top-down perceptual enhancement by word contexts, we then asked what the potential neural source of this top-down effect could be. We reasoned that if a candidate brain region was involved in the observed enhancement, then activity levels in this region would be expected to covary with the amount of letter information represented in early visual cortex. Moreover, this relationship should not be driven by a categorical difference between conditions (e.g. that both

BOLD amplitude in a candidate region and informational content in visual cortex are higher for words than nonwords, while the two are not related within conditions). Taking the two requirements together, we expected regions implicated in the top-down effect to show *increased functional coupling* between local BOLD activity and representational information in early visual cortex, for words compared to nonwords.

To test for this increased information-activation coupling, we used a GLM-based approach to model regional BOLD amplitude in both conditions as a function of early visual cortex classification evidence, and tested for an increased slope for words compared to nonwords (see Figure 2.4b). This is analogous to the well-established PPI analysis (Friston et al., 1997) but uses classifier evidence instead of BOLD activity as the 'seed' time course. Classifier evidence here corresponds to the predicted probability of the correct (presented) letter stimulus for each brain volume (TR) (see *Methods*).

We first tested for increased coupling in a hypothesis-driven, ROI-based fashion. We tested two candidate regions: the visual wordform area (VWFA) and the left posterior middle temporal gyrus (pMTG), associated with orthographic/visual (Cohen et al., 2000; Dehaene and Cohen, 2011)) and lexical/semantic processing (Davey et al., 2016; Turken and Dronkers, 2011) respectively. Activity of all voxels was averaged to obtain a single BOLD time course per ROI. This BOLD timecourse was then modelled as a function of visual cortex classification strength to obtain separate coupling parameters for word and nonword conditions. We indeed observed a significantly increased coupling in both VWFA (Wilcoxon sign rank, $T_{34} = 80$, $p = 2.00 \times 10^{-4}$, $r = 0.73$) and pMTG (paired t-test $t_34 = 2.83$, $p = 8.2 \times 10^{-3}$, $d = 0.48$; see Figure 2.4a**) for words. The increase in coupling appeared stronger in VWFA, but the difference in effects between regions was not statistically significant (paired t-test, $t_{34} = 0.62$, $p = 0.54$, $d = 0.11$). Finally, we carried out an exploratory analysis by testing for increased functional coupling across the entire brain. In essence, the GLM procedure was identical to the one above but carried out at the individual voxel level. This yielded, for each participant, a map of estimated *changes* in functional coupling for every voxel. These functional maps were then registered to a standard space after which we tested whether there were clusters of voxels that showed an increase in functional coupling for words compared to nonwords. We found two significant (FWE-corrected, cluster-forming $P < 0.001$, cluster-level $P < 0.05$) left-lateralised clusters at key nodes of the language network: one in pMTG and one in IFG (Figure 2.4c; Figure S2.11). No significant cluster was found at VWFA, possibly due to individual neuro-anatomical variability in VWFA size and location (Glezer and Riesenhuber, 2013).

Altogether, these results demonstrate increased functional coupling between visual cortical classification evidence and neural activity in VWFA, pMTG and IFG. In all of these regions, we found a significant increase in functional coupling (here,

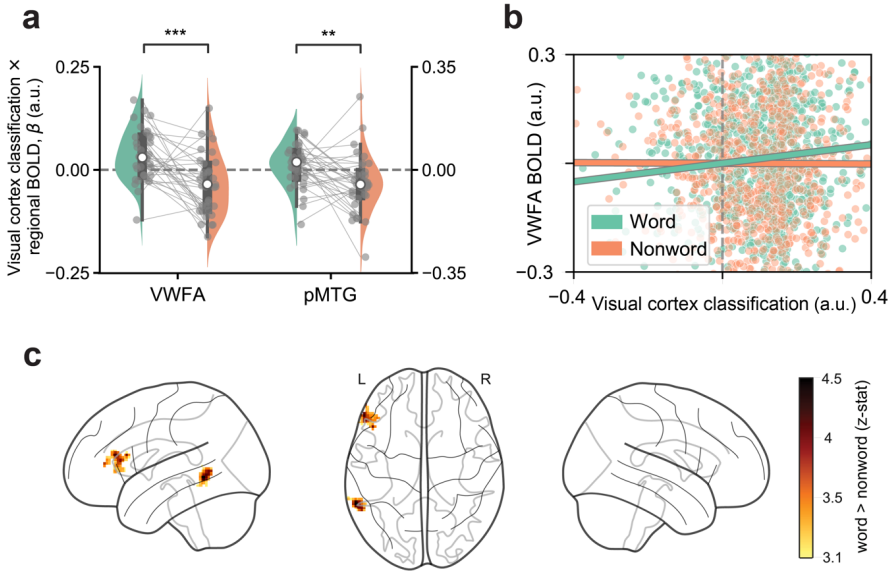

Figure 2.4. Information-activation coupling analysis. **(a)** ROI-base coupling analysis. For two ROIs, GLMs were fitted to estimate coupling between early visual cortex classification evidence and regional BOLD amplitude for words and nonwords separately. We then tested for increased coupling (higher coefficients) in words compared to nonwords. **(b)** Illustration for example participant. A single (averaged) time course was extracted from each ROI and regressed against visual cortex classification evidence to test for increased slopes in words compared to nonwords. For illustration purposes, only the predicted slopes based on the regressor of interest are shown. Note that classification evidence was defined as probability, but here expressed in arbitrary units due to the whitening operation. **(c)** Whole brain results. Same analysis as in **(a)** and visualised in **(b)** but performed for each voxel independently. Resulting contrast images (word-nonword) were tested at the group level for increases in coupling in words compared to nonwords. This revealed statistically significant clusters ($P < 0.05$ FWE corrected), in the left pMTG and in the left IFG. Glass brain plot rendered with nilearn (Abraham et al., 2014). For a non-thresholded slice-by-slice rendering of the whole brain results in panel **(c)**, see Figure S2.11. Grey dots indicate coefficients of individual participants, and lines the within-subject differences; white dots, boxes and whiskers are between-subject medians, quartiles and interquartile ranges, respectively. Significance levels correspond to $p < 0.01$ (**) or $p < 0.001$ (***) in a paired, 2-tailed Student's t or Wilcoxon sign rank test. *pMTG*, posterior Medial Temporal Gyrus; *IFG*, Inferior Frontal Gyrus.

meaning that classifier evidence increased when the regions became more active, and vice versa) for words compared to nonwords, which is consistent with the idea that these regions might constitute the neural source of the top-down effect.

Discussion

Visual context facilitates perception (Bar, 2004). Letter perception offers a striking example of such facilitation, as letters are more easily recognised when embedded in a word. Dominant, 'interactive' models of word recognition assume this facilitation occurs in the visual system already, proposing that linguistic knowledge can enhance perception in a top-down fashion (McClelland and Rumelhart, 1981). Here we tested this perceptual enhancement hypothesis for the first time at the neural level. We presented streams of words or nonwords with a fixed middle letter while recording fMRI. Simulations of this novel paradigm confirmed that top-down models of word recognition uniquely predict that perceptual representations of the middle letter should be enhanced when embedded in a word. In line with the top-down account, information about the middle letter, probed using multivariate pattern analysis in early visual cortex, was enhanced when the letter was embedded in words compared to nonwords. Further, we found increased functional coupling between the informational pattern in early visual cortex, and regional BOLD amplitude in three key regions of the left-lateralized language network, i.e. VWFA, left pMTG and IFG. This points to these regions as potential neural sources of the representational enhancement effect. Together, these results constitute the first neural evidence for representational enhancement of letters by word contexts, as hypothesized by top-down accounts of word recognition (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982). The results also fit naturally with theoretical frameworks of top-down perceptual inference, such as hierarchical predictive coding (Friston, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999; see de Lange, Heilbron, and Kok, 2018; Heilbron and Chait, 2018 for review) and with the broader literature on top-down, predictive effects in language processing (Davis and Johnsrude, 2007; Kuperberg and Jaeger, 2016).

Our results are in line with a large behavioural literature on context effects in letter perception that support interactive activation (top-down) models. These works have demonstrated, for instance, that the word advantage persists when the guessing advantage afforded by words is constrained experimentally (Reicher, 1969; Wheeler, 1970 see Balota, Yap, and Cortese, 2006 for review); that readers subjectively perceive letters embedded in real words as sharper(Lupyan, 2017); and that readers are better at detecting subtle perceptual changes in real words than in nonwords (Lupyan, 2017). While the behavioural literature has extensively investigated top-down effects, work

on the neural basis of visual word recognition has focussed almost exclusively on its bottom-up component, most notably by probing the bottom-up selectivity profile of VWFA (or ventral occipitotemporal cortex more broadly) to various visual and orthographic properties (Dehaene et al., 2005; Glezer, Jiang, and Riesenhuber, 2009; Vinckier et al., 2007). One study tried to disentangle word and letter encoding at a neural level (Thesen et al., 2012), but did not probe individual letter representations and their enhancement by word contexts. A recent study did investigate top-down processing, but was limited to attention-based response modulations in decision contexts (Kay and Yeatman, 2017).

Beyond the domain of language, but converging with the results presented here, are results from object perception, where it was recently found that the facilitation of object recognition by familiar contexts was accompanied by enhancement of object representations in object selective cortex (Brandman and Peelen, 2017). The similarity to the current findings speaks to the idea that representational enhancement reflects a more general principle of contextual effects in perception. These contextual effects have been extensively studied, and range from neurons in macaque early visual cortex responding differently to identical lines when presented as parts of different figures (Zhou, Friedman, and Heydt, 2000; Zipser, Lamme, and Schiller, 1996) to neurons in mouse V1 encoding contextually expected but omitted stimuli (see Khan and Hofer, 2018, for review). As such, it may be that although here the context is linguistic, the contextual effect in visual cortex reflects a more general mechanism that is not specific to reading or unique to humans. Interestingly, the idea that contextual enhancement reflects a more general perceptual mechanism was a key motivation to develop models of word recognition in the first place (McClelland and Rumelhart, 1981).

When viewed as a more general principle of perception, contextual enhancement touches on an even broader question: are objects recognised by their parts or as wholes? On the one hand, word superiority has historically been taken as an example of 'holistic' perception (Balota, Yap, and Cortese, 2006) and the enhancement we observed (in which 'wholes' enhance representations of 'parts') indeed seems to contradict a strictly letter-based (part-based) account. But on the other hand, it has been convincingly demonstrated, both for word and face recognition (Martelli, Majaj, and Pelli, 2005; Pelli, Farell, and Moore, 2003) that the identifiability of parts poses a bottleneck on the identification of wholes, and hence, even for the most common words, recognition cannot be truly holistic(Pelli et al., 2003). Moreover, effects of wholes on the identification of parts are not always faciliatory: facial arrangements, for instance, have been both reported to have positive and negative effects on search performance (Suzuki and Cavanagh, 1995; Wenger and Townsend, 2006). Developing a theoretical framework that naturally accounts for top-down, contextual

enhancement as reported here (see e.g. Lee and Mumford, 2003) while being properly constrained so as to incorporate feature-based bottlenecks, and the occasional detrimental effects of context, provides an important challenge for future research.

A limitation of the current study is that we cannot access sensory representations directly, and instead have to infer them by estimating sensory information from measured neural activity patterns. By itself, the fact that letter identity could be more readily decoded from words than nonwords could in principle merely reflect confounding differences in the BOLD signal between conditions, or in the our ability to extract information from that signal (Smith, Kosillo, and Williams, 2011). Importantly, however, we obtained converging evidence using two complementary techniques of probing representational content. One of these (classification) used an independent data set for training purposes in which only single letters were presented without noise, which suggests that our MVPA techniques were picking up relevant information about the middle letters, rather than irrelevant signals that only covaried with reading (non)words with the respective middle letter. Moreover, enhancement was consistently found in multiple visual areas, was retinotopically specific, but not contingent on exact ROI definitions, and could not be explained by other confounds such as signal amplitude or eye movements. As such, we believe that the most parsimonious explanation of the observed effect is as reflecting an enhancement in the underlying sensory information available to visual cortex itself – in other words, a representational enhancement.

We interpret this representational enhancement as a neural signature of the perceptual enhancement of letters – a process formalised by top-down models of word recognition, and widely characterised in behavioural literature (Balota, Yap, and Cortese, 2006; Lupyan, 2017). However, a limitation of fMRI is that it cannot differentiate between earlier and later activity. Hence it is possible that the observed effects arise late, perhaps even much later than what is typically considered 'perceptual' (e.g. > 400 ms); instead perhaps reflecting what one might call iconic memory encoding. Although distinguishing between perceptual and post-perceptual effects is notoriously difficult, future studies might address this by probing perception more directly using an objective measure of perceptual sensitivity, or by using a high temporal resolution method (e.g. ECoG or MEG) in combination with a temporal criterion to arbitrate between perceptual and post-perceptual enhancement of sensory representations.

An apparent disconnect between our study and the existing literature concerns the level of representation at which enhancement occurs. We probed enhancement in early visual cortex (representing visual features such as edges and simple line conjunctions) while in theoretical models (McClelland and Rumelhart, 1981; Spratling, 2016) enhancement is probed at the level of letters, not features. However, percep-

tual enhancement is a generic mechanism and should not be unique to a specific level of representation. In fact, the main reason (McClelland and Rumelhart, 1981, p.378) that in the classic models, enhancement occurs only at the letter level is simplicity: because features comprise the input to the network and are hence not *recognized*, the possibility of enhancement occurring at the feature level is excluded by design.

Apart from model simplicity, one might argue there are more substantive, cognitive reasons that enhancement is primarily described at the letter level and not at the level of simple features. Specifically, word superiority has been reported with stimuli consisting of mixed case and font (Adams, 1979), implying that sometimes the word superiority effect can be independent of the visual features that define exact letter shape, and may instead act at a more abstract level of letter identity. However, when the exact letter shape is well known and especially under visually noisy conditions (like in our experiment), enhancement of simple low-level visual features appears useful for letter recognition, thereby incentivizing top-down enhancement to reach the (functionally well-localised) visual feature level. As such, we do not claim that our experiment shows that word superiority *always* acts at the level of sensory features. Rather, it demonstrates that in principle these enhancement effects can extend even to the earliest sensory cortical regions, contradicting purely bottom-up accounts in which such top-down enhancement is ruled out by design. What kind of information is driving the observed representational enhancement effect? A possible source of information is lexical knowledge, although sublexical (orthographic/phonological) knowledge may be an equally plausible candidate. Indeed, behaviourally letters are also more easily recognised when embedded in pronounceable nonwords (pseudowords) than in unpronounceable, orthographically illegal nonwords (Baron and Thurston, 1973). Theoretically, such a pseudoword superiority effect can be either understood as originating via top-down connections in a dedicated sublexical route (Coltheart et al., 2001) or as arising as a by-product from co-activations of lexical items with overlapping letter combinations (McClelland and Rumelhart, 1981; Seidenberg and McClelland, 1989). The fact that we found increased functional coupling in both VWFA (associated with sublexical orthography) and pMTG (associated with lexical access) makes our data consistent with both types of feedback, originating from the sublexical and the lexical route. Future studies could examine the relative contributions of these forms of feedback, by examining the neural activity patterns to pronounceable pseudowords.

A final point of discussion concerns the interpretation of the information-activation coupling analysis. We interpret the results as pointing to putative candidate sources, because the observed increases in functional coupling match the expected pattern of results if the associated regions were indeed a neural source of the enhancement. However, we acknowledge that since the functional coupling analysis is correlational

in nature, the direction of causality implied by this interpretation remains speculative. To get a better understanding of the sources involved this effect, future studies could either directly perturb candidate sources, use a more indirect method for inferring directionality such as laminar fMRI (Lawrence et al., 2017) or a directional connectivity analysis (Friston, Moran, and Seth, 2013).

In conclusion, we have observed that word contexts can enhance sensory letter representations in early visual cortex. These results provide the first neural evidence for top-down enhancement of sensory letter representations by word contexts, and suggest that readers can better identify letters in context because they might, quite literally, see them better.

Methods

Participants

Thirty-six participants were recruited from the participant pool at the Donders Centre for Cognitive Neuroimaging. Sample size was chosen to detect a within-subject effect of at least medium size ($d > 0.5$) with 80% power using a two-tailed one-sample or paired t-test. The study was in accordance with the institutional guidelines of the local ethical committee (CMO region Arnhem-Nijmegen, The Netherlands, Protocol CMO2014/288), all participants gave informed consent and received monetary compensation. Participants were invited for an fMRI session and a prior behavioural training session, that took place no more than 24 hours before the fMRI session. For one participant, who moved excessively between runs, decoding accuracy was never above chance; this participant was excluded from all fMRI analyses. One additional participant had their eyes closed for an extended duration during more than 20 trials, and was excluded from both the behavioural and fMRI analyses. All remaining participants (n=34, 12 male, mean age = 23 ± 3.32) were included in all analyses. Due to technical problems, one participant only completed 4 instead of 6 blocks, all of which were analysed.

Stimuli

Stimuli were generated using Psychtoolbox-3 (Kleiner et al., 2007), running on MATLAB (MathWorks, MA, USA). Stimuli were rear-projected using a calibrated EIKI (EIKI, Rancho Santa Margarita, CA) LC XL 100 projector (1024 x 768, 60 Hz). Each stimulus was a 5 letter word or nonword presented in a custom-made monospaced typeface. To prevent that the multivariate analyses would pick up on global low-level features (such as overall luminance or contrast) to discriminate between middle letter identity, the middle characters (U or N) were chosen to be identical in shape and

size, but flipped vertically with respect to each other. Words were presented in a large font size, each letter $3.6°$ wide and with $0.6°$ spacing between letters. This size was chosen to make the middle letter as large as possible while retaining readability of all letters when fixating at the centre. In addition to the words and nonwords, a fixation dot of $0.8°$ in diameter was presented at the centre of the screen.

To make reading visually challenging and incentivize top-down enhancement of low-level visual features, words were embedded in visual noise. The noise consisted of pixelated squares, each 1.2 deg wide, offset so that the pixels were misaligned with the letter strokes. Letters were presented on top of the noise with 80% opacity. We chose this type of noise after finding it impacted readability strongly even when the letters were presented at high physical luminance. Brightness values (in the range 0-255) of the noise 'pixels' were randomly sampled from a Gaussian distribution with a mean of 128 and an SD of 50. To make sure that the local brightness was on average identical for each trial and across the screen, the noise patches were generated using a pseudo-random procedure. In each trial, 10 noise patches were presented, 5 of which were independent and randomly generated, while the other 5 were copies of the random patches, but polarity-inverted in terms of their relative brightness with respect to the mean. This way the brightness of each noise pixel was always 128 (grey) on average in each trial. The order of noise patches was pseudo-random, with the constraint that copied patches were never presented directly before or after their original noise patch. This way the re-use of noise patches was not noticeable and all patches seemed randomly sampled anew.

In the main experiment, we used a blocked design, in which we presented blocks of four long trials (one of each of the four conditions), followed by a null-trial. Each trial was 14 s long, during which 10 stimuli were presented. Of those stimuli, 9 or occasionally (in 25% of trials) 8 were (non-)word items and 1 or 2 were (learned) targets. A single presentation consisted of 900 ms of (non)word item plus noise background, and 500 ms of blank screen plus fixation dot (Figure 1c). Targets were either presented in their regular (learned) form or with one of the non-middle letters permuted, and participants had to discriminate whether the target was regular or permuted. Target correctness and occurrence within the trial were counterbalanced and randomised, with the constraint that targets were never presented directly after each other. The order of word items was shuffled pseudo-randomly, with the constraint that the same letter never repeated twice at the same position (except for the middle letter).

In the functional localiser run, only the middle letters (U and N) plus fixation bulls' eye were presented. We again used a blocked design, with long trials that had a duration of 14 s during which one of letters was repeated at 1 Hz (500 ms on, 500 ms off; see Figure 2.1b). During the localiser, each trial was followed by a null-trial

in which only the fixation dot was presented for 9.8 seconds. This was repeated 18 times for each letter.

Two different sets of words and nonwords were used for the training and experimental session. For the experimental session, we used 100 5-letter words with a U or N as third character in Dutch (see Table S2.1), plus equally many nonword items. This particular subset was chosen because they were the 100 most frequent 5 letter words with a U or N in Dutch, according to the subtlex database (Keuleers et al., 2010). Each item occurred at least 4 times and maximally 5 times (4.2 on average) during the entire experimental session; to ensure repetitions were roughly equally spaced, items were only repeated once all other items were presented equally often. Because we wanted to familiarise participants with the task and the custom-font, but not with the (non)word stimuli themselves (especially because there was considerable variation in the amount of training between participants), we used different (non)words for the training session. For the training session, we used the remaining 50 less frequent 5 letter Dutch words with a U and N. For the nonwords, letters were randomly sampled according to the natural frequency of letters in written Dutch (Broecke, 1988), with the constraint that adjacent letters were never identical. The resulting nonwords were then hand-selected to ensure all created strings we unpronounceable, orthographically illegal nonwords. The four learned target stimuli were CLUBS and ERNST for the words, and KBUOT and AONKL for the nonwords. These were learned during the prior training session.

Procedure

Each participant performed one behavioural training and one experimental fMRI session. The goal of the training was for participants to learn the 4 target items and learn how to perform the task while maintaining fixation at the centre of the screen. The fMRI session consisted of a brief practice of 5 minutes during which the anatomical scan was acquired. This was followed by 6 experimental runs of 9-10 minutes, which were followed by a localiser run of 15 minutes. We used a blocked design, in which we presented blocks of four long trials (one of each of the four conditions), followed by a null-trial experimental run consisted of 40 trials of 14 s. Trials were presented in blocks consisting of 5 trials: one of each condition (U-word, U-nonword, N-word,N-nonword), plus a null trial during which only the fixation dot was present. The order of trial types within blocks was randomised and equalised: over the entire experiment, each order was presented twice, resulting in a total number of 240 trials (192 excluding nulls). In the functional localiser, single letters were presented blockwise: one letter was presented for 14 s, followed by a null-trial (9.8 s), followed by a trial of the other letter. Which letter came first was randomised and counterbalanced across

participants.

Statistical testing

For each (paired/one-sample) statistical comparison we first verified that the distribution of the data did not violate normality and was outlier free, determined by the D'Agostino and Pearson's test implemented in SciPy and the 1.5 IQR criterion, respectively. If both criteria were met, we used a parametric test (e.g. paired t-test); otherwise we resorted to a non-parametric alternative (e.g. Wilcoxon sign rank). All statistical tests were two-tailed and used an alpha of 0.05. For effect sizes, we report Cohen's d for the parametric and biserial correlations for the non-parametric tests.

fMRI acquisition

Functional and anatomical images were collected with a 3T Skyra MRI system (Siemens), using a 32-channel headcoil. Functional images were acquired using a whole-brain T2-weighted multiband-4 sequence (TR/TE = 1400/33.03 ms, voxel size = 2 mm isotropic, 75° flip angle, A/P phase encoding direction). Anatomical images were acquired with a T1-weighted MP-RAGE (GRAPPA acceleration factor = 2, $TR/TE = 2300/3.03$ ms, voxel size 1 mm isotropic, 8° flip angle).

fMRI preprocessing

fMRI data pre-processing was performed using FSL 5.0.11 (FMRIB Software Library; Oxford, UK; Smith et al., 2004). The pre-processing pipeline included brain extraction (BET), motion correction (MCFLIRT), temporal high-pass filtering (128 s). For the univariate and univariate-multivariate coupling analyses, data was spatially smoothed with a Gaussian kernel (4mm FWHM). For the multivariate analysis, no spatial smoothing was applied. Functional images were registered to the anatomical image using boundary based registration as implemented in FLIRT and subsequently to the MNI152 T1 2 mm template brain using linear registration with 12 degrees of freedom. For each run, the first 4 volumes were discarded to allow for signal stabilization. Most FSL routines were accessed using the nipype framework (Gorgolewski et al., 2017). Using simple linear registration to align between participants can result in decreased sensitivity compared to more sophisticated methods like cortex-based alignment (Weiner et al., 2018). However, note that using a different inter-subject alignment method would not affect any of the main analyses, which were all performed in native EPI space. The only analysis that could be affected is the whole-brain version of the information-activation coupling analysis (Figures 2.4,S2.11). However, this was only an exploratory follow-up on the pre-defined ROI-based coupling anal-

ysis, intended to identify potential other regions displaying the signature increase in coupling. For this reason the simple linear method was deemed sufficient.

Univariate data analysis

To test for differences in univariate signal amplitude between conditions, voxelwise GLMs were fit to each run's data using FSL FEAT. For the experimental runs, GLMs included four regressors of interest, one for each condition (U-word, U-nonword, etc). For the functional localiser runs, GLMs included two regressors of interest (U, N). Regressors of interest were modelled as binary factors and convolved with a double-gamma HRF. In addition, (nuisance) regressors were added for the first-order temporal derivatives of the regressors of interest, and 24 motion regressors (6 motion parameters plus their Volterra expansion, following Friston et al. (1996). Data were combined across runs using FSL's fixed effects analysis. All reported univariate analyses were performed on an ROI basis by averaging all parameter estimates within a region of interest, and then comparing conditions within participants (see Figures S2.4-S2.5).

Multivariate data analysis

For the multivariate analyses, spatially non-smoothed, motion-corrected, high-pass filtered (128s) data was obtained for each ROI (see below for ROI definitions). Data were temporally filtered using a third-order Savitzky-Golay low-pass filter (window length 21) and z-scored for each run separately. Resulting time courses were shifted by 3 TRs (i.e. 4.2 seconds) to compensate for HRF lag, averaged over trials, and null-trials discarded. For each participant, this resulted in 18 samples per class for the localiser (i.e. training data) and 96 samples per condition (word/nonword) for the main runs (i.e. testing data).

For the classification analysis we used a logistic regression classifier implemented in in sklearn 0.2 (Pedregosa et al., 2011) with all default settings. The model was trained on the time-averaged data from the functional localiser run and tested on the time-averaged data from the experimental runs. Because we had the same number of samples for each class, binary classification performance was evaluated using accuracy (%).

For the pattern correlation analysis, only the time-averaged data from the main experiment was used. Data was randomly grouped into 2 arbitrary splits that both contained an equal number of trials of all 4 conditions (U-word, U-nonword, N-word, N-nonword). Within each split, the time-averaged data of each trial was again averaged to obtain a single average response for each condition per split. For both word/nonword conditions separately, these average responses were then correlated

across splits. This resulted, for both word and nonword conditions, in two (Pearson) correlation coefficients: ρ_{within} and ρ_{between}, obtained by correlating the average response to stimuli with the same or different middle letter, respectively. This process was repeated 12 times, each time using a different random split of the data, and all correlation coefficients were averaged to obtain a single coefficient per comparison, per condition, per participant. Finally, pattern letter information for each condition was quantified by subtracting the two average correlation coefficients ($\rho_{\text{within}} - \rho_{\text{between}}$).

For the searchlight variant of the multivariate analyses, we performed exactly the same procedure as described in the manuscript. However, instead of using a limited number of a priori defined ROIs, we used a spherical searchlight ROI that slid across the brain. A searchlight radius of 6mm was used, yielding an ROI size of about 170 voxels on average, similar to the 200 voxels in our main ROI. For both analyses, this resulted in a map for each outcome metric for each condition for each subject, defined in native EPI space. These maps were then used for subsequent analyses (see *Supplementary Note 1*).

Information-activation coupling analysis

For the information-activation coupling analysis, we used a GLM based approach to predict regional BOLD amplitude as a function of early visual cortex classification evidence, and tested for an increase in coupling (slope) for words compared to nonwords (see Figure 2.4b). The GLM had one variable of interest, visual cortex classification evidence (see below for definition) that was defined on a TR-by-TR basis, and split over two regressors, corresponding to both conditions (word/nonword). In addition, first-order temporal derivatives of the two regressors of interest and the full set of motion regressors (from the FSL FEAT GLM) were included to capture variability in HRF response onset and motion-related nuisance signals, respectively. Because the classification evidence was undefined for null-trials, these were omitted. To compensate for temporal autocorrelation in the data, pre-whitening of the data was applied using the AR(1) noise model as implemented in nistats (Abraham et al., 2014). The resulting GLM yielded two regression coefficients (one per condition) for each participant which were then compared at the group level to test for an increase in coupling in word contexts. Conceptually, this way of testing for condition-dependent changes in functional coupling is analogous to PPI (Friston et al., 1997) but using a multivariate time-course as a 'seed'. This timecourse, classification evidence, was defined as the probability assigned by the logistic regression model to the correct outcome – or $\widehat{p}(\ A \mid y = A)$. This probabilistic definition combines aspects of both prediction accuracy and confidence into a single quantity. Mathematically it is defined via the

logistic sigmoidal function:

$$\hat{p}(A \mid y = A) = \begin{cases} \frac{1}{1+e^{-\theta^T \mathbf{x}}} & \text{if } y = 1 \\ 1 - \frac{1}{1+e^{-\theta^T \mathbf{x}}} & \text{if } y = 0 \end{cases} \tag{2.1}$$

where θ are the model weights, y is the binary stimulus category, $\mathbf{X}$ are the voxel response patterns for all trials, and the letter 'U' is coded as 1 and 'N' as 0. Note that while the value of $\widehat{p}(\ A \mid y = A)$ itself is bounded between 0 and 1, the respective regressors were not after applying prewhitening to the design matrix (see Fig 2.4b).

Two variants of the GLM analysis were performed: one on timecourses extracted from two candidate ROIs and one on each voxel independently. For the ROI-based approach, timecourses were extracted by taking the average timecourse of all amplitude-normalised (z-scored) data from two ROIs: left pMTG and VWFA (see *ROI definition* for details). For the brain-wide variant, the same GLM was estimated voxelwise for each voxel independently. This resulted in a map with the difference in coupling parameters for each voxel, for each participant ($\beta_{\text{word}} - \beta_{\text{nonword}}$) defined in native MRI space. These maps were then transformed to MNI space, after which a right-tailed one-sample t-test was preformed to test for voxels showing an increase in coupling in word conditions. The resulting p-map was converted into a z-map and thresholded using FSL's Gaussian random-field based cluster thresholding, using the default cluster-forming threshold of $z > 3.1$ (i.e., $p < 0.001$) and a cluster significance threshold of $p < 0.05$.

ROI definition

For the ROIs of V1-V4, fusiform cortex and inferior temporal cortex, Freesurfer 6.0 (Dale, Fischl, and Sereno, 1999) was used to extract labels (left and right) per subject based on their anatomical image, which were transformed to native space and combined into a bilateral mask. Labels for V1-V2 were obtained from the default atlas (Desikan et al., 2006) whereas V3 and V4 were obtained from Freesurfer's visuotopic atlas (Van Essen and Dierker, 2007). Early visual cortex (EVC) was defined as the union of V1 and V2.

The visual wordform area (VWFA) was functionally defined following a procedure based on earlier work (Kay and Yeatman, 2017). Briefly, first we took the union of left fusiform cortex and left inferior temporal cortex that were defined via individual cortical parcellations obtained from freesurfer, and trimmed the anterior parts of the resulting mask. Within this broad, left-lateralised ROI we then selected the 200 voxels that were most selective to words over nonwords (i.e. words over orthographically illegal, unpronounceable letter strings) as defined by the highest Z-statistics in the respective word-nonword contrast in the

univariate GLM. Similarly to Kay and Yeatman (2017) we found for most participants this resulted in a single, contiguous mask and in other participants in multiple word-selective patches. There are two main reasons we used the simple contrast word-nonword from the main experiment, rather than running a separate, dedicated VWFA localiser. First, using the main task strongly increased statistical power per subject as we could a full hour of data per participant to localise VWFA. Second, the comparison of words and unpronounceable letter strings (with matched unigram letter frequency) solely targets regions that are selective to lexical and orthographic information (i.e. the more anterior parts of VWFA, according to the VWFA hierarchy reported by (Vinckier et al., 2007). As such, the localiser only targets regions selective to the type of linguistic (lexical or orthographic) knowledge that could underlie the observed effect. This stands in contrast to other, less restrictive VWFA definitions (such as words > phase scrambled words, or words > false fonts).

For the multivariate stimulus representation analyses we did not use the entire anatomical ROIs defined above, but performed a selectivity-selection to ensure we probed voxels that were selective to the relevant part of the visual field. In this procedure, we defined the most selective voxels as those with the k highest Z-statistics when we contrasted any letter (U or N) versus baseline in the functional localiser GLM. Following (Richter et al., 2018) we took 200 voxels as our predefined value for k. To verify that our results were not contingent on this specific (but arbitrary) value, we also made a large range of masks for early visual cortex by varying k between 50 and 1000 with steps of 10. Repeating the classification and pattern correlation analyses over all these masks revealed that the same pattern of effects was obtained over almost the full range of mask definitions, and that the best classification performance was in fact at our predefined value of $k = 200$ (See Figure S2.3).

For the peripheral visual ROI voxels were selected based on the functional criterion that they showed a strong response to stimuli in the main experiment (which spanned a large part of the visual field), but a weak or no response to stimuli in the localiser (which were presented near fixation). Specifically, voxels were selected if they were both in the top 50% of Z-stats for the contrast visual stimulation > baseline in the main experiment, and in the bottom 50% of Z-scores for visuals stimulation > baseline in the localiser. This resulted in masks that contained on average 183 voxels, similar to the 200 voxels in the central ROI. In our initial analysis we focussed on V1 (see Figure S2.9) because it has the strongest retinotopy. However, the same was also applied to early visual cortex with similar results (see Supplementary Note 1).

To define pMTG we performed an automated meta-analysis using Neurosynth (Yarkoni et al., 2011). Because we were interested in pMTG as a hub for lexical access, we searched for the keyword 'semantic'. This resulted in a contrast map based on 1031 studies which we thresholded at an arbitrarily high Z-value of Z>9. The

resulting map was mainly restricted to two hubs, in the IFG and pMTG. We selected left pMTG by overlaying the map with an anatomical mask of medial temporal gyrus from FSL's Harvard-Oxford Atlas. The resulting map was brought to native space by applying the registration matrix for each participant.

Behavioural data analysis

Participants had 1.5 seconds after target onset to respond. Reaction times under 100 ms were considered spurious and discarded. If two non-spurious responses were given, only the first response was considered and evaluated. Median reaction times and mean accuracies were computed for both (word and nonword) conditions and compared within participants.

Eye tracking

Eye movements were recorded using an SMI iView X eye monitor with a sampling rate of 50 Hz. Data was pre-processed and submitted to two analyses: number of trials during which eyes were closed for extended periods, and comparison of horizontal (reading-related) eye movements between conditions.

During pre-processing all data points during which there was no signal (i.e. values were 0) were omitted. After omitting periods with no signal, data points with spurious, extreme values (which sometimes occurred just before or after signal loss) were omitted. To determine which values were spurious or extreme we computed the z-score for each points, over the entire run and ignoring the periods where signal was 0, and considered all values higher than 4 extreme and spurious. Similar to the periods with no signal, these timepoints were also omitted in following analysis. The resulting 'cleaned' timecourses were then visually inspected to evaluate their quality. For two participants the data was of insufficient quality to include in any analysis. For 6 participants, there was enough data of sufficient quality to perform the overall amount of reading-related eye movements between conditions, but signal quality was insufficient to quantify the number of trials during which the eyes were shut for an extended period. This is because in these participants there were various periods of intermittent signal loss that were related to signal quality, not to the eyes being closed. To compare eye movements between conditions, we took the standard deviance of the gaze position over the reading (horizontal) direction, and averaged this over each trial. Because the resulting data contained outliers (i.e. trials during which the participants failed to maintain fixation) we took the median over trials in each condition (word/nonword), and compared them within participants (Figure S2.6). For the participants where the data was consistently of sufficient quality, periods of signal loss longer than 1.2 seconds were considered 'eyes closed for extended

period'. As an inclusion criterion we allowed no more than 25 trials during which eyes were closed for an extended period. This led to the exclusion of 1 participant, who had 33 trials during which the eyes were closed for an extended period. This participant was a clear outlier: of all participants with sufficient quality eye tracking data to be included in this analysis, 14 had no trials during which eyes were closed for an extended period, and in the remaining 12 with at least one such trial the median number of trials was 3.5.

Neural network model

Simulations were performed using a predictive coding formulation of the classic interactive activation model (Rumelhart and McClelland, 1982; Spratling, 2016). We begin by explaining the model at an abstract level, then outline the algorithmic and mathematical details in generic terms, and then specify the exact settings we used for our model architecture, and how we used them in our simulations.

The interactive activation model is a hierarchical neural network model which takes visual features as inputs, integrates these features to recognise letters, and then integrates letters to recognise words. Critically, activity in word-units is propagated back to the letter-level, making the letter detectors sensitive not only to the presence of features (such as the vertical bar in the letter E), but also to neighbouring letters (such as the orthographic context HOUS_ preceding the letter E). This provides a top-down explanation for context effects in letter perception, such as (pseudo)word superiority. The predictive coding formulation of this model was first described by Spratling(2016) . It uses a particular implementation of predictive coding – the PC/BC-DIM algorithm – that reformulates predictive coding (PC) to make it compatible with Biased Competition (BC) and uses Divisive Input Modulation (DIM) as the method for updating error and prediction activations. The goal of the network is to infer the hidden cause of a given pattern of inputs (e.g. the 'hidden' letter underlying a pattern of visual features) and create an internal reconstruction of the input. Note that the reconstruction is model-driven and not a copy of the input. Indeed, when the input is noisy or incomplete the reconstruction will ideally be a denoised or pattern-completed version of the input pattern. Inference can be done hierarchically: at the letter-level, predictions represent latent letters given patterns of features, whilst at the word-level predictions represent latent words given patterns of letters (and reconstructions, inversely, represent reconstructed patterns of letters given the predicted word).

Mathematically the network can be conveniently described as consisting of 3 components: prediction units (**y**), reconstruction units (**r**), and error units (e) that can be captured in only three equations. First, at each level error units combine the

input pattern (x) and the reconstruction of the input (**r**) to compute the prediction error (e) :

$$\mathbf{e} = \mathbf{x} \oslash [\mathrm{r}]_{\varepsilon_2} \tag{2.2}$$

Here, x is a (m by 1) input vector; r is a (m by 1) vector of reconstructed input activations, $\oslash$ denotes pointwise division and the square brackets denote a max operator: $[v]_\epsilon = \max(\epsilon, v)$. This max-operator prevents division-by-zero errors when all prediction units are silent and there is no reconstruction. Following Spratling (2016) we set ϵ_2 at 1×10^{-3}. Division sets the algorithm apart from other versions of predictive coding that use subtraction to calculate the error (see Spratling (2016) for review). The prediction is computed from the error via pointwise and matrix multiplication:

$$\mathbf{y} \leftarrow [\mathbf{y}]_{\epsilon_1} \otimes \mathbf{We} \tag{2.3}$$

Here, W is a (n by m) matrix of feedforward weights that map inputs onto latent causes (e.g. letters), $\otimes$ denotes pointwise multiplication, square brackets represents a max operator and ϵ_1 is set at 1×10^{-6}. Each row of W maps the pattern of inputs to a specific prediction unit representing a specific latent cause (such as the letter) and can hence be thought of as the 'preferred stimulus' or basis vector for that prediction unit. The entire W matrix is then best thought of as comprising the layer's model of its environment. Finally, from the distribution of activities of the prediction units (y), the reconstruction of expected input features (r) is calculated as a simple linear generative model:

$$\mathbf{r} = \mathbf{Vy} \tag{2.4}$$

Where V is a (m by n) matrix of feedback weights that map predicted latent causes (e.g. letters) back to their elementary features (e.g. strokes) to create an internal reconstruction of the predicted input, given the current state estimate. The model adheres to a form of weight symmetry: V is almost identical to $\mathrm{W}^{\mathbf{T}}$, but its values are values normalised so that each column sums to one. To perform inference, prediction units can be initialised at zero (or with random values) and the Equations (2,3,4) are updated iteratively. To perform top-down hierarchical inference, reconstructions from a higher-order stage (e.g. recognizing words) can be sent back to the lower-order stage (e.g. recognising letters) as additional input. To accommodate these recurrent inputs, additional weights have to be defined that are added to W and V as extra columns and rows respectively. The strength of these weights is scaled to control the reliance on top-down predictions.

Architecure specification

The interactive activation architecture we used was a modification of the network described by (Spratling, 2016) extended to recognise 5-letter words, trained on the Dutch subtlex vocabulary, and with a slight change in letter composition. Letters are presented to the network using a simulated font adapted from the one described by Rumelhart and Siple (1974) that composes any character using 14 strokes (Figure S2.12). For our 5-letter network, the input layer comprises five 14-dimensional vectors (one per character) that each represent the presence of 14 line segments for one letter position. Note that conceptually it is easier to partition the input into five 14-dimensional vectors, in reality these were concatenated into a single 70-dimensional vector x

At the first level, weight matrix W has 180 rows 250 columns: rows comprise 5 slots of 36 alphanumeric units (5 × 36 = 180); the first columns comprise 5 slots of 14 input features (5 × 14 = 70) and the last 180 columns route the top-down reconstruction from the word level. To define the weights of 70 (feedforward) columns, we used encoding function $\varphi(c)$ that takes an alphanumeric character and maps it into a binary visual feature vector. For each alphanumeric character, the resulting feature vector was concatenated 5 times and the resulting 70 dimensional vector comprised the first row. This was repeated for all 36 alphanumeric characters and concatenated 5 times. The resulting numbers were then normalised so that the columns summed to one. Then we added the weights of the second 180 columns (inter-regional feedback coming from 5x36 letter reconstructions) were simply a 180 by 180 identity matrix multiplied by a scaling factor to control top-down strength. For our 'top-down model' (Fig 2.3b) we set the scaling factor at 0.4; in the 'bottom-up model' we set it to 10^{-6} to effectively cancel the influence of feedback, resulting in a 'bottom-up' model. At the second level, weight matrix W had 6778 rows and 180 columns, representing 6776 Dutch 5 letter words from the subtlex corpus, plus the 2 learned nonword targets (that we included in the vocabulary as participants learned these during training) and 5 times 36 alphanumeric characters. The orthographic frequency of letters as specified by the corpus was hard coded into the weights and then normalised to sum to one.

Although there are substantial implementational differences between this model and the classic connectionist version of the interactive activation model (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) the version described here has been shown to capture all key experimental phenomena of the original model (see Spratling, 2016 for details). Since our simulations only tried to validate and demonstrate a qualitative principle, not subtle quantitative effects, the exact numerical differences related to the differences in implementation should not matter for the

effect we demonstrate here.

Simulations

Because our paradigm is different from classical paradigms, we performed simulations to confirm that the top-down account indeed predicts the representational enhancement we set out to detect. Although the main simulation result (2.3a) is not novel, our simulation, by mirroring our paradigm, departs from earlier simulations in some aspects, which we will clarify before going into the implementation details. First, most word superiority studies present stimuli near-threshold: words are presented briefly, followed by a mask, and average identification accuracies typically lie between 60% and 80%. This is mirrored in most classic simulations, where stimuli are presented to the network for a limited number of iterations and followed by a mask, leading to similar predicted response accuracies (McClelland and Rumelhart, 1981; Spratling, 2016). In our task, stimuli are presented for almost a second, and at least the critical middle letter is always clearly visible. This is mirrored in our simulations, where stimuli are presented to the network until convergence and predicted response accuracies of the network are virtually 100% in all conditions (see Fig S2.2). As such, an important aspect to verify was that enhancement of a critical letter can still occur when it is well-above threshold and response accuracy would be virtually at 100% already. Second, our simulations used the same Dutch word and nonword materials used in the experiment. This includes the occurrence of *learned* targets in the nonword condition which we added to the vocabulary of the network and were hence a source of contamination as 12% of the items in the nonword condition were in fact in the vocabulary. Finally, unlike classical simulations, stimuli were corrupted by visual noise.

For 2.3a, we simulated 34 artificial 'runs'. In each run, 48 words and 48 nonwords were presented to a network with feedback connections (feedback weight strength 0.4) and without word-to-letter feedback (feedback weight strength 10^{-6}). The same Dutch, 5 letter (non)words were used as in the main experiment, and like in the experiment 12% of the (non)word items were replaced by target items. Critically, the nonword targets were learned and hence were part of the vocabulary of the network. To present a (non)word to the network each character c has to be first encoded into a set of visual features and then corrupted by visual noise to produce an input vector $\mathbf{x}$

$$\mathrm{x} = \varphi(c) + \mathcal{N}\left(\mu, \sigma^2\right) \tag{2.5}$$

For μ we used 0, σ was set to 0.125, and any values of x that became negative after adding white noise were zeroed. The network then tried to recognise the word

by iteratively updating its activations using Equation (2), (3) and (4), for 60 iterations. To compute the 'relative evidence' metric we used in Fig 2.3a to quantify representational quality $q(\mathbf{y})$ we simply take the fraction of activation for the correct letter $(\mathbf{y}_i)$ of the sum of letter activations for all characters at the third slot:

$$q(\mathbf{y}) = \frac{\mathbf{y}_i}{\sum_{j=37}^{73} \mathbf{y}_j} \tag{2.6}$$

Finally, to compute predicted response probabilities as in Figure S2.2, we followed McClelland and Rumelhart to use Luce's decision rule to compute responses probabilistically:

$$p(R_j) = \frac{e^{\beta y_i}}{\sum_{j=37}^{73} e^{\beta y_j}} \tag{2.7}$$

The β parameter (or inverse softmax temperature) determines how rapidly the response probability grows as $\mathbf{y}_i$ increases (i.e. the 'hardness' of the argmax operation) and was set at 10, following Rumelhart and McClelland (1982); but results are similar for any typical beta value that is approximately in the same order of magnitude.

All simulations were performed using custom MATLAB code, which was an adaptation and extension of the implementation by (Spratling, 2016).

Data availlability

All raw data required to reproduce all analyses and figures are uploaded onto the Donders Data Repository and can be found at http://hdl.handle.net/11633/aacjymw7.

Acknowledgements

This work was supported by The Netherlands Organisation for Scientific Research (NWO Research Talent grant to M.H; NWO Vidi grant to F.P.d.L.; 016.Veni.195.435 to M.E.; Gravitation Program Grant Language in Interaction no. 024.001.006 to P.H.) and the European Union Horizon 2020 Program (ERC Starting Grant 678286, "Contextvision" to F.P.d.L). We thank Ashley Lewis for helpful comments on and discussions of an earlier version of this manuscript.

Author contributions

M.H., F.P.d.L., P.H., D.R. and M.E. designed the study. M.H. and D.R. collected the data. M.H., D.R., M.E. and F.P.d.L. conceived of the analysis plan. M.H. analysed the data. M.H. performed simulations. M.H. wrote the initial draft. All authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

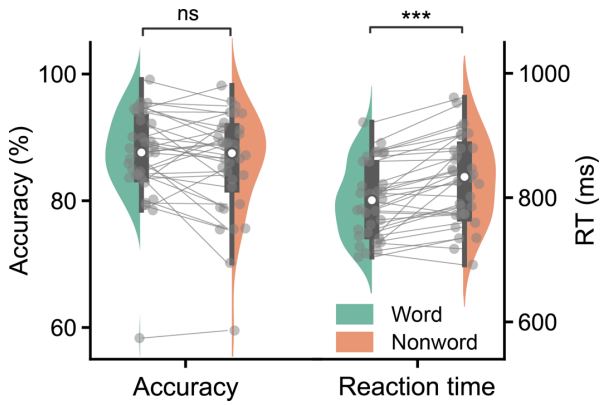

Figure S2.1. Behavioural results. To make sure participants kept reading and were equally attentive of words and nonwords, they performed a challenging orthographic discrimination task. The task was performed on specific, learned targets that were presented about once per trial at an unpredictable moment. Targets were learned during a separate training session and were presented either in their regular (learned) form or with one of the non-middle letters permuted. Whenever a target was presented participants had to report whether it was correctly 'spelled'. Participants were faster (Wilcoxon signed rank, T=40, $p = 1.07 \times 10^{-5}, r = 0.87$) but not statistically significantly more accurate (two-tailed t-test, $t_{34} = 1.70, p = 0.098$, $d = 0.29$) for word compared to nonword targets. This is in line with word superiority, although the perceptual nature of this advantage cannot be established from behavioural results on this task alone as there might also be memory or decisional factors contributing to the observed facilitation. Grey dots with connecting lines are individual participants. Colours are estimated densities, white dots are group medians, boxes are quartiles and whiskers are 1.5 interquartile range. Significance stars indicate $p < 0.001$ (***) in a (paired) two-tailed Wilcoxon sign rank test.

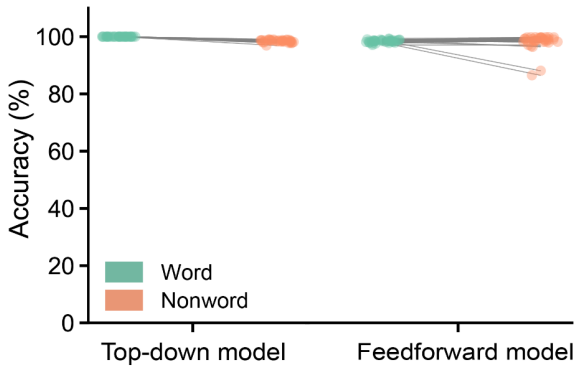

Figure S2.2. Simulated letter identification accuracies. All simulation parameters were identical to the simulation of Figure 3a, except that median predicted response accuracy, rather than representational strength, for the middle letter was computed (see *Methods*). The fact that the accuracies are virtually at 100% in all conditions shows that stimuli were, despite the visual noise, clearly 'visible' to the network (note that chance level would be 3.84% or 1/26). This reflects a key difference between our paradigm – in which stimuli were presented well-above threshold – and the majority of studies in the literature – where stimuli are presented near-threshold. These results confirm that even when the critical letter is clearly visible and predicted letter identification responses are virtually at 100%, theoretical models still predict that enhancement of representations can occur. The accuracy values here might appear in conflict with the accuracies in Figure S2.10. Note however that in the behavioural task, performance did not purely rely on perception of letters but also on their comparison to a memory template, and that the task was performed on the outer letters while participants maintained fixation at the centre of the screen. The middle letter was therefore always well-identifiable, making the predicted near-perfect accuracies a reasonable approximation of experimental viewing conditions.

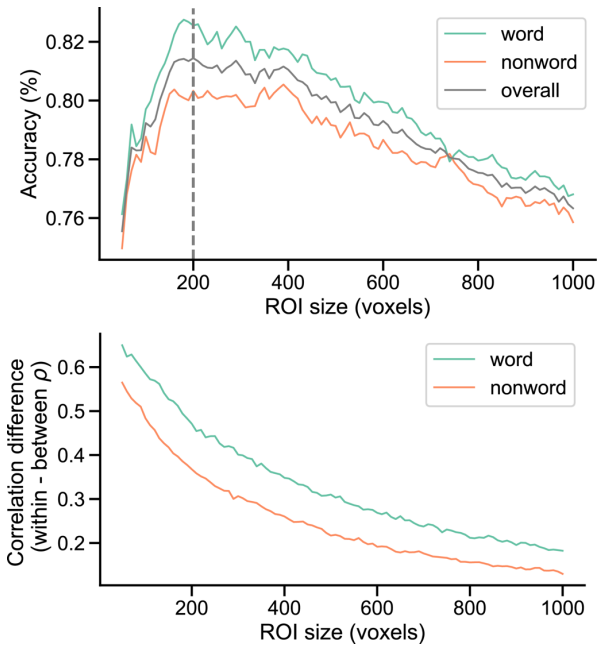

Figure S2.3. Key contrast in main region of interest is stable over a range of ROI sizes. Same analysis as in Figure 3b, but performed over a wide range of ROI sizes, from 50 to 1000 voxels, with steps of 10. For both classification accuracy (upper panel) and pattern correlation difference (lower panel), the same pattern of effects was found practically over the full range of ROIs. Strikingly, the highest overall classification accuracy (vertical dashed line, corresponding to the maximum value of the solid grey line) was found at the pre-defined ROI of 200 voxels – a number that we based on a previous study (Richter and Ekman, 2018). Although the difference with other, similar ROI sizes is negligible, this result confirms that the choice for 200 voxels was justified in the sense that choosing a different number could not have considerably improved the decoding performance.

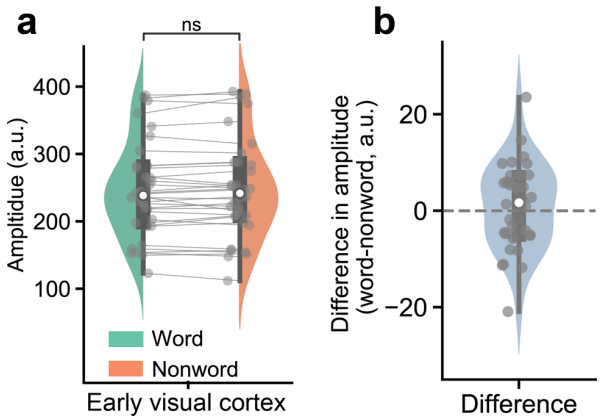

Figure S2.4. No difference in amplitude between conditions. (A) Same analysis as in Figure 3b, but performed over a wide range of ROI sizes, from 50 to 1000 voxels, with steps of 10. For both classification accuracy (upper panel) and pattern correlation difference (lower panel), the same pattern of effects was found practically over the full range of ROIs. Strikingly, the highest overall classification accuracy (vertical dashed line, corresponding to the maximum value of the solid grey line) was found at the pre-defined ROI of 200 voxels – a number that we based on a previous study (Richter and Ekman, 2018). Although the difference with other, similar ROI sizes is negligible, this result confirms that the choice for 200 voxels was justified in the sense that choosing a different number could not have considerably improved the decoding performance.

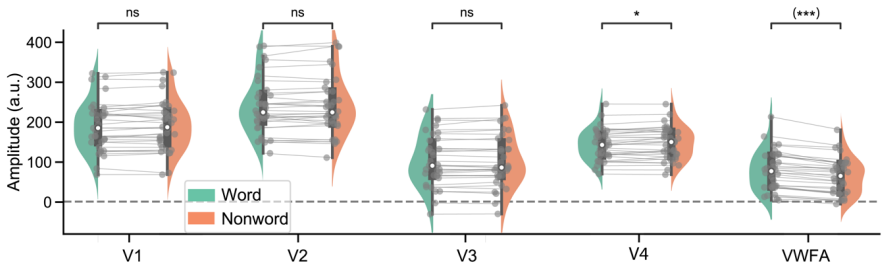

Figure S2.5. Univariate results for various ROIs. Same as Supplementary Figure 4 but for 4 anatomically defined visual regions (V1-V4) and one functionally defined region (VWFA). Overall, there were no strong amplitude differences between conditions in most regions of interest, except for VWFA where BOLD amplitude was by definition higher for words than nonwords in each subject. Significance levels: * indicates $p < 0.05$ (uncorrected), and $(***)$ indicates difference-by-definition (no stats).

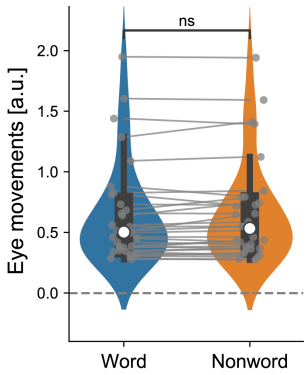

Figure S2.6. Comparison of reading-related eye movements across conditions. Horizontal eye movements were quantified for each trial and then averaged for both conditions and compared within participants. Grey dots and connecting lines represent single participants, white dots group medians, boxes and whiskers represent quartiles and 1.5 interquartile ranges. No statistically significant difference between conditions was found (paired t-test, $t_{32} = -1.43$, $P = 0.16$). Two participants were not included because there was no eye tracking data of sufficient quality.

Supplementary Note 1

If the letter information extracted from visual cortex, and its enhancement by word contexts, indeed reflect sensory representations, then the MVPA results should be retinotopically specific. If, on the other hand, letter identity could be decoded from voxels throughout much of the brain, or if the enhancement was not retinotopically specific (e.g. reflecting a more general increase in signal-to-noise ratio) it would be more difficult to conclude that the MVPA results reflect sensory representations. We therefore tested for spatial specificity by running a searchlight version of the classification and pattern correlation analyses. Figures S2.7 and S2.8 depict the group averaged results of both analyses. In both figures, the *colour* of the overlay represents the difference in letter decoding between conditions (word minus nonword), while the *opacity* represents the extent to which the overall letter decoding is above chance (irrespective of condition).

This way, the difference between conditions is only visible when the overall decoding was above chance. From Figures S2.7 and S2.8, two things become clear. First, opacity is nonzero almost exclusively in visual regions, implying that only there decoding was above chance, and that the letter decoding was could not have relied on a global pattern, but only on information from visual cortex. Second, most of the overlay is red. This means that in the regions with above-chance decoding, the difference between conditions is almost always positive. This converges with Figure S2.3, by confirming that this pattern of effects was not contingent on the specific (but arbitrary) ROI definition we employ.

Figures S2.7 and S2.8 clearly show that letter decoding is specific to visual cortex. However, from the maps it is difficult to see if, *within* visual cortex, the letter decoding and representational enhancement peak the expected (foveal) location. This is because the individual maps got smeared out during averaging in standard space. Therefore, we ran a more sensitive ROI analysis in native EPI space. Here, we use the resulting searchlight maps (containing classification and pattern correlation results for each voxel in a participant's native EPI space). We compared the classification in the central ROI (using the functional definition described earlier) to a functionally defined peripheral ROI. Voxels were deemed peripheral when they showed a strong response to stimuli in the main experiment (which spanned a large part of the visual field), but showed a weak or no response to stimuli in the localiser (which were presented near fixation). For this analysis we focused on V1, because it has the strongest retinotopy. Indeed, as can be seen in Fig S2.9 overall letter decoding was greatly reduced for the peripheral ROI compared to the central ROI, both for the classification analysis (paired t-test, t_{34}=15.59, p = 8.86×10^{-17}, d = 2.67) and pattern correlation analysis (paired t-test, $t_{34} = 8.06$, p = 2.65×10^{-9}, d = 1.38).

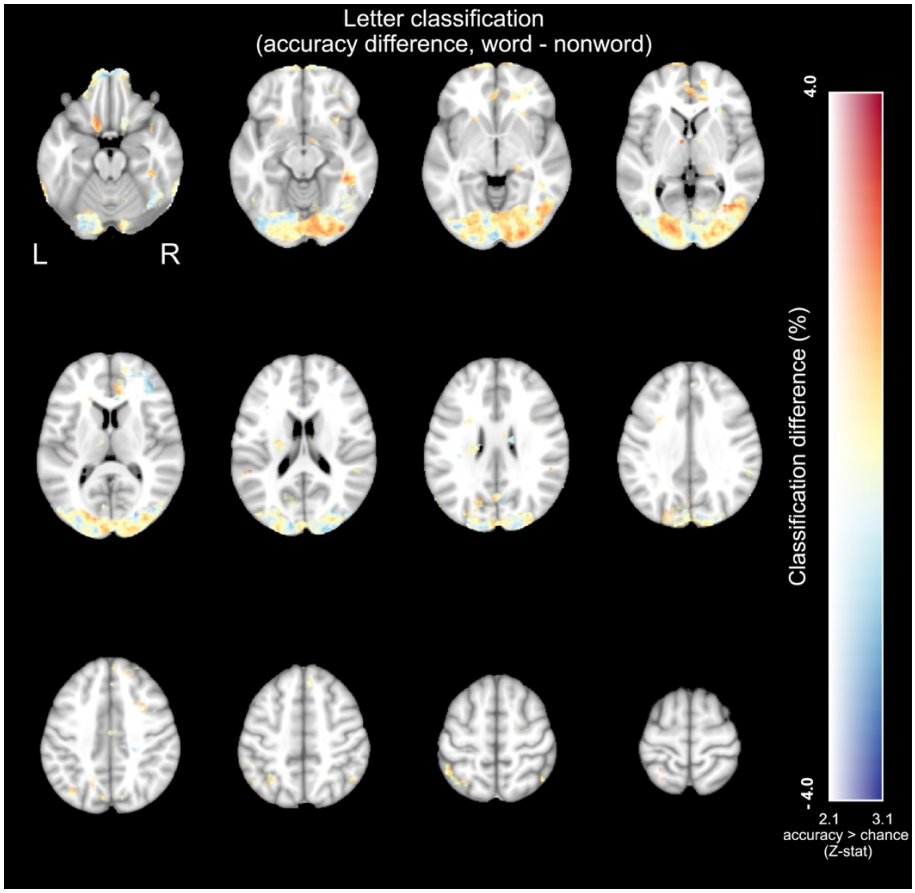

Figure S2.7. Spatial specificity of classification analysis. Group averaged result of the searchlight version of the classification analysis. This figure uses a dual-coding scheme in which the opacity of the overlay is determined by the average decoding accuracy with respect to chance (averaged over subjects), and the colour indicates the average decoding difference (word-nonword) between conditions. See text (*Supplementary Note 1*) for interpretation.

Critically, a similar reduction in the peripheral ROI was found for the enhancement effect (the difference in decoding between conditions), again both for the classification analysis (paired t-test, t_{34}=2.56, p = 0.015, d = 0.44) and pattern correlation analysis (paired t-test, t_{34}=2.92, p = 6.31×10^{-3}, d = 0.50). Importantly, although we initially (Figure S2.9) focussed on V1 – because it has the strongest retinotopy and because it was requested by the reviewer – a similar reduction was observed for our main ROI of interest, early visual cortex (i.e. the conjunction of V1 and V2). Specifically, here too we found greatly reduced overall letter decoding, both for the classification analysis (paired t-test, t_{34}=18.49, $p = 5.52 \times 10^{-19}$, d = 3.17) and pattern

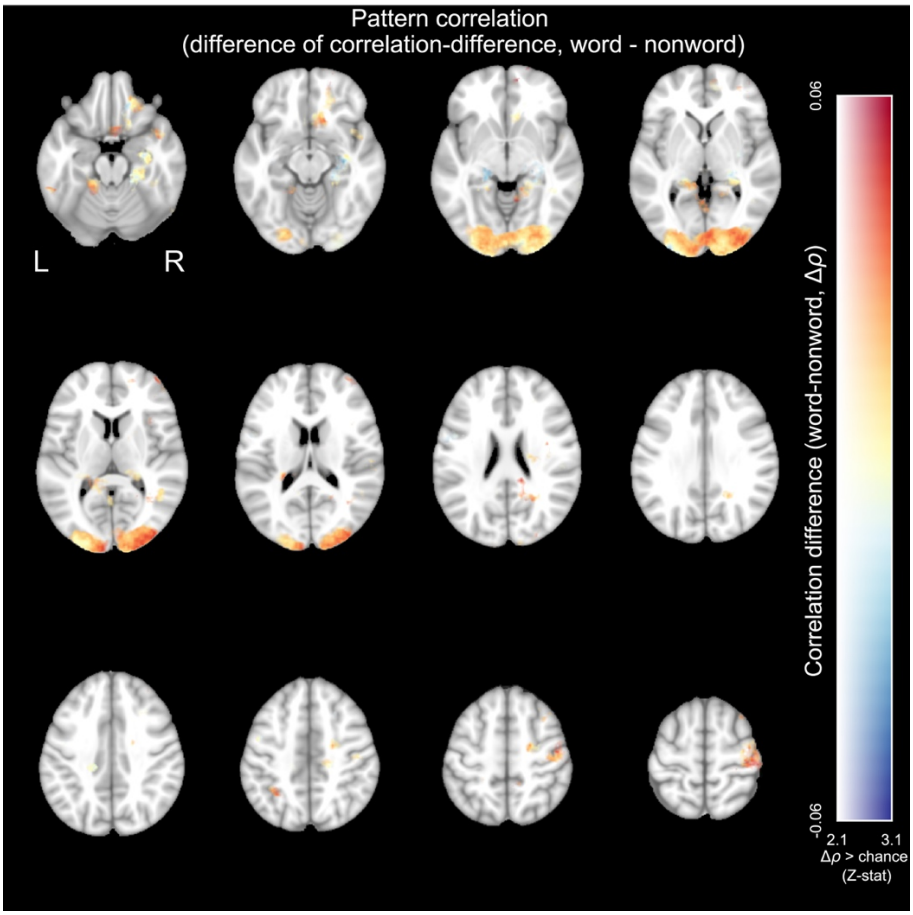

Figure S2.8. Spatial specificity of pattern correlation analysis. Group averaged result of the searchlight version of the pattern correlation analysis. Results are displayed using a dual-coding scheme in which the opacity of the overlay is determined by the average letter decoding performance (quantified as pattern correlation difference) with respect to chance, and the colour indicates the decoding difference between conditions (word-nonword). See text (*Supplementary Note 1*) for interpretation.

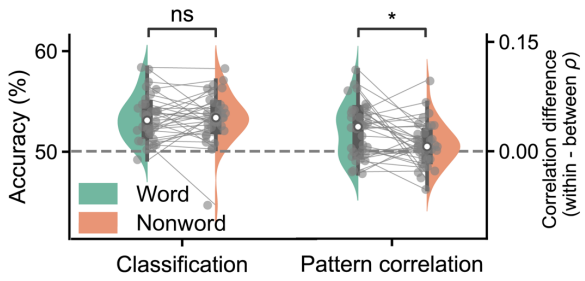

Figure S2.9. Reduced letter decoding and representational enhancement in the periphery. Same analysis as in 2.3b, but now for the peripheral V1 ROI (individually defined for each participant). Compared to the central V1 ROI, both classification and pattern correlation analyses revealed a reduction, both for overall letter decoding (both p's$< 10^{-8}$, paired t-test), and representational enhancement (both p's< 0.016, paired t-test). This reduction suggests both analyses relied on retinotopically specific, early sensory information. The same effect is found when this analysis is performed on early visual cortex (see text). Grey dots with connecting lines are individual participants. Colours are estimated densities, white dots are group medians, boxes are quartiles and whiskers are 1.5 interquartile range. Significance stars indicate $p < 0.05$ (*) in a (paired)two-tailed t-test

correlation analysis (paired t-test, t_{34}=8.86, $p = 3.02\times10^{-10}$, d = 1.52). Moreover, we again found a reduction of the enhancement effect, again both for the classification analysis (paired t-test, t_{34}=2.44, $p = 0.02, d = 0.42$) and pattern correlation analysis (paired t-test, t_{34}=3.21, $p = 2.90\times 10^{-3}$, d = 0.55). Together, these analyses show that MVPA results exhibit spatial and retinotopic sensitivity, which suggests that the MVPA results indeed reflect early visual representations, as expressed in BOLD activity.

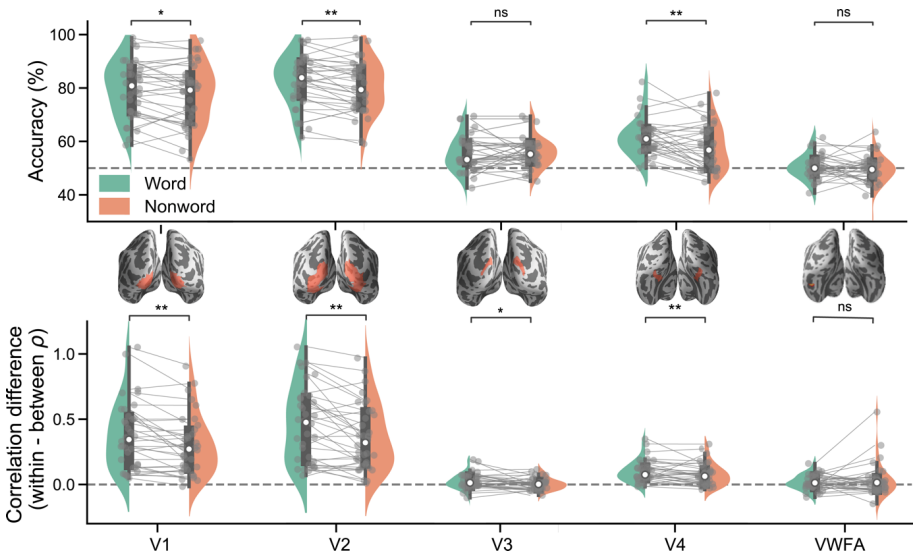

Figure S2.10. Enhancement throughout the visual hierarchy. Same analysis as in Fig 2.3b, over the same ROIs as in Fig S2.5. Overall, in all three ROIs where overall letter decoding was well-above chance, the key enhancement effect was found; in no ROI was the pattern inverted. Specifically, both classification and pattern correlation analyses revealed evidence for word enhancement in V1 (classification analysis t_{34} = 2.35, p = 0.025, d = 0.40; correlation difference: Wilcoxon signed rank t_{34} = 115, p = 1.81 × 10^{-3}, r = 0.61) V2 (classification difference: t_{34}=3.043; P = 4.57 × 10^{-3}, d = 0.52; correlation difference: Wilcoxon's t_{34} = 99.0, p = 6.90 × 10^{-4}, r = 0.68) and V4 (classification difference: t_{34} = 3.42, p = 1.67 × 10^{-3}, d = 0.59; correlation difference: Wilcoxon's t_{34} = 151.0, p = 0.012, r = 0.49). However, no consistent differences were found for V3 (classification difference, Wilcoxon's t_{34} = 176, p = 0.54, r = 0.13; correlation difference: Wilcoxon's t_{34} = 172, p = 0.032, r = 0.42; see figure and note difference in direction); and VWFA (classification difference: t_{34} = 1.18, p = 0.25, d = 0.20; correlation difference: Wilcoxon's t_{34} = 151.0, p = 0.012, r = 0.49). Brain images are surface plots with anatomical ROI overlays created using the pysurfer plotting engine (Ramachandran and Varoquaux, 2011).

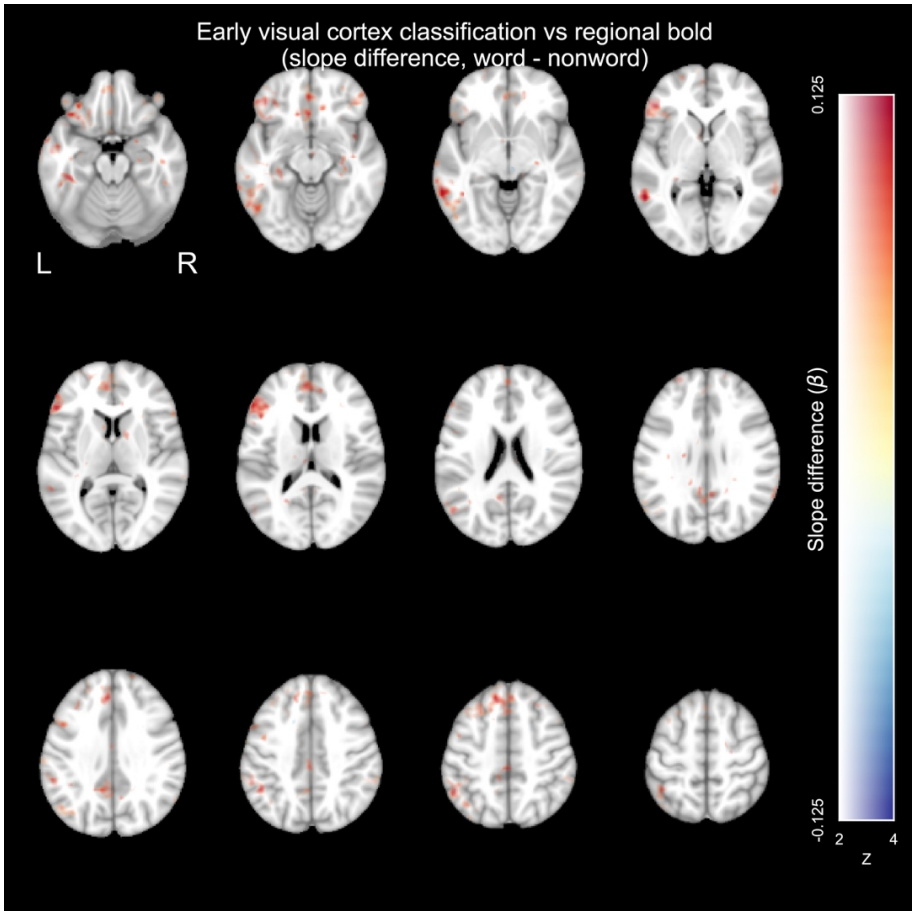

Figure S2.11. Non-thresholded whole brain result of the information-activation coupling analysis. Same results as in 2.4c, but using a dual coding scheme in which the overlay is opacity-weighted by statistical values instead of a binarily thresholded at statistical significance. Colour indicates the numerical difference in the information activation coupling parameter between conditions (word-nonword), opacity represents the consistency of this difference over participants, expressed using the Z-statistic. From the results it becomes evident that even without thresholding, the lateralisation, and two statistically significant clusters in left MTG and IFG, clearly stand out.

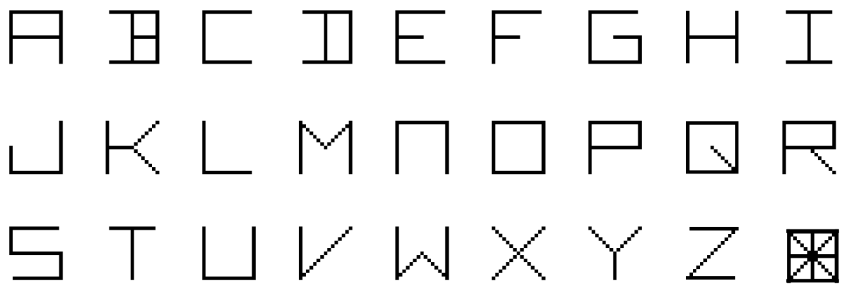

Figure S2.12. Illustration of virtual font. Illustration of the virtual font presented to the network. In this font all 36 alphanumeric characters can be formed from only 14 line segments. This allows each character to be encoded as a 14-dimensional input vector representing visual features. Font is adapted from Rumelhart and Siple (1974), slightly modified to increase similarity between U and N, and overlap with other letters, to simulate what we used in our experiment.

Uword		Unonword		Nword		Nnonword	
ABUIS	KRUIK	REUJZ	AEUEI	AGNES	LYNCH	NMNNS	DSNEN
ACUTE	KRUIP	KNUUE	IOUST	BANDS	MANDY	DTNAI	INNTE
ACUUT	KRUIS	DGUNE	RNUAH	BANEN	MANEN	ILNTN	IENNW
AZUUR	KRUIT	ITUOD	DGUWD	BANGE	MENEN	HSNND	MTNSA
BEURS	LAUDE	TGUAE	OEUAT	BANJO	MENGT	NKNSE	NTNBW
BEURT	LEUKE	LNUOT	ASUEA	BANKS	MENIG	AINKH	ETNKD
BLUES	LEUKS	EDUTB	ENUDP	BENDE	MINST	JDNIV	VJNTS
BLUFT	LEUNT	NIUDL	TKUEP	BENEN	MINUS	ARNWT	ITNEI
BOUWT	MEUTE	ONUHB	OAUPI	BENUL	MONTE	IHNTR	MDNJT
BRUID	MOUTH	NPUAO	JNUCE	BINDT	NANNY	LONRH	ERNLM
BRUIN	NAUWE	FDUDE	DHUJD	BINGO	NINJA	NKNWV	DTNCA
BRULT	NEURO	EIUSP	EDUSJ	BONEN	OPNAM	GNNRT	MUNJE
BRUTE	PAUZE	LNUME	MLUHN	BONES	PANTY	RTNBE	AONRL
BRUTO	PLUIM	AGUEK	OWUAO	BONUS	PINDA	ENNTL	KRNBG
BRUUT	PLUIS	RZUNI	MNUDV	CONGO	PUNCH	DINRD	NMNCN
BUURT	PLUKT	EAUYI	ONUIE	DANDY	RANCH	NRNMI	ZDNNH
COUPE	PRUIK	WVUGN	TDUER	DANKT	RENDE	WVNVS	NDNEA
DEUGD	PRUIM	HLUOR	NTURN	DANST	RENTE	RVNNE	MVNAM
DEUGT	RAUWE	OAUWV	ENUAW	DENKT	RONDE	RDNRA	RTNXV
DRUGS	REUMA	ITUNB	DZUEO	DINER	RUNDE	EWNDZ	DJNET
DRUIF	REUZE	AIUVS	NLURE	DONOR	SANDS	IHNOI	LHNNE
DRUKT	ROUGE	RHUEJ	JDUNE	DONUT	SENOR	TPNLK	AJNCN
DRUMS	ROUTE	EHUDB	EBUUI	DUNNE	SINDS	ZTNZE	TNNSE
DUURT	ROUWT	IEUOI	NMURF	EINDE	SONAR	ZGNRE	NLNUI
EEUWS	SAUNA	NHUEZ	WVUNI	FONDS	SONDE	KDNNA	DRNLZ
ERUIT	SLUIP	DEUEO	NUUDA	GENAS	SONGS	ENNRH	DLNEN
FAUNA	SLUIS	AEUVR	SUUET	GENEN	TANGO	DRNEG	NCNEH
FLUIT	SLUIT	ZKUEN	EOUUN	GENIE	TANKS	VNNAE	CNNWI
FOUTE	SLURF	FWUTE	TAULR	GENOT	TANTE	IENWR	ARNNK
FOUTS	SLUWE	SBUAI	EIUAW	GENRE	TENEN	EINAT	RDNMH
FRUIT	SNUIF	AEUVO	NMUEN	GINDS	TONEN	JVNNR	JDNNS
GEUIT	SNUIT	HRUEN	TKUES	GUNST	TONIC	OCNEO	EDNRG
GOUWE	SNURK	GHUOW	VAUAO	HANGT	VANAF	FZNND	NMNTP
HEUSE	SPUIT	TLULZ	UNUEA	HINTS	VANGT	TBNRK	EONNI
HOUDT	SPUUG	EAUAG	VTUNL	INNEN	VENUS	PTNVO	WTNHE
HOUSE	SPUWT	EVUNE	EIUOA	JONGE	VINDT	KSNGI	DTNWT
HUURT	SQUAD	RHUET	ZMUVT	KANON	WENEN	THNLR	DLNTE
JEUGD	STUFF	NPUAL	THUIJ	KENDE	WENST	UONTD	DRNAE
JEUKT	STUIT	JHUTZ	ETUDL	KINDS	WINDT	TNNRI	RDNZJ
JOUWE	STUKS	TXUEM	VHUOR	KINKY	WINST	DRNNM	SINNO
KAUWT	STUNT	HRUMN	EAUAR	KUNST	WONDE	MNNGO	OTNUE
KEURT	STUUR	NIUEH	AUULS	LANDT	WONEN	GRNEM	WLNEH
KEUZE	THUIS	NTUEL	OGUBN	LANGE	ZENDT	DTNJI	VWNOE
KLUIF	TRUCK	TGURV	ANUET	LANGS	ZENUW	EANAC	TLNEG
KLUIS	TRUCS	HDUPM	NLUAR	LENEN	ZINGT	IENNG	EONNA
KLUNS	TRUST	MNUHC	ODUAL	LENTE	ZINKT	ETNIR	NDNTN
KLUTS	TRUUK	DLUEI	EHUWJ	LINIE	ZONDE	TZNRO	TDNLT
KOUDE	VUURT	VNUDW	PEUEA	LINKS	ZONEN	EMNSC	IHNSE
KOUDS	ZEURT	ZUUAH	TNUEV	LONEN	ZONES	IGNEM	ODNRB
KRUID	ZOUTE	HGUTO	ENUIZ	LUNCH	ZONET	VNNAR	KMNHT

Table S2.1. Word and nonword stimuli used in main experiment. Words were taken from a corpus scraped from a large number of subtitles and hence also contains names and highly common English terms that are not Dutch in a strict sense. But critically, all word items are highly familiar and pronounceable, whereas all nonword items are unfamiliar and unpronounceable.

Chapter 3

Tracking naturalistic linguistic predictions with deep neural language models

Abstract

Prediction in language has traditionally been studied using simple designs in which neural responses to expected and unexpected words are compared in a categorical fashion. However, these designs have been contested as being 'prediction encouraging', potentially exaggerating the importance of prediction in language understanding. Recent studies have begun to address these worries by using model-based approaches to probe the neural effects of linguistic predictability in naturalistic stimuli (e.g. continuous narrative). However, these studies so far only looked at very local forms of prediction, using models that take no more than the prior few words into account when computing a word's predictability. Here, we extend this approach using a state-of-the-art neural language model that can take roughly 500 times longer linguistic contexts into account. Predictability estimates from the neural network offer a much better fit to EEG data from subjects listening to naturalistic narrative than simpler models, and reveal strong surprise responses akin to the P200 and N400. These results show that predictability effects in language are not a side-effect of simple designs, and demonstrate the practical use of recent advances in AI for the cognitive neuroscience of language.

This chapter is based on:
Heilbron, M., Ehinger, B., Hagoort, P., de Lange, FP. (2019). Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. Conference on Cognitive Computational Neuroscience, 424-427.

Introduction

In a typical conversation, listeners perceive (or produce) about 3 words per second. It is often assumed that prediction offers a powerful way to achieve such rapid processing of often-ambiguous linguistic stimuli. Indeed, the widespread use of language models – models computing the probability of upcoming words given the previous words – in speech recognition systems demonstrates the in-principle effectiveness of prediction in language processing (Jurafsky and Martin, 2014).

Linguistic predictability has been shown to modulate fixation durations and neural response strengths, suggesting that the brain may also use a predictive strategy. This dovetails with more general ideas about predictive processing (de Lange, Heilbron, and Kok, 2018; Friston, 2005; Heilbron and Chait, 2018) and has lead to predictive interpretations of classical phenomena like the N400 (Kuperberg and Jaeger, 2016; Rabovsky, Hansen, and McClelland, 2018). However, most neural studies on prediction in language used hand-crafted stimulus sets containing many highly expected and unexpected sentence endings – often with tightly controlled (predictable) stimulus timing to allow for ERP averaging. These designs have been criticised as 'prediction encouraging' (Huettig and Mani, 2016), potentially distorting the importance of prediction in language.

A few recent studies used techniques from computational linguistics combined with regression based deconvolution to estimate predictability effects on neural responses to naturalistic, continuous speech. However, these pioneering studies probed very local forms of prediction by quantifying word predictability based on only the first few phonemes (Brodbeck, Hong, and Simon, 2018) or the prior two words (Armeni et al., 2019; Willems et al., 2016). Recently, the field of artificial intelligence has seen major improvements in neural language models that predict the probability of an upcoming word based on a variable-length and (potentially) arbitrarily-long prior context. In particular, self-attentional architectures (Vaswani et al., 2017) like GPT-2 can keep track of contexts of up to a thousand words long, significantly improving the state of the art in long-distance dependency language modelling tasks like LAMBADA and enabling the model to generate coherent texts of hundreds of words (Radford et al., 2019). Critically, these pre-trained models can achieve state-of-the art results on a wide variety of tasks and corpora without any fine-tuning. This stands in sharp contrast to earlier (ngram or recurrent) language models which were trained on specific tasks or linguistic registers (e.g. fiction vs news). As such, deep self-attentional language models do not just coherently keep track of long-distance dependencies, but also exhibit an unparalleled degree of *flexibility*, making them arguably the closest approximation of a 'universal model of English' so far.

Here we use a state-of-the art pre-trained neural language model (GPT-2 M) to

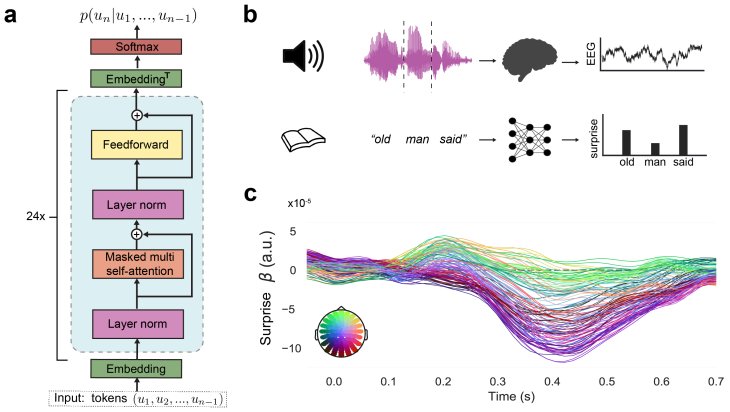

Figure 3.1. a) GPT-2 architecture. For more info on individual operations, see Vaswani et al. (2017). (Note that this panel is a re-rendered version of the original GPT schematic, with subcomponents re-arranged to match the architecture of GPT-2.) **b)**Analysis pipeline overview. **c)** Obtained series of β coefficients (TRF) of lexical surprise (from GPT-2), averaged over participants.

generate word-by-word predictability estimates of a famous work of fiction, and then regress those predictability estimates against publicly-available EEG data of participants listening to a recording of that same work.

Methods

Stimuli, data acquisition and preprocessing

We used publicly available EEG data of 19 native English speakers listening to Hemingway's *The Old Man and the Sea*. Participants listened to 20 runs of 180s long, amounting to the first hour of the book (11,289 words, ~3 words/s). Participants were instructed to maintain fixation and minimise all motor activities but were otherwise not engaged in any task.

The dataset contains raw 128-channel EEG data downsampled to 128 Hz, plus on/offset times of every content word. The raw data was visually inspected to identify bad channels, decomposed using ICA to remove blinks, after which the rejected channels were interpolated using MNE-python. For all analyses, we focussed on the slow dynamics by filtering the z-scored, cleaned data between 0.5 and 8 Hz using a bidirectional FIR. This was done to keep the analysis close to earlier papers using the same data to study how EEG tracks acoustic and linguistic content of speech; but note that changing the filter parameters does not qualitatively change the results.

For more information on the dataset and prior analyses, see (Broderick et al., 2018).

Computational models

Word-by-word unpredictability was quantified via lexical surprise – or $-\log\left(p(word|context)\right)$ – estimated by GPT-2 and by a trigram language model. We will describe each in turn.

GPT-2

GPT-2 is a decoder-only variant of the *Transformer* (Vaswani et al., 2017). In the network, input tokens $U = (u_{i-k}, ..., u_{i-1})$ are passed through a token embedding matrix W_e after which a position embedding W_p is added to obtain the first hidden layer: $h_0 = UW_e + W_p$. Activities are then passed through a stack of transformer blocks, consisting of a multi-headed self attention layer, a position-wise feedforward layer, and layer normalisation (Fig 3.1a). This is repeated n times for each block b, after which (log)probabilities are obtained from a (log)softmax over the transposed token embedding of h_n:

$$h_b = \text{transformer\_block}\left(h_{b-1}\right) \forall i \in [1, n] \tag{3.1}$$

$$P(u_i|U) = \text{softmax}\left(h_n W_e^\top\right) \tag{3.2}$$

We used the largest public version of GPT-2 (345M parameter, released May 9)[1] which has a number of layers (blocks) of $n = 24$ and a context length of $k = 1024$. Note that k refers to the number of Byte-Pair Encoded *tokens*. A token can be either a word or (for less frequent words) a word-part, or punctuation. How many words actually fit into a context window of length k therefore depends on the text. We ran predictions on a run-by-run basis – each containing about 600 words, implying that in each run the entire preceding context was taken into account to compute a token's probability. For words spanning multiple tokens, word probabilities were simply the joint probability of the tokens obtained via the chain rule. The model was implemented in PyTorch with the Huggingface BERT module[2].

Trigram

As a comparison, we implemented an n-gram language model. N-grams also compute $p(w_i|w_{i-k}, ..., w_{i-1})$ but are simpler as they are based on counts. Here we used

[1] For more details on GPT-2, see https://openai.com/blog/better-language-models/ or Radford et al. (2019)
[2] see https://github.com/huggingface/pytorch-pretrained-BERT

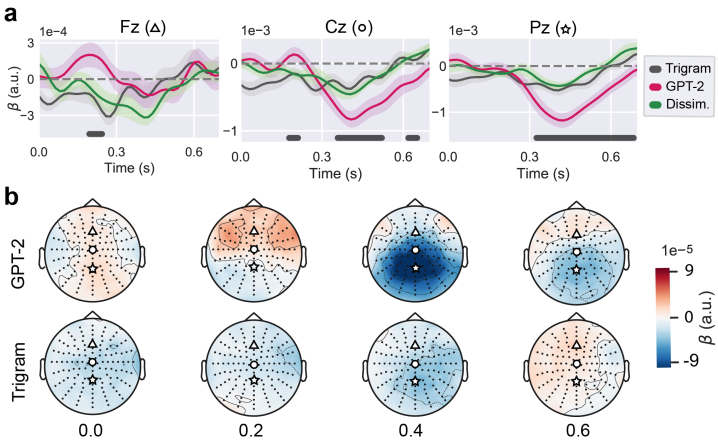

Figure 3.2. a) Grand averaged TRFs for trigram surprise,GTP-2 surprise and semantic dissimilarity for three channels of interest. At each time point, the GPT-2 TRF was compared to both the trigram and semantic dissimilarity TRF with a 2-tailed paired t-test; black bars indicating that both tests were significant at $p < 0.01$, FDR-corrected. Error bars indicate the between-subject SEM. **b)** Topographic maps of grand averaged TRFs for surprise, computed by GPT-2 (top) and the trigram language model (bottom).

a trigram ($k = 2$) – which was perhaps the most widely used language model before the recent rise of neural alternatives.[3] To deal with sparsity we used modified Knesner-Ney, the best-performing smoothing technique (Jurafsky and Martin, 2014). The trigram was implemented in NLTK and trained on its Gutenberg corpus, chosen to closely approximate the test set.

Non-predictive controls

We included two non-predictive and potentially confounding variables: first, frequency which we quantified as unigram surprise ($-\log p(w)$) which was based on a word's lemma count in the CommonCrawl corpus, obtained via spaCy. Second, following Broderick et al. (2018), we computed the semantic dissimilarity for each content word: $\text{dissim}(w_i) = 1 - \text{corr}\Big(\text{GloVe}(w_i), \frac{1}{n}\sum_{i=i}^{n}\text{GloVe}(c_i)\Big)$, where $(c_1, ..., c_n)$ are the content words preceding a word in the same or – if w_i is the first content word of the sentence – the previous sentence, and $\text{GloVe}(w)$ is the embedding. As shown by Broderick et al. (2018) this variable covaries with an N400-like component. However, it only captures how semantically dissimilar a word is from the preceding words (represented as an 'averaged bag of words'), and not how unexpected a

[3] While $k = 2$ might seem needlessly restrictive, training ngrams beyond $k = 2$ becomes exponentially difficult due to sparsity issues.

word is in its context, making it an interesting comparison, especially for predictive interpretations of the N400.

Time resolved regression

Variables were regressed against EEG data using time-resolved regression. Briefly, this involves temporally expanding a design matrix such that each predictor column C becomes a series of columns over a range of lags $C_{t_{min}}^{t_{max}} = (C_{t_{min}}, ..., C_{t_{max}})$. For each predictor one thus estimates a series of weights $\beta_{t_{min}}^{t_{max}}$ (Fig 3.1c) which, under some assumptions, corresponds to the isolated ERP that would have been obtained in an ERP paradigm. In all analyses, word onset was used as time-expanded intercept and other variables as covariates. All regressors were standardised and coefficients were estimated with Ridge regression. Regularisation was set at $\alpha = 1000$ since this lead to the highest R^2 in a leave-one-run-out CV procedure (Fig. 3.3) Analyses were performed using custom code adapted from MNE's linear_regression module.

Results

We first inspected our main regressor of interest: the surprise values computed by GPT-2, estimated with a regression model that included frequency (unigram surprise) and semantic dissimilarity as nuisance covariates. As can be seen in Figure 3.1C, the obtained TRF revealed a clear frontal positive response around 200 ms and a central/posterior negative peak at 400 ms after word onset. These peaks indicate that words that were more surprising to the network tended to evoke stronger positive responses at frontal channels at 200 ms and stronger negative potentials at central/posterior channels 400 ms after word onset. Note that while Figure 3.1C only shows the TRF obtained using one regularisation parameter, we found the same qualitative pattern for any alpha we tested.

We then compared this to an alternative regression model, in which the surprise regressor was based on the trigram model, but that was otherwise identical. Although the TRFs exhibited the same negativity at 400 ms, it was a lot weaker overall, as can be seen from Figure 3.2B. One anomalous feature is that the TRF is not at 0 at word onset. We suspect this is because 1) we only had onset times for content words, and not for function words typically preceding content words; and 2) for neighbouring words the log-probabilities from the trigram model were correlated ($\rho = 0.24$) but those from GPT-2 were not ($\rho = -0.002$), explaining why only the trigram TRF displays a baseline effect. Further analyses incorporating onset times for all words should correct this issue.

The negative surprise response at 400ms revealed by both the trigram and GPT is similar to the effect of semantic dissimilarity reported by Broderick et al. (2018) using

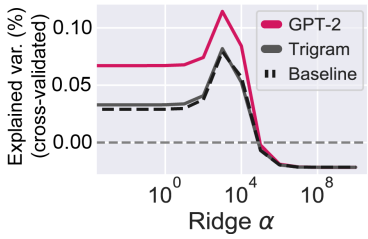

Figure 3.3. Predictive performance of three regression models. We compared a baseline regression model with only unigram surprise and semantic dissimilarity as covariates (dotted line) to two other models that also included surprise values, either obtained from the trigram model (grey) or from GPT-2 (red).

the same dataset. We therefore also looked at the TRF of semantic dissimilarity, for simplicity focussing on the three main channels of interest analysed by Broderick et al. (2018). At each time-point we compared the GPT-2 TRF to both the trigram and semantic dissimilarity TRF with a 2-tailed paired t-test to find time-points where both tests where significant at $\alpha = 0.01$ (FDR-corrected). As visible in Figure 3.2b, we observed timepoints in all three channels where the GPT-2 TRF was significantly more positive or negative than both other TRFs, confirming that the surprise values from the neural network covary more strongly with EEG responses than the other models.

Finally, to make sure that the difference in coefficients were not related to overfitting or some other estimation problem, we compared the predictive performance of the GPT-2 regression model to the alterntives using a leave-one-run-out cross-validation procedure. As can be seen in Figure 3.3, this revealed that cross-validated R^2 of the trigram regression model was not significantly higher than that of a baseline model that included only the two nuisance covariates (paired t-test, $t_{19} = -0.25, p = 0.8$); by contrast, R^2 of the GPT-2 regression model was significantly higher than both the trigram regression model (paired t-test, $t_{19} = 5.38, p = 4.1 \times 10^{-4}$) and the baseline model (paired t-test, $t_{19} = 3.10, p = 6.2 \times 10^{-3}$).

Discussion and conclusion

We have shown that word-by-word (un)predictability estimates obtained with a state-of-the-art self-attentional neural language model systematically covary with evoked brain responses to a naturalistic, continuous narrative, measured with EEG. When this relationship was plotted over time, we observed a frontal positive response at 200 ms, and a central negative response at 400 ms, akin to the N400. Unpredictability

estimates from the neural network were a much better predictor of EEG responses than those obtained from a trigram that was specifically trained on works of fiction, and than a non-predictive model of semantic incongruence, that simply computed the dissimilarity between a word and its context.

These results bear strong similarities to earlier work demonstrating a relationship between the N400 and semantic expectancy. However, we observed the responses in participants passively listening to naturalistic stimuli, without many highly expected or unexpected sentence endings typically used in the stimulus sets of traditional ERP studies. This suggests that linguistic predictability effects are not just a by-product of simple (prediction encouraging) designs, underscoring the importance of prediction in language processing.

Future analyses will aim at modelling all words, looking at different frequency bands, disentangling different forms of linguistic prediction (e.g. syntactic vs semantic), and trying to replicate these results in different, independent datasets.

Acknowledgments

We want to thank Michael Broderick and the Lalor lab for sharing the data, and all authors of open source software we used. This work was supported by NWO (Vidi grant to FdL, Research Talent grant to MH), the James S. McDonell Foundation (JSMF scholar award to FdL), and the EU Horizon 2020 Program (ERC starting grant 678286 to FdL).

Chapter 4

A hierarchy of linguistic predictions during natural language comprehension

Abstract

Understanding spoken language requires transforming ambiguous acoustic streams into a hierarchy of representations, from phonemes to meaning. It has been suggested that the brain uses prediction to guide the interpretation of incoming input. However, the role of prediction in language processing remains disputed, with disagreement about both the ubiquity and representational nature of predictions. Here, we address both issues by analysing brain recordings of participants listening to audiobooks, and using a deep neural network (GPT-2) to precisely quantify contextual predictions. First, we establish that brain responses to words are modulated by ubiquitous, probabilistic predictions. Next, we disentangle model-based predictions into distinct dimensions, revealing dissociable signatures of syntactic, phonemic and semantic predictions. Finally, we show that high-level (word) predictions inform low-level (phoneme) predictions, supporting hierarchical predictive processing. Together, these results underscore the ubiquity of prediction in language processing, showing that the brain spontaneously predicts upcoming language at multiple levels of abstraction.

This chapter is based on:
Heilbron M, Armeni K, Schoffelen JM, Hagoort P, de Lange FP. 2021. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*

Introduction

Understanding spoken language requires transforming ambiguous stimulus streams into a hierarchy of increasingly abstract representations, ranging from speech sounds to meaning. It is often argued that during this process, the brain relies on prediction to guide the interpretation of incoming information (Kuperberg and Jaeger, 2016; Kutas, DeLong, and Smith, 2011). Such a 'predictive processing' strategy has not only proven effective for artificial systems processing language (Graves, Mohamed, and Hinton, 2013; Jelinek, 1998), but has also been found to occur in neural systems in related domains such as perception and motor control and might constitute a canonical neural computation (Friston, 2005; Keller and Mrsic-Flogel, 2018).

There is a considerable amount of evidence that appears in line with predictive language processing. For instance, behavioural and brain responses are highly sensitive to violations of linguistic regularities (Hagoort, Brown, and Groothusen, 1993; Kutas and Hillyard, 1984) and to deviations from linguistic expectations more broadly (Armeni et al., 2019; Donhauser and Baillet, 2020; Henderson et al., 2016; Smith and Levy, 2013; Willems et al., 2016). While such effects are well-documented, two important questions about the role of prediction in language processing remain unresolved (Ryskin, Levy, and Fedorenko, 2020).

The first question concerns the *ubiquity* of prediction. While some models cast prediction as a routine, integral part of language processing (Fitz and Chang, 2019; Kuperberg and Jaeger, 2016; Levy, 2008), others view it as relatively rare, pointing out that apparent widespread prediction effects might instead reflect other processes like semantic integration difficulty (Brown and Hagoort, 1993; Huettig and Mani, 2016); or that such prediction effects might be exaggerated by the use of artificial, prediction-encouraging experiments focussing on highly predictable 'target' words (Huettig and Mani, 2016; Nieuwland, 2019). The second question concerns the representational nature of predictions: Does linguistic prediction occur primarily at the level of syntax (Brennan et al., 2020; Hale, 2001; Hale et al., 2018; Levy, 2008) or rather at the lexical (Fitz and Chang, 2019; Fleur et al., 2020), semantic (Federmeier, 2007; Rabovsky, Hansen, and McClelland, 2018) or the phonological level (Brodbeck, Hong, and Simon, 2018; Di Liberto et al., 2019; Donhauser and Baillet, 2020; Gagnepain, Henson, and Davis, 2012; Gwilliams et al., 2018)? ERP studies have described brain responses to violations of, and deviations from, both high and low-level expectations, suggesting prediction might occur at all levels simultaneously (Kuperberg and Jaeger, 2016; Nieuwland, 2019), although see (Nieuwland et al., 2018). However, it has been disputed whether these findings would generalise to natural language, where violations are rare or absent and with few highly predictable words. In these cases, prediction may be less relevant or might perhaps be limited to the most ab-

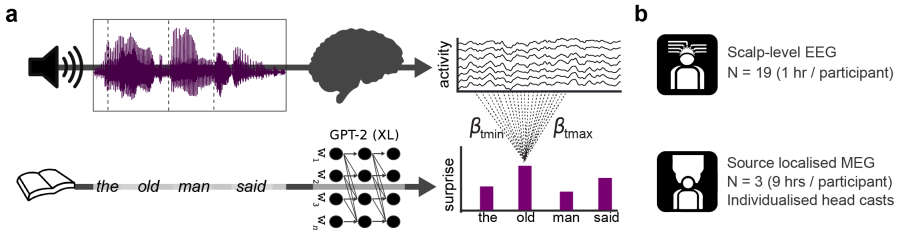

Figure 4.1. Schematic of experimental and analytical framework. a) Top row: in both experiments participants listened to continuous recordings from audiobooks while brain activity was recorded. Bottom row: the texts participants listened to were analysed by a deep neural network (GPT-2) to quantify the contextual probability of each word. A regression-based technique was used to estimate the effects of (different levels of) linguistic unexpectedness on the evoked responses within the continuous recordings. **b)** Datasets analysed: one group-level EEG dataset, and one individual subject source-localised MEG dataset.

stract levels (Huettig and Mani, 2016; Nieuwland, 2019; Nieuwland et al., 2018).

Here, we address both issues, probing the ubiquity and nature of linguistic prediction during natural language understanding. Specifically, we analysed brain recordings from two independent experiments of participants listening to audiobooks, and use a state-of-the-art deep neural network (GPT-2) to quantify linguistic predictions in a fine-grained, contextual fashion. First, we obtain evidence for predictive processing, confirming that brain responses to words are modulated by *probabilistic* predictions. Critically, the effects of prediction were found over and above those of non-predictive factors such as integration difficulty, and were not confined to a subset of predictable words, but were widespread – supporting the notion of *ubiquitous* prediction. Next, we investigated at which level prediction occurs. To this end, we disentangled the model-based predictions into distinct dimensions, revealing dissociable neural signatures of syntactic, phonemic and semantic predictions. Finally, we found that higher-level (word) predictions constrain lower-level (phoneme) predictions, supporting hierarchical prediction. Together, these results underscore the ubiquity of prediction in language processing, and demonstrate that prediction is not confined to a a single level of abstraction but occurs throughout the language network, forming a hierarchy of predictions across all levels of analysis, from phonemes to meaning.

Results

We consider data from two independent experiments, in which brain activity was recorded while participants listened to natural speech from audiobooks. The first experiment is part of a publicly available dataset (Broderick et al., 2018), and contains

1 hour of electroencephalographic (EEG) recordings in 19 participants. The second experiment collected 9 hours of magneto-encephalographic (MEG) data in three individuals, using individualised head casts that allowed us to localise the neural activity with high precision. While both experiments had a similar setup (see Figure 4.1), they yield complementary insights, both at the group level and in three individuals.

Neural responses to speech are modulated by probabilistic linguistic predictions

We first tested for evidence for linguistic prediction in general. We reasoned that if the brain is constantly predicting upcoming language, neural responses to words should be sensitive to violations of contextual predictions, yielding 'prediction error' signals which are considered a hallmark of predictive processing (Keller and Mrsic-Flogel, 2018). To this end, we used a regression-based deconvolution approach to estimate the effects of prediction error on evoked responses within the continuous recordings. We focus on this event-related, low-frequency evoked response because it connects most directly to earlier influential neural signatures of prediction in language (Frank et al., 2015; Kutas and Hillyard, 1984; Nieuwland et al., 2018; Van Petten and Luka, 2012).

To quantify linguistic predictions, we analysed the books participants listened to with a state-of-the-art neural language model: GPT-2 (Radford et al., 2019). GPT-2 is a large transformer-based model that predicts the next word given the previous words, and is currently among the best publicly-available models of its kind. Note that we do not use GPT-2 as a model of human language processing, but purely as a tool to quantify how expected each word is in context.

To test whether neural responses to words are modulated by contextual predictions, we compared three regression models (see S4.5). The baseline model formalises the hypothesis that natural, passive language comprehension does not invoke prediction. This model did not include regressors related to contextual predictions, but did include several potentially confounding variables (such as word frequency, semantic integration, and acoustics). The *constrained guessing* model formalised the hypothesis that language processing *sometimes* (in constraining contexts) invokes prediction, and that such predictions are an all-or-none phenomenon – together representing how the notion of prediction was classically used in the psycholinguistic literature (Van Petten and Luka, 2012). This model included all non-predictive variables from the baseline model, plus, in constraining contexts, a linear estimate of word improbability (since all-or-none predictions result in a linear relationship between word probability and brain responses; see methods for details). Finally, the *probabilistic prediction* model included all confounding regressors from the baseline model, plus

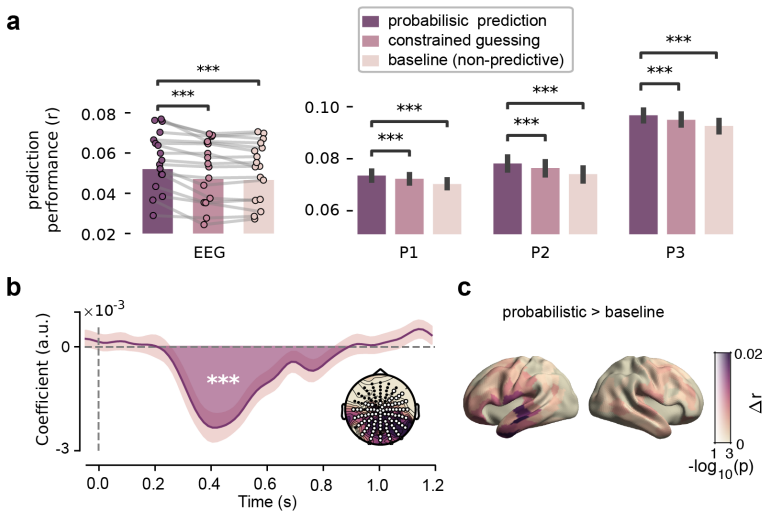

Figure 4.2. Neural responses are modulated by probabilistic predictions. a) Model comparison. Cross-validated correlation coefficients for EEG (left) and each MEG participant (right). EEG: dots with connecting lines represent individual participants (averaged over all channels). MEG: bars represent median across runs, bars represent bootstrapped absolute deviance (averaged over language network sources). **b)** EEG: coefficients describing the significant effect of lexical surprise (see Figure S4.3 for the full topography over time). Highlighted area indicates extent of the cluster, shaded error bar indicates bootstrapped SE. Inset shows distribution of absolute t-values and of channels in the cluster. **c)** Difference in prediction performance across cortex (transparency indicates FWE-corrected p-values). Significance levels correspond to P<0.001 (***) in a two-tailed one-sample Student's *t* or Wilcoxon sign rank test.

for every word a logarithmic estimate of word improbability (i.e. *surprise*). This formalises the hypothesis that the brain constantly generates *probabilistic predictions*, as proposed by predictive processing accounts of language (Frank et al., 2015; Kuperberg and Jaeger, 2016) and of neural processing more broadly (Friston, 2005; Keller and Mrsic-Flogel, 2018).

When we compared the ability of these models to predict brain activity using cross-validation, we found that the probabilistic prediction model performed better than both other models (see Figure 4.2a). The effect was highly consistent, found in virtually all EEG participants (probabilistic vs constrained guessing, $t_{18} = 5.34$, $p = 4.46 \times 10^{-5}$; probabilistic vs baseline, $t_{18} = 6.43$, $p = 4.70 \times 10^{-6}$) and within each MEG participant (probabilistic vs constrained guessing, all $p's < 1.54 \times 10^{-6}$; probabilistic vs baseline, all $p's < 5.17 \times 10^{-12}$).

As the *constrained guessing* model differed from the probabilistic model in two ways – by assuming that predictions are (i) categorical and (ii) limited to constraining contexts – we also considered a control model. Like the constrained guessing model,

this extended guessing model included a linear estimate of word probability, but for every word rather than only for constraining contexts. Although this model did not outperform the probabilistic prediction model, it did substantially outperform the constrained model (Fig S4.5). This demonstrates that the effects of prediction are not limited to constraining contexts, but apply much more broadly – in line with the idea that predictions are ubiquitous and automatic.

Having established that word unexpectedness modulates neural responses, we characterised this effect in space and time. In the MEG dataset, we asked for which neural sources lexical surprise was most important in explaining neural data, by comparing the prediction performance of the baseline model to the predictive model in a spatially resolved manner. This revealed that overall word unexpectedness modulated neural responses throughout the language network (see Figure 4.2c). To investigate the temporal dynamics of this effect, we inspected the regression coefficients, which describe how fluctuations in lexical surprise modulate the neural response at different time lags – together forming a modulation function also known as the *regression evoked response* (Smith and Kutas, 2015) or Temporal Response Function (TRF) (Brodbeck, Hong, and Simon, 2018; Ding and Simon, 2012). When we compared these across participants in the EEG experiment, cluster-based permutation tests revealed a significant effect ($p = 2 \times 10^{-4}$) based on a posterio-central cluster with a negative polarity between 0.2 and 0.9 seconds (see Figure 4.2b and S4.8). This indicates that surprising words lead to a stronger negative deflection of evoked responses, an effect peaking at 400 ms post word onset and strongly reminiscent of the classic N400 (Kutas and Hillyard, 1984; Nieuwland et al., 2018; Rabovsky, Hansen, and McClelland, 2018). Coefficients for MEG subjects revealed a similar, slow effect at approximately the same latencies (see Fig S4.4).

Together, these results constitute clear evidence for predictive processing by confirming that brain responses to words are modulated by predictions. These modulations are not confined to constraining contexts, occur throughout the language network, evoke an effect reminiscent of the N400, and are best explained by a probabilistic account of prediction. This suggests the brain predicts constantly and probabilistically – even when passively listening to natural language.

Linguistic predictions are feature-specific

The results so far revealed modulations of neural responses by *overall* word unexpectedness. What type of linguistic prediction might be driving these effects? Earlier research suggests a range of possibilities, with some proposing that the effect of overall word surprise primarily reflects syntax (Hale, 2001; Levy, 2008), while others propose that prediction unfolds at the semantic (Federmeier, 2007; Rabovsky,

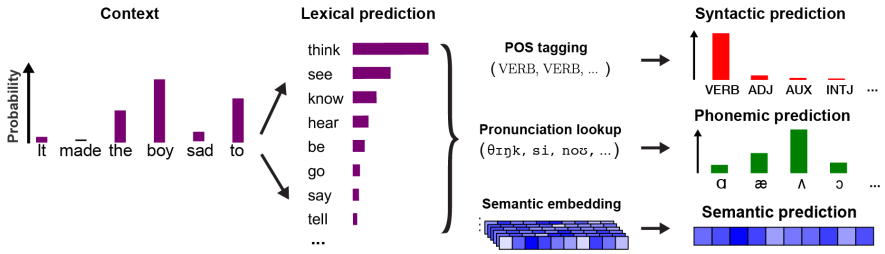

Figure 4.3. Partitioning model-derived predictions into distinct linguistic dimensions. To disentangle syntactic, semantic and phonemic predictions, the lexical predictions from GPT-2 were analysed. For the syntactic prediction, part-of-speech was tagging performed over all potential sentences (e.g. "It made the boy sad to *think*"). To compute the phonemic prediction, each predicted word was decomposed into its constituent phonemes, and the predicted probabilities were used as a contextual prior in a phoneme model (see Figure 4.6). For the semantic prediction, a weighted average was computed over the GLoVE embeddings of all predicted words.

Hansen, and McClelland, 2018), or the phonemic level (Brodbeck, Hong, and Simon, 2018; Donhauser and Baillet, 2020; Gagnepain, Henson, and Davis, 2012) – or at all levels simultaneously (Kuperberg and Jaeger, 2016).

To evaluate these possibilities, we factorised the aggregate, word-level linguistic predictions from the artificial neural network into distinct linguistic dimensions (Fig 4.3). This allows us to derive model-based estimates of three feature-specific predictions: the syntactic prediction (defined as the conditional probability distribution over parts-of-speech, given context), semantic prediction (defined as the predicted semantic embedding) and phonemic prediction (i.e. the conditional probability of the next phoneme, given the phonemes within the word so far and the prior context). By comparing these predictions to the presented words, we derived *feature-specific prediction errors* which quantified not just the extent to which a word is surprising overall, but also in what way: semantically, syntactically or phonemically (see Methods for definitions).

We reasoned that if the brain is generating predictions at a given level (e.g. syntax), then the neural responses should be sensitive to prediction errors specific to this level. Moreover, because these different features are processed by partly different brain areas over different timescales, the prediction errors should be at least partially dissociable. To test this, we formulated a new regression model (Figure S4.6). This included all variables from the lexical prediction model as nuisance regressors, and added three regressors of interest: syntactic surprise (defined for each word), semantic prediction error (defined for each content word), and phonemic surprise (defined for each word-non-initial phoneme).

Because these regressors were to some degree correlated, we first asked whether, and in which brain area, each of the feature-specific prediction errors explained any unique variance, not explained by the other regressors. In this analysis, we turn to the MEG data because of its spatial specificity. As a control, we first performed the analysis for a predictor with a known source: the acoustics. This revealed a clear peak around auditory cortex (Fig S4.7) especially in the right hemisphere. This aligns with prior work (Abrams et al., 2008) and confirms that this approach can localise which areas are especially sensitive to a given regressor. We then tested the three prediction errors, finding that each type of prediction error explained significant unique variance in each individual (Figure 4.4), except in participant 1 where phonemic surprise did not survive multiple comparisons correction (but see Figure 4.6c and Discussion). This shows that the brain responds differently to different types of prediction errors, implying that linguistic predictions are feature-specific and occur both at high and low levels of processing simultaneously.

Although we observed considerable variation in lateralisation and exact spatial locations between individuals, the overall pattern of sources aligned well with prior research on the neural circuits for each level. For instance, only for semantic prediction errors we observed a widely distributed set of neural sources – consistent with the fact that the semantic (but not the syntactic or phonological) system is widely distributed (Binder et al., 2009; Huth et al., 2016). Moreover, the temporal areas showing the strongest effect of syntactic surprise are indeed key areas for syntactic processing (Matchin and Hickok, 2020) and for the posterior temporal areas predictive syntax in particular (Brennan et al., 2020; Lopopolo et al., 2017; Matchin et al., 2019; Nelson et al., 2017) – though a clear syntactic effect in the inferior frontal gyrus (IFG) was interestingly absent. When we compared the sources of phonemic surprise to those obtained for lexical surprise, we observed a striking overlap in all individuals (see Chapter S4.7, S4.4 and S4.13), suggesting that the phonemic predictions as formalised here mostly relate to predictive (incremental) word recognition at the phoneme level rather than describing phonological or phonotactic predictions *per se*.

Dissociable signatures of syntactic, semantic and phonemic predictions

Having established that syntactic, phonemic and semantic prediction errors independently modulated neural responses in different brain areas, we further investigated the nature of these effects. This was done by inspecting the coefficients (or modulation functions), which describe how fluctuations in a given regressor modulate the response over time. We first turn to the EEG data because there the sample size allows for population-level statistical inference on the coefficients. We fitted the same integrated model (Figure S4.6) and performed cluster-based permutation tests on the

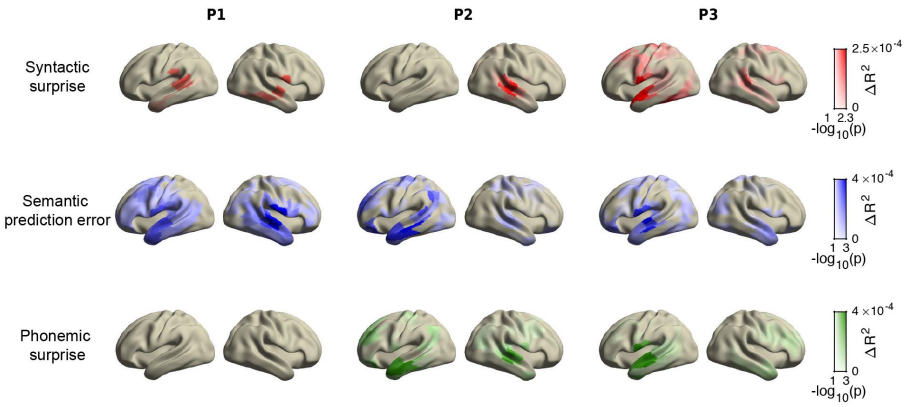

Figure 4.4. Dissociable patterns of explained variance by syntactic, semantic and phonemic predictions. Unique variance explained by syntactic, semantic and phonemic unexpectedness (quantified via surprise or prediction error) across cortical sources in each MEG participant. In all plots, colour indicates amount of additional variance explained; opacity indicates FWE-corrected statistical significance. Note that $p < 0.05$ is equivalent to $-\log_{10}(p) > 1.3$.

modulation functions. This revealed significant effects for each type of prediction error (Figure 4.5).

First, syntactic surprise evoked an early, positive deflection ($p = 0.027$) based on a frontal cluster between 200 and 500 ms. This early frontal positivity converges with two recent studies that investigated specifically syntactic prediction using models trained explicitly on syntax (Brennan and Hale, 2019; Hale et al., 2018). We also observed a late negative deflection for syntactic surprise ($p = 0.025$; Figure S4.9), but this was neither in line with earlier findings nor replicated in the MEG data. The semantic prediction error also evoked a positive effect ($p = 9.1 \times 10^{-3}$) but this was based on a much later, spatially distributed cluster between 600 and 1100 ms. Although such a late positivity has been prominently associated with *syntactic* violations (Hagoort, Brown, and Groothusen, 1993), there is also a considerable body of work reporting such late positivities for purely semantic anomalies (Herten, Kolk, and Chwilla, 2005) which is more in line with the semantic prediction error as quantified here (see Discussion). Notably, we did not find a significant N400-like effect for semantic prediction error – possibly because this negative deflection was already explained by the overall lexical surprise, which was included as a nuisance regressor (Figure S4.10). Finally, the phonemic surprise evoked a negative effect ($p = 3 \times 10^{-4}$) based on an early, distributed cluster between 100 and 500 ms. This effect was similar to the word-level surprise effect (Figure 4.2C and S4.10) but occurred earlier. This timecourse corresponds to recent studies using similar regression-based techniques

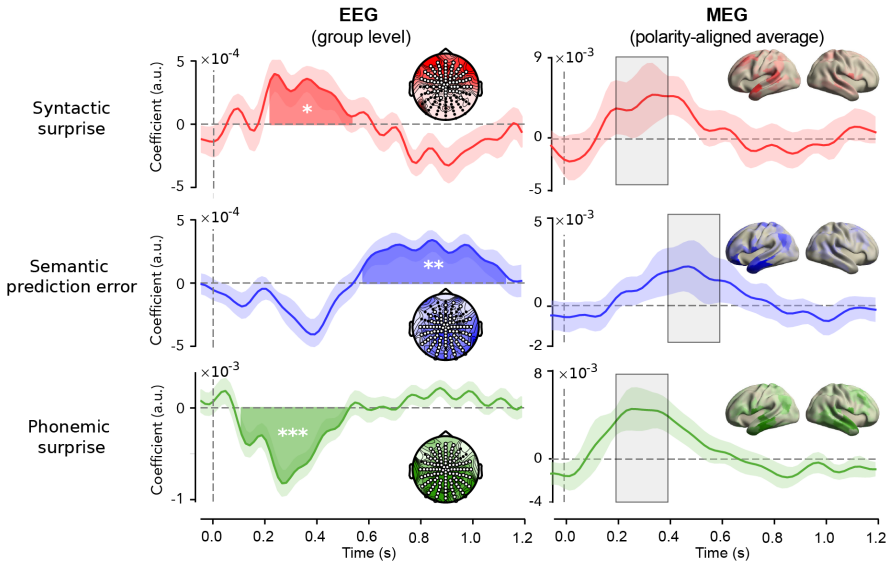

Figure 4.5. Spatiotemporal signatures of syntactic, semantic and phonemic prediction errors. Coefficients describing the effects of each prediction-error. EEG (left column): modulation functions averaged across the channels participating for at least one sample in the three main significant clusters (one per predictor). Highlighted area indicates temporal extent of the cluster. Shaded area around waveform indicates bootstrapped standard errors. Stars indicate cluster-level significance; $p < 0.05$ (*) , $p < 0.05$ (**), $p < 0.001$ (***). Insets represent selected channels and distribution of absolute t-values. Note that these plots only visualise the *effects*; for the full topographies of the coefficients and respective statistics, see Figure S4.8. MEG (right column): polarity aligned responses averaged across the sources with significant explained variance (Figure 4.4) across participants. Shaded area represents absolute deviation. Insets represent topography of absolute value of coefficients averaged across the highlighted period. Note that due to polarity alignment, sign information is to be ignored for the MEG plots. For average coefficients for each source, see Figure S4.10; for coefficients of each individual, see Figs S4.11 - S4.14.

to study (predictive) phoneme processing in natural listening (Di Liberto et al., 2019; Donhauser and Baillet, 2020; Gwilliams et al., 2020).

When we performed the same analysis on the MEG data, we observed striking differences in the exact shape and timing of the modulation functions between individuals (see Figure S4.11- S4.14). While this might partly reflect variance in the coefficients due to inherent correlations between the variables, it clearly also reflects true individual differences, demonstrated by one of the strongest and least correlated regressors (the acoustics) also showing considerable variability (see Figure S4.14). Overall however, we could recover a temporal pattern of effects similar to the EEG results: phonemic and syntactic surprise modulating early responses, and semantic prediction error modulating later responses – although not as late in the EEG

data. This temporal order holds on average (Figures 4.5, S4.10) and is especially clear within individuals (Figure S4.11 - S4.13).

Overall, our results (Figure 4.4,4.5) demonstrate that syntactic, phonemic and semantic prediction errors evoke brain responses that are both temporally and spatially dissociable. Specifically, while phonemic and syntactic predictions modulate relatively early neural responses (100-400 ms) in a set of focal temporal (and frontal) areas that are key for syntactic and phonetic/phonemic processing, semantic predictions modulate later responses (>400 ms) across a widely distributed set of areas across the distributed semantic system. These results reveal that linguistic prediction is not implemented by a single system but occurs throughout the speech and language network, forming a hierarchy of linguistic predictions across all levels of analysis.

Phoneme predictions reveal hierarchical inference

Having established that the brain generates linguistic predictions across multiple levels of analysis, we finally asked whether predictions at different levels might interact. One option is that they are encapsulated: Predictions in separate systems might use different information, for instance unfolding over different timescales, rendering them independent. Alternatively, predictions at different levels might inform and constrain each other, effectively converging into a single multilevel prediction – as suggested by theories of hierarchical cortical prediction (Friston, 2005; Keller and Mrsic-Flogel, 2018; Kiebel, Daunizeau, and Friston, 2008).

One way to adjudicate between these hypotheses is by evaluating different schemes of deriving phoneme predictions. One possibility is that such predictions are only based on information unfolding over short timescales. In this scheme, the predicted probability of the next phoneme is derived from the *cohort* of words that are compatible with the phonemes presented so far, with each candidate word weighted by its overall frequency of occurrence (see Figure 4.6A). As such, this scheme proposes a *single-level model*: phoneme predictions are based only on information at the level of within-word phoneme sequences unfolding over short timescales, plus a fixed frequency-based prior (capturing statistical knowledge of word frequencies within a language).

Alternatively, phoneme predictions might not only be based on sequences of phonemes within a word, but also on the longer prior linguistic context. In this case, the probability of the next phoneme would still be derived from the cohort of words compatible with the phonemes presented so far, but now each candidate word is not weighted by its overall frequency but by its *contextual probability* (Figure 4.6A). Such a model would be hierarchical, in the sense that predictions are based both –

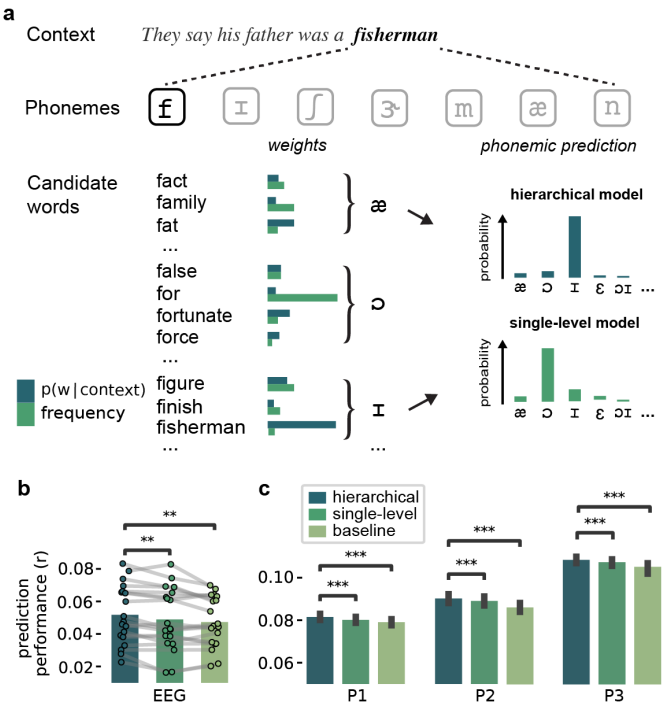

Figure 4.6. Evidence for hierarchical inference during phoneme prediction. a) Two models of phoneme prediction during incremental word recognition. Phonemic predictions were computed by grouping candidate words by their identifying next phoneme, and weighting each candidate word by its prior probability. This weight (or prior) could be either based on a word's overall probability of occurrence (i.e. frequency) or on its conditional probability in that context (from GPT-2). Critically, in the frequency-based model, phoneme predictions are based on a single level: short sequences of within words phonemes (hundreds of ms long) plus a fixed prior. By contrast, in the contextual model, predictions are based not just on short sequences of phonemes, but also on a contextual prior which is itself based on long sequences of prior words (up to minutes long), rendering the model hierarchical (see Methods). **b-c)** Model comparison results in EEG (**b**) and all MEG participants (**c**). EEG: dots with connecting lines represent individual participants (averaged over all channels). MEG: bars represent median across runs, error bars represent bootstrapped absolute deviance (averaged over language network sources). Significance levels correspond to P<0.01 (**) or P<0.001 (***) in a two-tailed paired *t* or Wilcoxon sign rank test.

at the first level – on short sequences of phonemes (i.e. of hundreds of milliseconds long), and on a contextual prior which itself is based – at the higher level – on long sequences of words (i.e. of tens of seconds to minutes long).

Here, the first model is more in line with the classic Cohort model of incremental (predictive) word recognition, which suggests that context is only integrated after the selection and activation of lexical candidates (Marslen-Wilson, 1989). By contrast,

the second model is more in line with contemporary theories of hierarchical predictive processing which propose that high-level cortical predictions (spanning larger spatial or temporal scales) inform and shape low-level predictions (spanning finer spatial or temporal scales) (Kiebel, Daunizeau, and Friston, 2008; Rao and Ballard, 1999). Interestingly, recent studies of phoneme predictions during natural listening have used both the frequency-based single level model (Brodbeck, Hong, and Simon, 2018; Gwilliams et al., 2018) and a context-based (hierarchical) model (Donhauser and Baillet, 2020). However, the models have not been explicitly compared to test which model can best account for prediction-related fluctuations in neural responses to phonemes.

To compare these possibilities, we constructed 3 phoneme-level regression models (see Figure S4.15), which all only included regressors at the level of phonemes. First, the baseline model only included non-predictive control variables: phoneme onsets, acoustics, word boundaries and uniqueness points. This can be seen as the phoneme-level equivalent of the baseline model in Figures 4.2,S4.5. The baseline model was compared with two regression models which additionally included phoneme surprise. In one of the regression models, this was calculated using a single-level model (with a fixed, frequency-based prior), in the other regression model it was derived from a hierarchical model (with a dynamic, contextual prior derived from GPT-2). To improve our ability to discriminate between the hierarchical and single-level model, we not only included surprise but also phoneme entropy (calculated with either model) as a regressor (Donhauser and Baillet, 2020).

When we compared the cross-validated predictive performance, we first found that in both datasets the predictive model performed significantly better than the non-predictive baseline (Figure 4.6b-c hierarchical vs baseline, EEG: $t_{18} = 3.80$, $p = 1.31 \times 10^{-3}$; MEG: all $p's < 5.69 \times 10^{-12}$). This replicates the basic evidence for predictive processing but now at the phoneme rather than word level (Figure 4.2). Critically, when we compared the two predictive models, we found that the hierarchical model performed significantly better, both in EEG ($t_{18} = 3.03, p = 7.28 \times 10^{-3}$) and MEG (all $p's < 9.44 \times 10^{-4}$). This suggests that neural predictions of phonemes (based on short sequences of within-word speech sounds) are are informed by lexical predictions, effectively incorporating long sequences of prior words as contexts. This is a signature of hierarchical prediction, supporting theories of hierarchical predictive processing.

Discussion

Across two independent data sets, we combined deep neural language modelling with regression-based deconvolution of human electrophysiological (EEG and MEG)

recordings to ask if and how evoked responses to speech are modulated by linguistic expectations that arise naturally while listening to a story. Our results demonstrated that evoked responses are modulated by *probabilistic* predictions. We then introduced a novel technique that allowed us to quantify not just how much a linguistic stimulus is surprising, but also at what level – phonemically, syntactically and/or semantically. This revealed dissociable effects, in space and time, of different types of prediction errors: syntactic and phonemic prediction errors modulated early responses in a set of focal, mostly temporal areas, while semantic prediction errors modulated later responses across a widely distributed set of cortical areas. Finally, we found that phonemic prediction error signals were best modelled by a hierarchical model incorporating two levels of context: short sequences of within-word phonemes (up to hundreds of milliseconds long) and long sequences of prior words (up to minutes long). Together, these results demonstrate that during natural listening, the brain is engaged in prediction across multiple levels of linguistic representation, from speech sounds to meaning. The findings underscore the ubiquity of prediction during language processing, and fit naturally in predictive processing accounts of language (Kuperberg and Jaeger, 2016; Kutas, DeLong, and Smith, 2011) and neural computation more broadly (Friston, 2005; Heilbron and Chait, 2018; Keller and Mrsic-Flogel, 2018; Rao and Ballard, 1999).

A primary result of this paper is that evoked responses to words are best explained by a predictive processing model: regression models including unexpectedness performed better than strong non-predictive baseline models, demonstrating that the effects of prediction on brain responses cannot be reduced to confounding simple features like semantic incongruency. This aligns with recent ERP studies aimed specifically at distinguishing prediction from semantic integration (Mantegna et al., 2019; Nieuwland et al., 2020) and extends those findings by analysing not just specific (highly predictable) 'target' words, but *all* words in a natural story. Indeed, when we further compared different accounts of prediction, responses were best explained by a regression model casting linguistic predictions as ubiquitous and probabilistic. This supports the notion of continuous, graded prediction – as opposed to the classical view of prediction as the all-or-none pre-activation of specific words in highly constraining contexts (Van Petten and Luka, 2012).

Because our deconvolution analysis focussed on evoked responses, the results can be linked to the rich literature on linguistic violations using traditional ERP methods. This is powerfully illustrated by the modulation function of lexical surprise (Figure 4.2b) tightly following the N400 modulation effect, one of the first proposed, most robust and most debated ERP signatures of linguistic prediction (Kutas and Hillyard, 1984; Nieuwland et al., 2018; Rabovsky, Hansen, and McClelland, 2018). Similarly, the early negativity we found for phonemic surprise and later positivity for semantic

prediction error (Fig 4.5) align well with N200 and the semantic P600 or PNP effects of phonological mismatch and semantic anomaly respectively (Brink, Brown, and Hagoort, 2001; Van Petten and Luka, 2012). Unlike most ERP studies, we observed these effects in participants listening to natural stimuli – without any anomalies or violations – not engaged in any task. This critically supports the idea that these responses reflect deviations from *predictions* inherent to the comprehension process – rather than reflecting either detection of linguistic anomalies or expectancy effects introduced by the experiment (Huettig and Mani, 2016; Nieuwland, 2019).

While we found several striking correspondences between the modulation functions recovered from the data and classic effects from the ERP literature, there were also some differences. Specifically, for syntactic surprise, we found neither a late positive effect resembling the syntactic P600 (Hagoort, Brown, and Groothusen, 1993) nor an early negative effect akin to the ELAN (Friederici, 2002). One potential explanation for this is that our formalisation (part-of-speech surprise) might not fully capture syntactic violations used in ERP studies. Indeed, a recent paper on syntactic prediction using a similar model-based approach found a P600-like effect not for syntactic surprise but for the number of syntactic reinterpretation attempts a word induced (Hale et al., 2018). Conversely, the early positive effect of syntactic surprise we found – which replicated other model-based findings, despite using a different formalisation of syntactic surprise (Brennan and Hale, 2019; Hale et al., 2018) – does not have a clear counterpart in the traditional ERP literature. Better understanding such systematic differences between the traditional experimental and model-based approach provides an interesting challenge for future work.

Beyond the ERP literature, there has also been earlier model-based work on prediction. However, these studies have mostly quantified feature-unspecific lexical unexpectedness (Armeni et al., 2019; Frank et al., 2015; Heilbron et al., 2019; Weissbart, Kandylaki, and Reichenbach, 2020; Willems et al., 2016) or modelled feature-specific predictions at a single level such as syntax (Brennan and Hale, 2019; Hale et al., 2018; Henderson et al., 2016; Shain et al., 2020), phonemes (Brodbeck, Hong, and Simon, 2018; Di Liberto et al., 2019; Donhauser and Baillet, 2020) or semantics (Rabovsky, Hansen, and McClelland, 2018). We extend these studies by probing predictions at all these levels simultaneously. This is important because it allows to control for correlations between levels – since words that are, for instance, syntactically surprising are, on average, also semantically surprising. Moreover, prior modelling of feature-specific predictions used domain-specific models that had to be independently trained, and typically incorporated linguistic context in a limited way. By contrast, our method (Figure 3) allows to derive multiple predictions from a single, large pre-trained model (like GPT-2) which has a much deeper grasp of linguistic context. However, a limitation of this method is that the resulting predictions are

not independent. Therefore, you cannot test if levels interact without *also* creating a separate, domain-specific model. As such, the disentangling approach we used is complementary to the domain-specific modelling approach. Future work could combine the two, for instance to test if the hierarchical prediction we observed for phonemes applies to all linguistic levels – or whether predictions at some levels (e.g. syntax) might be independent.

In this study, we combined group-level analysis (of the EEG data) and individual-level analysis (of the MEG data). These approaches are complementary. While including more participants allows one to perform population-level inference, acquiring more data per participant allows one to evaluate effects within individuals. By combining both forms of analysis, we found that on the one hand, the basic effects of prediction and the comparison of hypotheses about its computational nature (probabilistic prediction, hierarchical prediction) were identical within and across each individual (Figure 2, 6, S5). But on the other hand, the exact spatiotemporal characteristics of these effects showed substantial variability (Figure 4, 5, S4, S7-S14). This suggest that while the prediction effects themselves at the EEG group-level are likely present in each individual, the precise spatiotemporal signatures (Figure 5) are probably best understood as a statistical average that is not necessarily representative of underlying individuals.

Because our analysis focused on evoked responses, we chose to probe predictions indirectly: via the neural markers of deviations from these predictions. As such, we cannot rule out that the effects might partly reflect 'postdiction'. However, a purely postdictive explanation appears unlikely as it implies that after recognition, the brain computes a prediction of the recognised stimulus based on information available *before* recognition. While the data therefore indirectly support pre-activation, the representational format of these pre-activations is still an open question. In our analyses – and many theoretical models (Friston, 2005; Rao and Ballard, 1999)) – predictions are formalised as *explicit* probability distributions, but this is almost certainly a simplification. It remains unclear whether the brain represents probabilities implicitly. Alternatively, it might use a kind of approximation: graded, anticipatory processing that is perhaps functionally equivalent to probabilistic processing, but avoids having to represent (and compute with) probabilities. A potential way to address this question is to try to decode predictions before word onset (Goldstein et al., 2021). Interestingly, this approach could be extended to assess whether predicted probabilities are represented before onset at different levels of the linguistic hierarchy, to test whether and which predicted distributions are reflected in pre-stimulus activity.

Why would the brain constantly predict upcoming language? Three – mutually non-exclusive – functions have been proposed. First, predictions can be used for *compression*: if predictable stimuli are represented succinctly, this yields an efficient

code (Friston, 2005; Rao and Ballard, 1999) – conversely, optimising efficiency can make predictive coding emerge in neural networks (Ali et al., 2021). A second, perhaps more studied function is that predictions can guide *inference*. Our analysis only probed prediction errors, and hence does not speak directly to such inferential effects of prediction – but earlier work suggests that linguistic context can indeed enhance neural representations in a top-down fashion (Broderick, Anderson, and Lalor, 2019; Heilbron et al., 2020); but see (Blank and Davis, 2016; Sohoglu and Davis, 2020). Finally, predictions may guide *learning*: prediction errors can be used to perform error-driven learning without supervision. While learning is perhaps the least-studied function of linguistic prediction in cognitive neuroscience (but see (Fitz and Chang, 2019)), it is its primary application in Artificial Intelligence (Manning et al., 2020; McClelland et al., 2020). In fact, the language model we used (GPT-2) was created to study such predictive learning. These models are trained only to predict words, but learn about language more broadly, and can then be applied to practically any linguistic task (Manning et al., 2020; Radford et al., 2019). Interestingly, models trained with this predictive objective also develop representations that are 'brain-like', in the sense that they are currently the best encoders of linguistic stimuli to predict brain responses (Caucheteux and King, 2020; Jain and Huth, 2018; Schrimpf et al., 2020; Toneva and Wehbe, 2019). And yet, these predictive models are also brain-unlike in an interesting way – they predict upcoming language only at a single (typically lexical) level.

When prediction is used for compression or inference, it seems useful to predict at multiple levels, since redundancies and ambiguities also occur at multiple levels. But if predictions drive learning, why would the brain predict at multiple levels, when effective learning can be achieved using simple, single-level prediction? One fascinating option is that it might reflect the brain's way to perform credit assignment within biological constraints. In artificial networks, credit assignment is typically done by first *externally* computing a single, global error term, and then 'backpropagating' this error through all levels of the network – but both these steps are biologically implausible (Whittington and Bogacz, 2017). Interestingly, it has been shown that hierarchical predictive coding networks can approximate or even implement classical backpropagation while using only Hebbian plasticity and local error computation (Friston, 2005; Millidge, Tschantz, and Buckley, 2020; Whittington and Bogacz, 2017). Therefore, if the brain uses predictive error-driven learning, one might expect such prediction to be hierarchical, so error-terms can be locally computed throughout the hierarchy – which is in line with what we find.

Beyond the domain of language, there have been other reports of hierarchies of neural prediction, but these have been limited to artificial, predictive tasks or to restricted representational spans, such as successive stages in the visual system (Issa,

Cadieu, and DiCarlo, 2018; Schwiedrzik and Freiwald, 2017; Wacongne et al., 2011). Our results demonstrate that even during passive listening of natural stimuli, the brain is engaged in prediction across disparate levels of abstraction (from speech sounds to meaning) based on timescales separated by three orders of magnitude (hundreds of milliseconds to minutes). These findings provide important evidence for hierarchical predictive processing in cortex. As such, they highlight how language processing in the brain is shaped by a domain-general neurocomputational principle: the prediction of perceptual inputs across multiple levels of abstraction.

Methods

We analysed EEG and source localised MEG data from two experiments. The EEG data is part of a public dataset that has been published about before (Brodbeck, Hong, and Simon, 2018).

Participants

All participants were native English speakers. In the EEG experiment, 19 subjects (13 male) between 19 and 38 years old participated; in the MEG experiment, 3 subjects participated (2 male) aged 35, 30, and 28. Both experiments were approved by local ethics committees (EEG: ethics committee of the School of Psychology at Trinity College Dublin; MEG: CMO region Arnhem-Nijmegen).

Stimuli and procedure

In both experiments, participants were presented continuous segments of narrative speech extracted from audiobooks. The EEG experiment used a recording of Hemingway's *The Old Man and the Sea*. The MEG experiment used 10 stories from the *The Adventures of Sherlock Holmes* by Arthur Conan Doyle. In total, EEG subjects listened to 1 hour of speech (containing 11,000 words and 35,000 phonemes); MEG subjects listened to 9 hours of speech (containing 85,000 words and 290,000 phonemes).

In the EEG experiment, each participants performed only a single session, which consisted of 20 runs of 180s long, amounting to the first hour of the book. Participants were instructed to maintain fixation and minimise movements but were otherwise not engaged in any task.

In the MEG experiment, each participant performed a total of ten sessions, each 1 hour long. Each session was subdivided in 6-7 runs of roughly ten minutes, although the duration varied as breaks only occurred at meaningful moments (making sure, for example, that prominent narrative events were not split across runs). Unlike in the EEG experiment, participants in the MEG dataset participants were asked to listen

attentively and had to answer questions in between runs: one multiple choice comprehension question, a question about story appreciation (scale 1-7) and a question about informativeness.

MRI acquisition and headcast construction

To produce the headcast, we needed to obtain accurate images of the participants's scalp surface, which were obtained using structural MRI scans with a 3T MAGNETOM Skyra MR scanner (Siemens AG). We used a fast low angle shot (FAST) sequence with the following image acquisition parameters: slice thickness of 1 mm; field-of-view of 256 × 256 × 208 mm along the phase, read, and partition directions respectively; TE/TR = 1.59/4.5 ms.

Data acquisition and pre-processing

The EEG data were originally acquired using a 128-channel (plus two mastoid channels) using an ActiveTwo system (BioSemi) at a rate of 512 Hz, and downsampled to 128 Hz before being distributed as a public dataset. We visually inspected the raw data to identify bad channels, and performed independent component analysis (ICA) to identify and remove blinks; rejected channels were linearly interpolated with nearest neighbour interpolation using MNE-python.

The MEG data were acquired using a 275 axial gradiometer system at 1200 Hz. For the MEG data, preprocessing and source modelling was performed in MATLAB 2018b using fieldtrip (Oostenveld et al., 2011). We applied notch filtering (Butterworh IIR) at the bandwidth of 49–51, 99–101, and 149–151 Hz to remove line noise. Artifacts related to muscle contraction and squidjumps were identified and removed using fieldtrip's semi-automatic rejection procedure. The data were downsampled to 150 Hz. To identify and remove eye blink artifacts, ICA was performed using the FastICA algorithm.

For both MEG and EEG analyses, we focus on the slow, evoked response and hence restricted our analysis to low-frequency components. To this end, we filtered the data between 0.5 and 8 Hz using a bidirectional FIR bandpass filter. Restricting the analysis to such a limited range of low frequencies (which are known to best follow the stimulus) is common when using regression ERP or TRF analysis, especially when the regressors are sparse impulses (Broderick et al., 2018; Di Liberto et al., 2019; Ding and Simon, 2012). The particular upper bound of 8 Hz is arbitrary but was based on earlier papers using the same EEG dataset to study how EEG tracks acoustic and linguistic content of speech (Broderick et al., 2018; Broderick, Anderson, and Lalor, 2019; Heilbron et al., 2019).

Head and source models

The MEG sensors were co-registered to the subjects' anatomical MRIs using position information of three localization coils attached to the headcasts. To create source models, FSL's Brain Extraction Tool was used to strip non-brain tissue. Subject-specific cortical surfaces were reconstructed using Freesurfer, and post-processing (downsampling and surface-based alignment) of the reconstructed cortical surfaces was performed using the Connectome Workbench command-line tools (v 1.1.1). This resulted in cortically-constrained source models with 7,842 source locations per hemisphere. We created single-shell volume conduction models based on the inner surface of the skull to compute the forward projection matrices (leadfields).

Beamformer and parcellation

To estimate the source time series from the MEG data, we used linearly constrained minimum variance (LCMV) beamforming, performed separately for each session, using Fieldtrip's source analysis routine. To reduce the dimensionality, sources were parcellated, based on a refined version of the Conte69 atlas, which is based on Brodmann's areas. We computed, for each session, parcel-based time series by taking the first principal component of the aggregated time series of the dipoles belonging to the same cortical parcel .

Self-attentional language model

Contextual predictions were quantified using GPT-2 – a large, pre-trained language model (Radford et al., 2019). Formally, a language model can be cast as a way of assigning a probability to a sequence of words (or other symbols), $(x_1, x_2, ..., x_n)$. Because of the sequential nature of language, the joint probability, $P(X)$ can, via the chain rule, be factorised as the product of conditional probabilities:

$$\begin{aligned} P(X) &= p(x_1) \times p(x_2 \mid x_1) \times \cdots \times p(x_n \mid x_{n-1}, \ldots, x_1) \\ &= \prod_{i=1}^{x} p\left(x_n \mid x_1, \ldots, x_{n-1}\right) \end{aligned} \tag{4.1}$$

Since the advent of neural language models, as opposed to statistical (Markov) models, methods to compute these conditional probabilities have strongly improved. Improvements have been especially striking in the past two years with the introduction of the *Transformer* (Vaswani et al., 2017) architecture, which allows efficient training of very large networks on large, diverse data. This resulted in models that dramatically improved the state-of-the art in language modelling on a range of domains.

GPT-2 (Radford et al., 2019) is one of these large, transformer-based language models and is currently among the best publcicly released models of English. The architecture of GPT-2 is based on the decoder-only version of the transformer. In a single forward pass, it takes a sequence of tokens $U = (u_1, \ldots, u_k)$ and computes a sequence of conditional probabilities, $(p(u_1), p(u_2|u_1), \ldots, p(u_k \mid u_1, \ldots, u_{k-1}))$. Roughly, the full model (see Figure S4.1) consists of three steps: first, an embedding step encodes the sequence of symbolic tokens as a sequence of vectors which can be seen as the first hidden state h_o. Then, a stack of transformer blocks, repeated n times, each apply a series of operations resulting in a new set of hidden states h_l, for each block l. Finally, a (log-)softmax layer is applied to compute (log-)probabilities over target tokens. Formally, then, the model can be summarised in three equations:

$$h_0 = UW_e + W_p \tag{4.2}$$

$$h_l = \text{transformer\_block}\left(h_{l-1}\right) \forall i \in [1, n] \tag{4.3}$$

$$P(u) = \text{softmax}\left(h_n W_e^T\right), \tag{4.4}$$

where W_e is the token embedding and W_p is the position embedding (see below).

The most important component of the transformer-block is the *masked multi-headed self-attention* (Fig S4.1). The key operation is self-attention, a seq2seq operation turning a sequence of input vectors $(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k)$ into a sequence of output vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k)$. Fundamentally, each output vector $\mathbf{y}_i$ is a weighted average of the input vectors: $\mathbf{y}_i = \sum_{j=1}^{k} w_{ij}\mathbf{x}_j$. Critically, the weight $w_{i,j}$ is not a parameter but is *derived* from a function over input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. The Transformer uses (scaled) *dot product attention*, meaning that the function is simply a dot product between the input vectors $\mathbf{x}_i^T\mathbf{x}_j$, passed through a softmax make sure that the weights sum to one, scaled by a constant determined by the dimensionality, $\frac{1}{\sqrt{d_k}}$ (to avoid the dot-products growing too large in magnitude): $w_{ij} = \left(\exp \mathbf{x}_i^T\mathbf{x}_j / \sum_{j=1}^{k} \exp \mathbf{x}_i^T\mathbf{x}_j\right)\frac{1}{\sqrt{d_k}}$.

In self-attention, then, each input $\mathbf{x}_i$ is used in three ways. First, it is multiplied by the other vectors to derive the weights for its own output, $\mathbf{y}_i$ (as the *query*). Second, it is multiplied by the other vectors to determine the weight for any other output $\mathbf{y}_j$ (as the *key*). Finally, to compute the actual outputs it is used in the weighted sum (as the *value*). Different (learned) linear transformations are applied to the vectors in each of these use cases, resulting in the Query, Key and Value matrices (Q, K, V). Putting this all together, we arrive at the following equation:

$$\text{self\_attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4.5}$$

where d_k is dimension of the keys/queries. In other words, self_attention simply computes a weighted sum of the values, where the weight of each value is determined

by the dot-product similarity of the query with its key. Because the queries, keys and values are linear transformations of the same vectors, the input *attends itself*.

To be used as a language model, two elements need to be added. First, the basic self-attention operation is not sensitive to the order of the vectors: if the order of the input vectors is permuted, the output vectors will be identical (but permuted). To make it position-sensitive, a position embedding W_p is simply added during the embedding step – see Equation 4.2. Second, to enforce that the model only uses information from one direction (i.e left), a mask is applied to the attention weights (before the softmax) which sets all elements above the diagonal to $-\infty$. This makes the self-attention *masked*.

To give the model more flexibility, each transformer block actually contains multiple instances of the basic self-attention mechanisms from (4.5). Each instance (each *head*) applies different linear transformations to turn the same input vectors into a different set of Q, K and V matrices, returning a different set of output vectors. The outputs of all heads are concatenated and then reduced to the initial dimensionality with a linear transformation. This makes the self-attention *multi-headed*.

In total, GPT-2 (XL) contains $n = 48$ blocks, with 12 heads each; a dimensionality of $d = 1600$ and a context window of $k = 1024$, yielding a total 1.5×10^9 parameters. We used the PyTorch implementation of GPT-2 provided by HuggingFace's *Transformers* package (Wolf et al., 2020).

Lexical predictions

We passed the raw texts through GPT-2 (Equations 4.2-4.4) for each run independently (assuming that listeners' expectations would to some extent 'reset' during the break). This resulted in a (log-)probability distribution over tokens $P(U)$. Since GPT-2 uses Byte-Pair Encoding, a token can be either punctuation or a word or (for less frequent words) a word-part. How many words actually fit into a context window of length k therefore depends on the text. For words spanning multiple tokens, we computed word probabilities simply as the joint probability of the tokens. 'For window-placement, we used the constraint that the windows had an overlap of at least 700 tokens, and that they could not start mid-sentence (ensuring that the first sentence of the window was always well-formed).

As such, for each word w_i we computed $p(w_i|\text{context})$, where 'context' consisted either of all preceding words in the run, or of a sequence of prior words constituting a well-formed context that was at least 700 tokens long.

Syntactic and semantic predictions

Feature-specific predictions were computed from the lexical prediction. To this end, we first truncated the unreliable tail from the distribution using a combination of top-k and nucleus truncation. The nucleus was defined as the "top" k tokens with the highest predicted probablility, where k was set dynamically such that the cumulative probability was at least 0.9. To have enough information also for very low entropy cases (where k becomes small), we forced k to be a least 40.

From this truncated distribution, we derived feature-specific predictions by analysing the predicted words. For the syntactic predictions, we performed part of speech tagging on every potential sentence (i.e. the context plus the predicted word) with Spacy to derive the probability distribution over parts-of-speech, from which the syntactic surprise was calculated as the negative log probability of the POS of a word, $-\log(P(\text{POS}_n\,|\text{context}))$.

For the semantic prediction, we took a weighted average of the glove embeddings of the predicted words to compute the expected vector: $\mathbb{E}[\text{G}(w_n)] = \sum_{i=1}^{k} P(x_i)\,\text{G}(x_i)$, where $\text{G}(w_i)$ is the GloVe embedding for predicted word w_i. From this prediction, we computed the semantic prediction error as the cosine distance between the predicted and observed vector:

$$\text{PE}_{\text{semantic}} = 1 - \frac{\mathbb{E}\left[\text{G}\left(w_n\right)\right]\text{G}\left(w_n\right)}{\left\|\mathbb{E}\left[\text{G}\left(w_n\right)\right]\right\|\left\|\text{G}\left(w_n\right)\right\|} \tag{4.6}$$

Phonemic predictions

Phonemic predictions were formalised in the context of incremental word recognition (Brodbeck, Hong, and Simon, 2018; Gwilliams et al., 2018). This process can be cast as probabilistic prediction by assuming that brain is tracking the *cohort* of candidate words consistent with the phonemes so far, each word weighted by its prior probability. We compared two such models that differed only in the prior probability assigned to each word.

The first model was the single-level or frequency-weighted model (Fig 4.6), in which prior probability of words was fixed and defined by a word's overall probability of occurrence (i.e. lexical frequency). The probability of a specific phoneme (A), given the prior phonemes within a word, was then calculated using the statistical definition:

$$P(\varphi_t = A \mid \varphi_{1:t-1}) = \frac{f(C_{\varphi_t = A})}{f(C_{\varphi_{1:t-1}})}. \tag{4.7}$$

Here, $f(C_{\varphi_t = A})$ denotes the cumulative frequency of all words in the remaining cohort of candidate words if the next phoneme were A, and $f(C_{\varphi_{(1:t-1)}})$ denotes the

cumulative frequency of all words in the prior cohort (equivalent to $f(C)$ of all potential continuations). If a certain continuation did not exist and the cohort was empty, $f(C_{\varphi_t=A})$ was assigned a laplacian pseudocount of 1. To efficiently compute (4.7) for every phoneme, we constructed a statistical phonetic dictionary as a digital tree that combined frequency information from SUBTLEX database and pronunciation from the CMU dictionary.

The second model was equivalent to the first model, except that the prior probability of each word was not defined by its overall probability of occurrence, but by its conditional probability in that context (based on GPT-2). This was implemented by constructing a separate phonetic dictionary for every word, in which lexical frequencies were replaced by implied counts derived from the lexical prediction. We truncated the unreliable tail from the distribution and replaced that by a flat tail that assigned each word a pseudocount of 1. This greatly simplifies the problem as it only requires to assign implied counts for the top k predicted words in the dynamic nucleus. Since all counts in the tail are 1, the cumulative implied counts of the nucleus is complementary to the the length of the tail, which is simply the difference between the vocabulary size and nucleus size ($V - k$). As such a little algebra reveals:

$$\text{freqs}_n = P_{tr}(w^{(i)}|\text{context})\frac{V-k}{1-\sum_{j=1}^{k} P(w_j^{(i)}|\text{context})}, \tag{4.8}$$

where $P_{tr}(w^{(i)}|\text{context})$ is the trunctated lexical lexical prediction, and $P(w_j^{(i)}|\text{context})$ is predicted probability that word i in the text is word j in the sorted vocabulary.

Although we computed probabilities using the simple statistical definition of probability, these two ways of assigning lexical frequencies are equivalent to two kinds of priors in a Bayesian model. Specifically, in the first model the prior over words is the fixed unconditional word probability, while in the second model the prior is the contextual probability, itself based on a higher level (lexical) prediction. This makes the second computation *hierarchical* because phoneme predictions are based on not just (at the first level) on short sequences of within-word phonemes, but also on a contextual prior which itself (at the second level) is based on long sequences of prior words.

Non-predictive control variables

To ensure we were probing effects of predictions, we had to control for various non-predictive variables: onsets, acoustics, frequency and semantic congruency. We will briefly outline our definitions of each.

For speech, it is known that the cortical responses are sensitive to fluctuations in the envelope – which is specifically driven by rapid increases of the envelope am-

plitude (or 'acoustic edges') (Daube, Ince, and Gross, 2019). To capture these fluctuations in a sparse, impulse-based regressor we quantified the amplitude of these edges as the variance of the envelope over each event (e.g. phoneme) following (Broderick, Anderson, and Lalor, 2019). A second non-predictive variable is frequency. We accounted for frequency as the overall base rate or unconditional probability of a word, defining it similarly to lexical surprise as the unigrams surprise $-\log P(\text{word})$ based on its frequency of occurrence in subtlex.

The final non-predictive variable was semantic congruency or integration difficulty. This speaks to the debate wether effects of predictability reflect prediction or rather post-hoc effects arising when integrating a word into the semantic context. This can be illustrated by considering a constraining context ('coffee with milk and …'). When we contrast a highly expected word ('sugar') and an unexpected word (e.g. 'dog'), the unexpected word is not just less likely, but also semantically incongruous in the prior context. As such, the increased processing cost reflected by effects like N400 increases might not (only) be due to a violated *prediction* but due to difficulty integrating the target word ('dog') in the semantic context ('coffee with milk') (Brown and Hagoort, 1993; Kutas and Hillyard, 1984; Mantegna et al., 2019; Nieuwland et al., 2020). As a proxy for semantic integration difficulty we computed the semantic congruency of a word in its context defined as the cosine dissimilarity (see (4.6)) between the average semantic vector of the prior context words and the target content word, following (Broderick et al., 2018). This metric is known to predict N400-like modulations and can hence capture the extent to which such effects can be explained by semantic congruency only (Broderick et al., 2018; Nieuwland et al., 2020).

Word-level regression models

The word-level models (see Fig S4.2 for graphical representation) captured neural responses to words as a function of word-level variables. The *baseline* model formalised the hypothesis that responses to words were not affected by word unexpectedness but only by the following non-predictive confounds: word onsets, envelope variability (acoustic edges), semantic congruency (integration difficulty) and word frequency.

The *probabilistic prediction* model formalised the hypothesis that predictions were continuous and probabilistic. This model was identical to the baseline model plus the lexical surprise (or negative log probability of a word), for every word. This was based on normative theories of predictive processing which state that the brain response to a stimulus should be proportional to the negative log probability of that stimulus (Friston, 2005).

The *constrained guessing* model formalised the classical psycholinguistic notion

of prediction as the all-or-none pre-activation of specific words in specific (highly constraining) contexts (Van Petten and Luka, 2012). We translated the idea of all-or-none prediction into a regression model using an insight by Smith and Levy which implied that all-or-none predictions result in a linear relationship between word probability and brain responses (Smith and Levy, 2013). The argument follows from two assumptions: (1) all predictions are all-or-none; and (2) incorrect predictions incur a cost, expressed as a prediction error brain response (fixed in size because of assumption 1). For simplicity, we first consider the unconstrained case (i.e. subjects make a prediction for *every* stimulus), and we bracket all other factors affecting brain responses by absolving them into an average brain response, y_{baseline}. As such, the response to any word is either y_{baseline} (if the prediction is correct) or $y_{\text{baseline}} + y_{\text{error}}$ (if it was false). For any individual stimulus, this equation cannot be used (as we don't know what a subject predicted). But if we assume that predictions are approximately correct, then the probability of a given prediction to be incorrect simply becomes $(1 - p)$. As such, *on average*, the response becomes $y_{\text{resp}} = y_{\text{baseline}} + (1 - p)y_{\text{error}}$. In other words, a linear function of word improbability. To extend this to the constrained case, we only define the improbability regressor for constraining contexts, and add a constant to those events to capture (e.g. suppressive) effects of correct predictions (Figure S4.2). To identify 'constraining contexts', we simply took the 10% of words with the lowest prior lexical entropy. The choice of 10% was arbitrary – however, using a slightly more or less stringent definition would not have changed the results because the naive guessing model (which included linear improbability for *every* word) performed so much better (see Figure S4.5).

Integrated regression model

For all analyses on feature-specific predictions, we formulated an integrated regression model with both word-level and phoneme-level regressors (Figure S4.6). To avoid collinearity between word and phoneme level regressors, phoneme-level regressors were only defined for word-non-initial phonemes, and word-level regressors were define for word-onset. As regressors of interest this model included phonemic surprise, syntactic surprise and semantic prediction error. In principle, we could have also included phoneme and syntactic entropy rather than just surprise (e.g. (Donhauser and Baillet, 2020)) – however, these were highly correlated with the respective surprise. Since this was already a complex regression model, including more correlated regressors would have made the coefficients estimates less reliable and hence more difficult to interpret. As such, we did not include both but focussed on surprise because it has the most direct relation to stimulus evoked effect.

Phoneme-level regression models

To compare different accounts of phoneme prediction, we formulated three regression models with only regressors at the individual phoneme level (Figure S4.15). In all models, following (Brodbeck, Hong, and Simon, 2018) we used separate regressors for word-initial and word-non-initial phonemes, to account for juncture phonemes being processed differently. The baseline model only included non-predictive factors of word-boundaries, phoneme onsets, envelope variability, and uniqueness points. The two additional models also included phoneme surprise and phoneme entropy from either the hierarchical model or non-hierarchical model. To maximise our ability to dissociate the hierarchical prediction and non-hierarchical prediction, we included both entropy and surprise. Although these metrics are correlated, adding both should add more information to the model-comparison, assuming that there is some effect of entropy (Donhauser and Baillet, 2020). (Note that here, we were only interested in model comparison, and not in comparing the coefficients, which may become more difficult when including both.)

Time resolved regression

As we were interested in the evoked responses, variables were regressed against EEG data using time-resolved regression, within a regression ERP/F (or impulse TRF) framework (Broderick et al., 2018; Smith and Kutas, 2015). Briefly, this involves using impulse regressors for both constants and covariates defined at event onsets, and then temporally expanding the design matrix such that each predictor column C becomes a series of columns over a range of temporal lags $C_{t_{min}}^{t_{max}} = (C_{t_{min}}, ..., C_{t_{max}})$. For each predictor one thus estimates a series of weights $\beta_{t_{min}}^{t_{max}}$ (Fig 4.1) which can be understood as the *modulation function* describing how a given regressor modulates the neural response over time, and which corresponds to the *effective* evoked response that would have been obtained in a time-locked ERP/ERF design. Here, we used a range between -0.2 and 1.2 seconds. All data and regressors were standardised and coefficients were estimated with ℓ_2-norm regularised (Ridge) regression:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{4.9}$$

using the scikit learn sparse matrix implementation. In both datasets, models were estimated by concatenating the (time-expanded) design matrix across all runs and sessions. Regularisation was set based on leave-one-run-out R^2 comparison; for inference on the weights in the EEG data this was done across subjects to avoid doing statistics over coefficients with different amounts of shrinkage.

Model comparison

In both datasets, model comparison was based on comparing cross-validated correlation coefficients. Cross-validation was performed in a leave-one-run-out cross-validation scheme, amounting to 19-fold cross-validation in the EEG data and between 63 and 65-fold cross-validation for the MEG data (in some subjects, some runs were discarded due to technical problems).

For the EEG data, models' cross-validated prediction performance was performed across subjects to perform population-level inference. To this end, we reduced the scores into a single n_{subs} dimensional vector by taking the median across folds and the mean across channels. Critically, we did not select any channels but used the average across the scalp. For the MEG data, models were only statistically compared on a within within-subject basis. Because the MEG data was source localised we could discard sources known to be of no interest (e.g. early visual cortex). To this end, we focussed on the language network, using a rather unconstrained definition encompassing all Brodmann areas in the temporal lobe, plus the temporo-parietal junction, and inferior frontal gyrus and dorsolateral prefrontal cortex; all bilaterally (see Figure S4.16).

Statistical testing

All statistical tests were two-tailed and used an alpha of 0.05. For all simple univariate tests performed to compare model-performance within and between subjects, we first verified that the distribution of the data did not violate normality and was outlier free, determined by the D'Agostino and Pearson's test implemented in SciPy and the 1.5 IQR criterion, respectively. If both criteria were met, we used a parametric test (e.g. paired t-test); otherwise, we resorted to a non-parametric alternative (e.g. Wilcoxon sign rank).

In EEG, we performed mass-univariate tests on the coefficients across participants between 0 and 1.2 seconds. This was firstly done using cluster-based permutation tests (Gramfort et al., 2014; Maris and Oostenveld, 2007) to identify clustered significant effects as in Figure 4.5 (10,000 permutations per test). Because the clustered effects as in Figure 4.5 only provide a partial view, we also reported more comprehensive picture of the coefficients across all channels (Figure S4.3,S4.8); there, we also provide multiple-comparison corrected p-values to indicate statistical consistency of the effects; these were computed using TFCE. In the MEG, multiple comparison correction for comparison of explained variance across cortical areas was done using Treshold Free Cluster Enhancement (TFCE). In both datasets, mass-univariate testing was performed based on one-sample t-tests plus the 'hat' variance adjustment method with $\sigma = 10^{-3}$.

Polarity-alignment

In the source localised MEG data, the coefficients in individuals (e.g. Figure S4.11-S4.14) are symmetric in polarity, with the different sources in a single response having an arbitrary sign due to ambiguity of the source polarity. To harmonise the polarities, and avoid cancellation when visualising the average coefficient, we performed a polarity-alignment procedure. This was based on first performing SVD, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where $\mathbf{A}$ is the $m \times n$ coefficient matrix, with m being the number of sources and n the number of regressors; and then multiplying each row of $\mathbf{A}$ by the sign of the first right singular vector. Because the right singular vectors (columns of $\mathbf{U}$) can be interpreted as the eigen vectors of the source-by-source correlation matrix, this can be thought of as flipping the sign of each source as a function of its polarity with respect to the dominant correlation. This procedure was used for visualisation purposes only (see Fig S4.4 and S4.11-S4.14).

Data and code availability

Data and code to reproduce all results will be made public at the Donders Repository. The full MEG dataset will be made public in a separate resource publication.

Acknowledgements

This work was supported by The Netherlands Organisation for Scientific Research (NWO Research Talent grant to M.H.; NWO Vidi grant to F.P.d.L.; NWO Vidi 864.14.011 to JMS; Gravitation Program Grant Language in Interaction no. 024.001.006 to P.H.) and the European Union Horizon 2020 Program (ERC Starting Grant 678286, 'Contextvision' to F.P.d.L). We wish to thank Michael P Broderick, Giovanni M. Di Liberto, and colleagues from the Lalor lab for making the EEG dataset openly available. We thank all the authors of the open source software we used and apologise for citation limits that prevent us from citing all tools used.

Contributions

Conceptualisation: MH, FPdL, PH; Formal analysis: MH; Data collection: KA, JMS; Source modelling: KA, JMS; Original draft: MH; Final manuscript: MH, FPdL, PH, JMS,KA.

Supplementary Figures

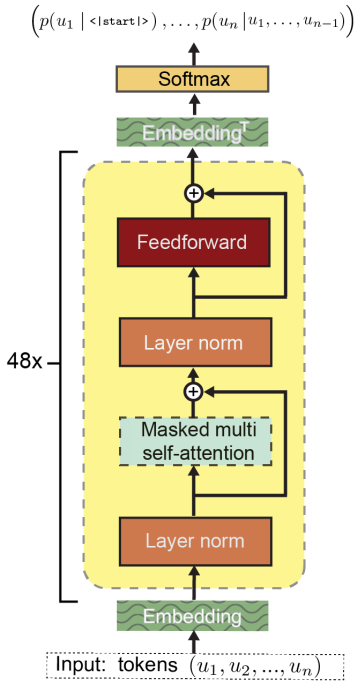

Figure S4.1. GPT-2 Architecture. Note that this panel is a re-rendered version of the original GPT schematic, slightly modifyied and re-arranged to match the architecture of GPT-2. For more details on the overall architecture and on the critical operation of self-attention, see *Methods*. In this graphic, Layer Norm refers to layer normalisation as described by Ba et al. Not visualised here is the initial tokenisation, mapping a sequence of characters into tokens.

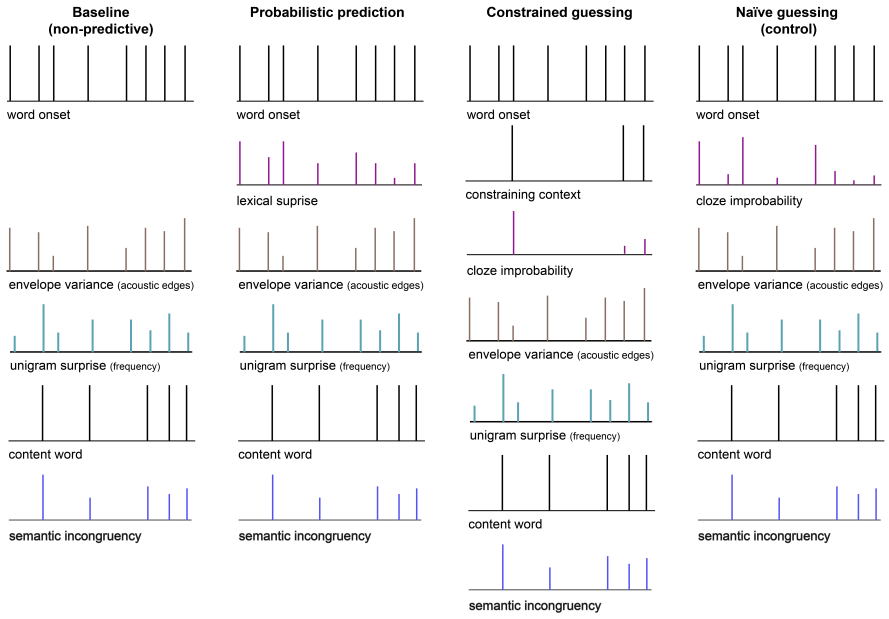

Figure S4.2. Word-level regression models. Schematic of the main models plus the control model of the initial model comparison to test for predictive processing at the word level. Because we use a regression ERP/ERF scheme (Smith and Kutas, 2015), aimed at capturing (modulations of) the evoked response to discrete events like words or phonemes, all regressors are modelled as impulses (see *Methods*).

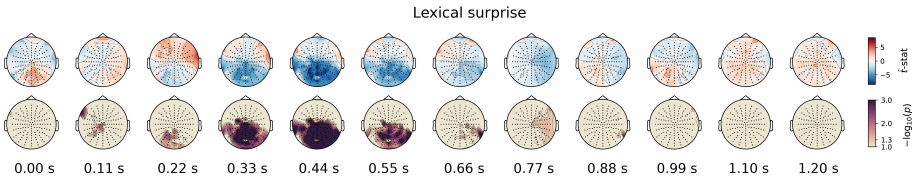

Figure S4.3. Full EEG topographies of the effects of lexical surprise These topographies show the average t-statistics of the coefficients (upper row) and respective FWE-corrected significance (lower row) of the lexical surprise regressor from the *probabilistic prediction* model (Figure S4.2). As such, while Figure 4.2b shows the coefficients averaged over channels participating in the cluster (thereby only visualising *the effect*) these topographies visualise the results comprehensively across all channels over time.

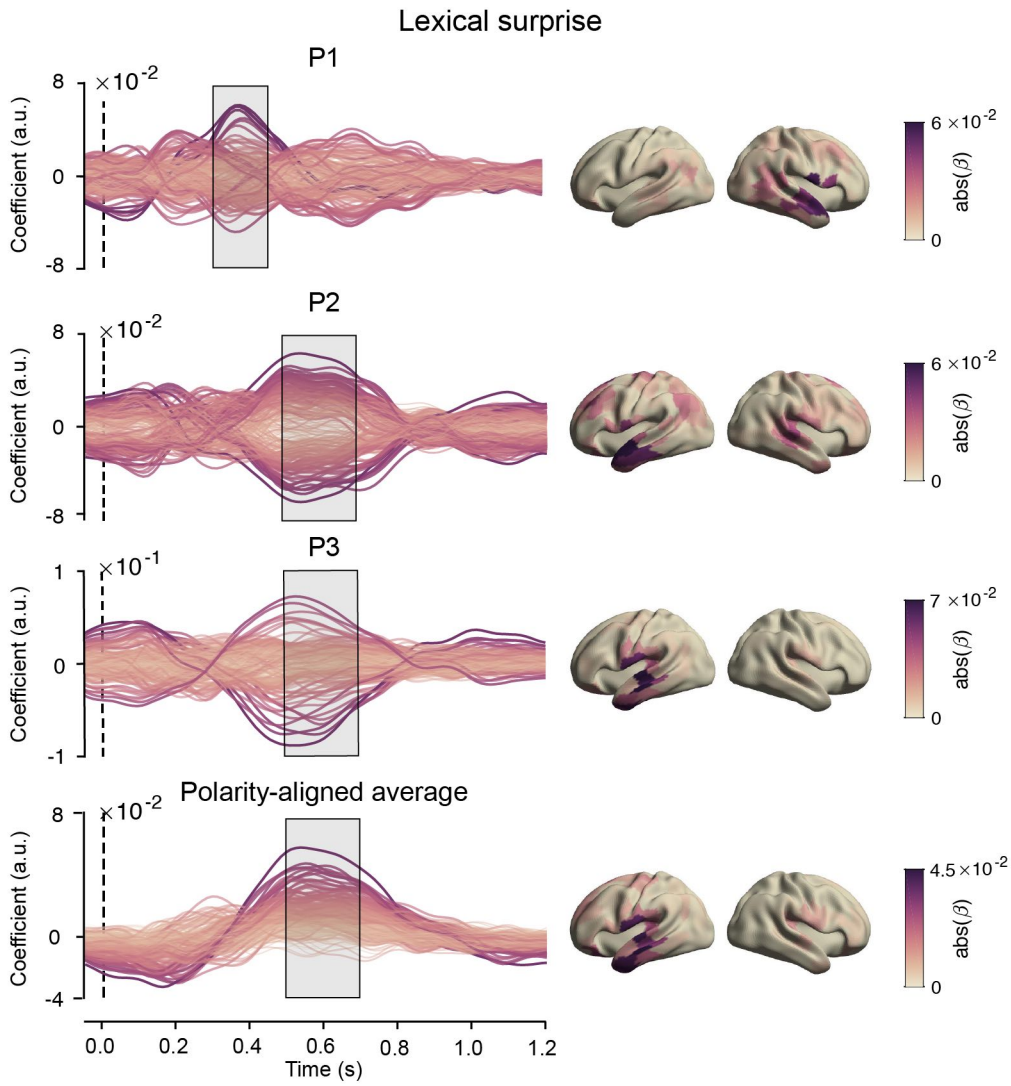

Figure S4.4. Coefficients for lexical surprise from the lexical model (Figure S4.2) Left column: timecourses of the coefficients at each MEG source-localised parcel for lexical surprise for all MEG participants, and the polarity-aligned average across them. Right column: Absolute value of the coefficients averaged across the highlighted period plotted across the brain.

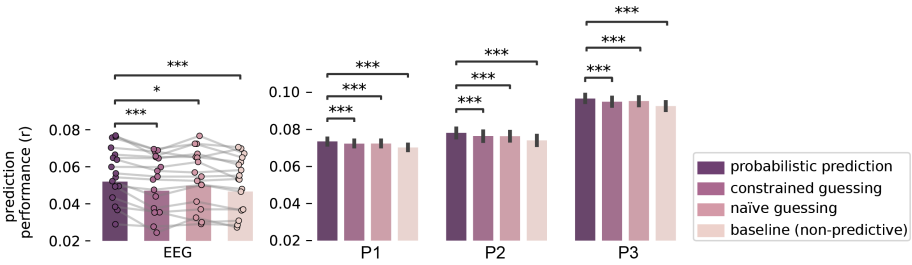

Figure S4.5. Model comparison results across all channels (EEG) and the full language network (MEG). Same as in Figure 4.2a, but now including the 'naive guessing' control model. Like the constrained guessing model, this model included a linear estimate of word probability, but defined for every word rather than only for constraining contexts. This model was introduced to identify which of the two differences between the *probabilistic prediction* and *constrained guessing* model – i.e. assuming that predictions are (i) categorical vs. probabilistic and (ii) occasional vs. continuous – made the largest difference in model performance. As can be seen, the *naive guessing* model performed considerably better than the *constrained guessing* model, but consistently worse than the *probabilistic prediction* model. This clearly shows that the modulatory effect of unexpectedness is not limited to only highly constraining contexts, but that that it applies much more generally – in line with the notion of continuous prediction.

Strictly speaking, the naive guessing model formalises the hypothesis that the brain 'naively' makes *all-or-none* guesses about *every* upcoming word. Given that this hypothesis is a-priori so implausible, it may seem surprising that the model still performs comparably well. However, we should note that the probabilistic prediction regressor (*surprise*) and the categorical prediction regressor (linear (im)probability) are highly correlated (0.7) because one is a monotonic function of the other. Therefore, we suggest the results are better interpreted the other way around: the fact that – despite being so correlated – the log-probability is consistently a better linear predictor of neural responses than the linear probability clearly supports predictive processing theories, which postulate that the neural response to a stimulus should be proportional to negative log-probability of that stimulus.

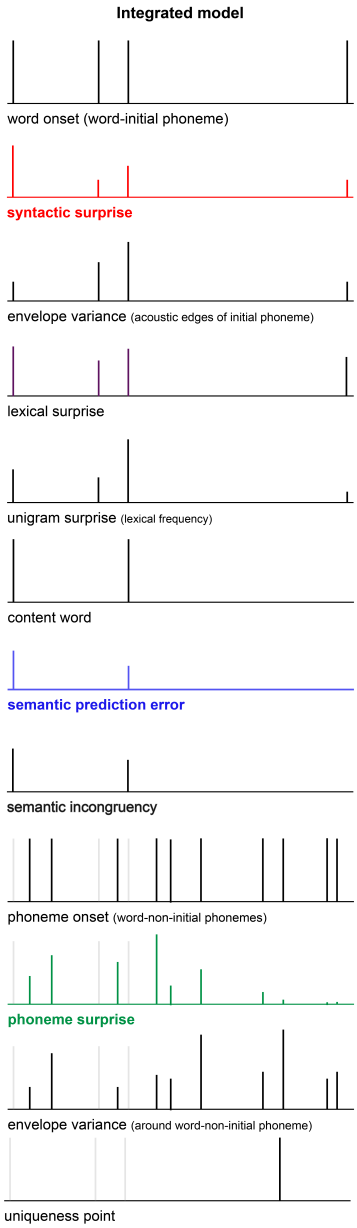

Figure S4.6. Regressors of the integrated feature-specific model. Same as Figure S4.5, but for the integrated feature-specific regression model. The three regressors of interest – syntactic surprise, semantic prediction error and phonemic surprise – are coloured, all control regressors are in black. Following the regression ERP/ERF scheme (Smith and Kutas, 2015), aimed at capturing (modulations of) the evoked response to discrete events like words or phonemes, all regressors are modelled as impulses (see *Methods*). To avoid collinearity between word an and phoneme regressors, phoneme regressors (both events and covariates) are restricted to all non-initial phonemes.

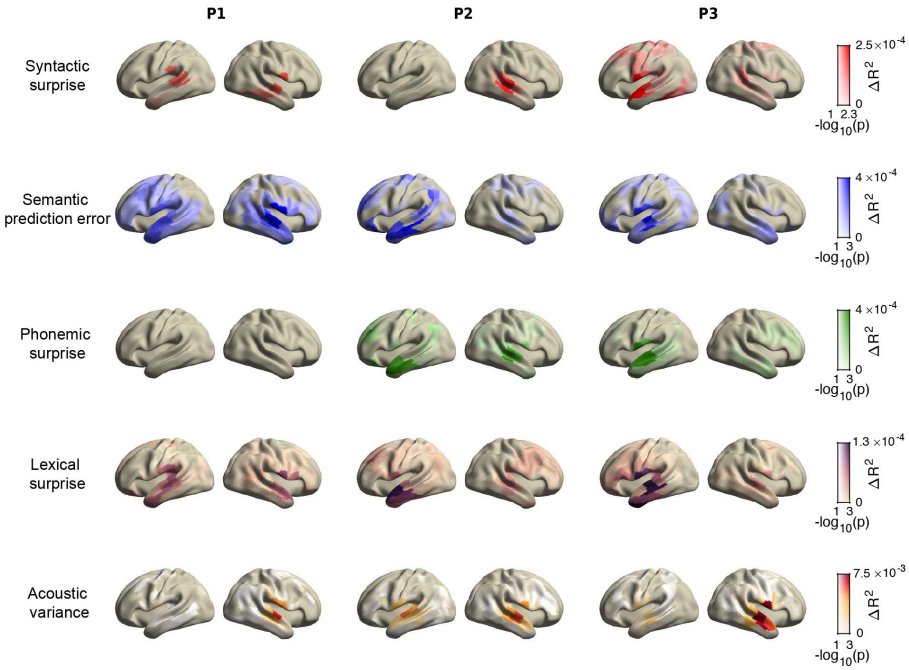

Figure S4.7. Unique explained variance for five regressors across the brain. Same as Figure 4.4, but including 2 control regressors (lexical surprise and acoustic variance) for comparison. Colours indicate amount of additional variance explained by each regressor; opacity indicates the FWE-corrected statitsical significance (across cross-validation folds). Note that $p < 0.05$ is equivalent to $-\log_{10}(p) > 1.3$.

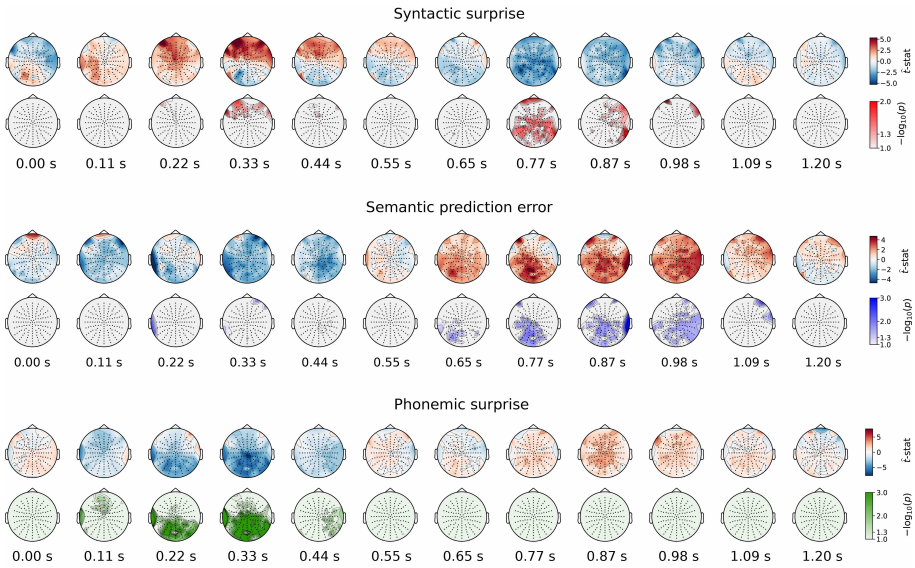

Figure S4.8. Full topographies of the coefficients and significance of feature-specific prediction errors For each feature-specific prediction error regressor, the topographies show the t-statistics of the coefficients (upper row) and the respective TFCE-corrected significance (lower row). So while Figure 4.5 only shows the coefficients averaged over channels participating in the cluster (thereby only visualising *the effect*) these topographies visualise the results comprehensively across all channels, over time.

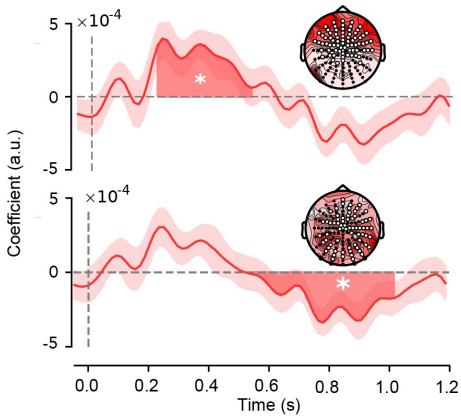

Figure S4.9. Significant effects of syntactic surprise in the EEG data. Two significant effects were observed in the modulation functions for syntactic surprise: an early positive effect with a frontal topography (upper panel) and a later negative effect based on a distributed cluster (lower panel). The early effect tightly replicates recent model-based studies on EEG effects of syntactic surprise, and was also found in the MEG data. By contrast, the late effect of syntactic surprise is not in line with any earlier study (note that it is negative unlike the syntactic P600) and importantly was not replicated in the MEG data. Therefore we only consider the early effect a 'main' effect of syntactic surprise (visualised in the main Figure 4.5) and we advice to refrain from interpreting the late effect before it is independently replicated.

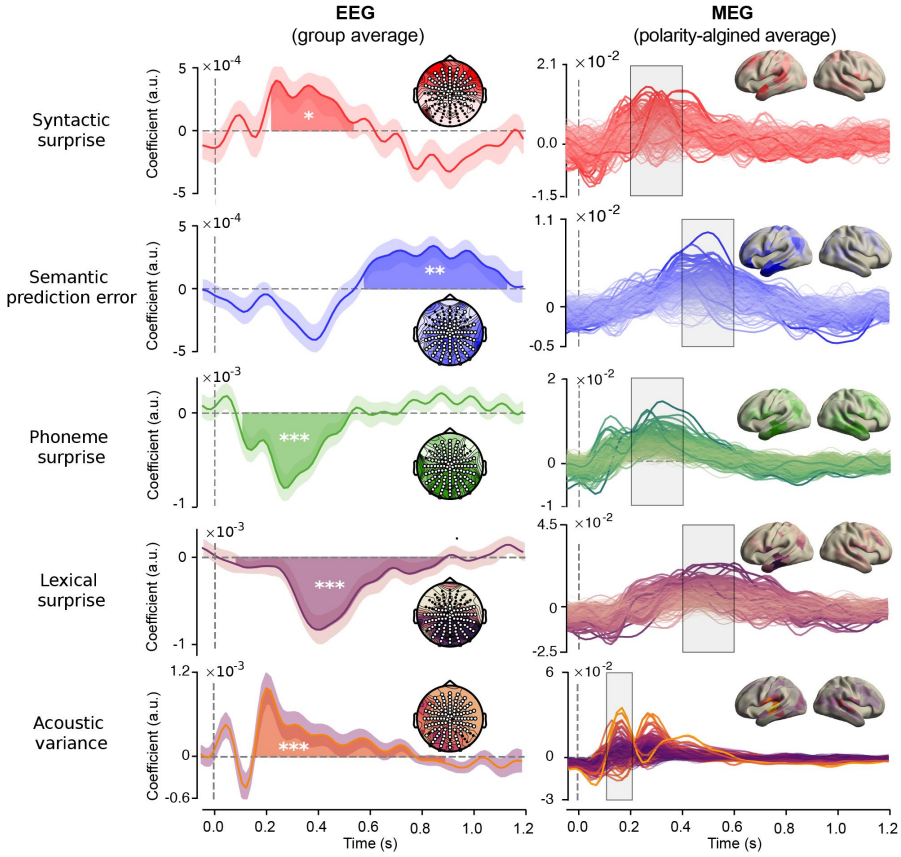

Figure S4.10. Coefficients for each prediction error, plus two control variables. EEG (left column): coefficient modulation function averaged across the channels participating for at least one sample in the significant clusters. Highlighted area indicates temporal extent of the cluster. Shaded area around waveform indicates bootstrapped standard errors. Stars indicate cluster-level significance; $p < 0.05$ (*), $p < 0.05$ (**), $p < 0.001$ (***). Insets represent channels assigned to the cluster (white dots) and the distribution of absolute values of t-statistics. MEG (right column): polarity aligned responses averaged across participants for all sources (same as in Figure 4.5 but without averaging over sources, and including two control variables). Insets represent topography of absolute value of coefficients averaged across the highlighted period. Note that due to polarity alignment, sign information is to be ignored for the MEG plots.

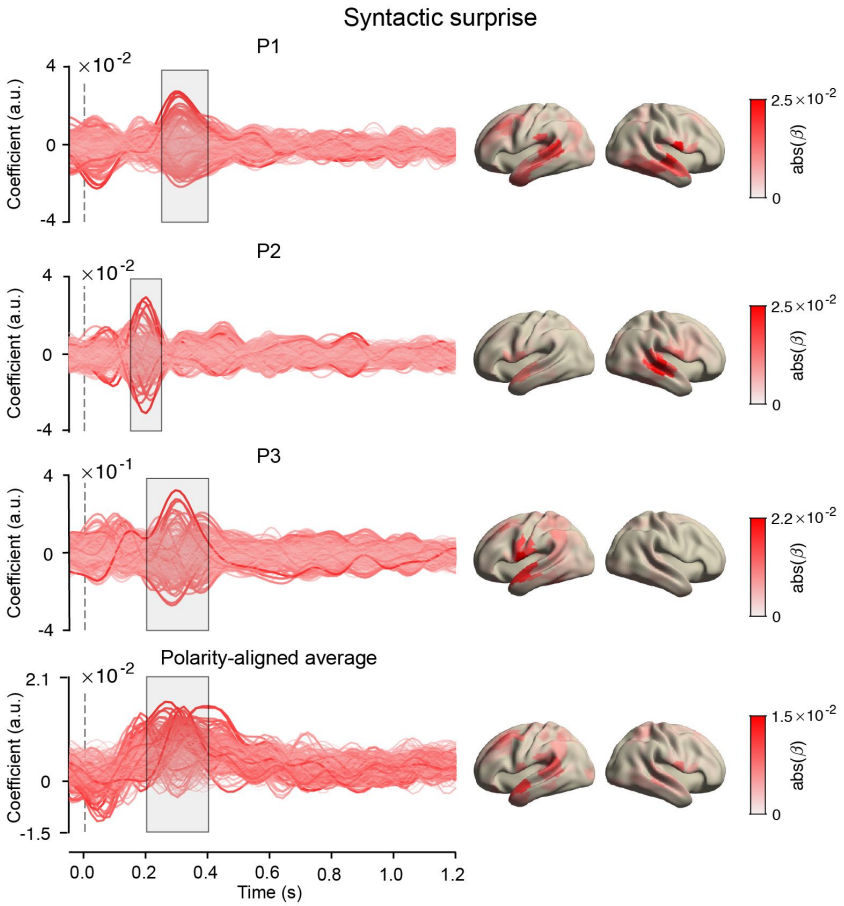

Figure S4.11. Coefficients for syntactic surprise from the integrated model (Figure S4.6) Left column: coefficients for each source for each individual in the MEG experiment, and the polarity-aligned average across participants. Right column: absolute value of the coefficients across the brain, averaged across the highlighted time-period.

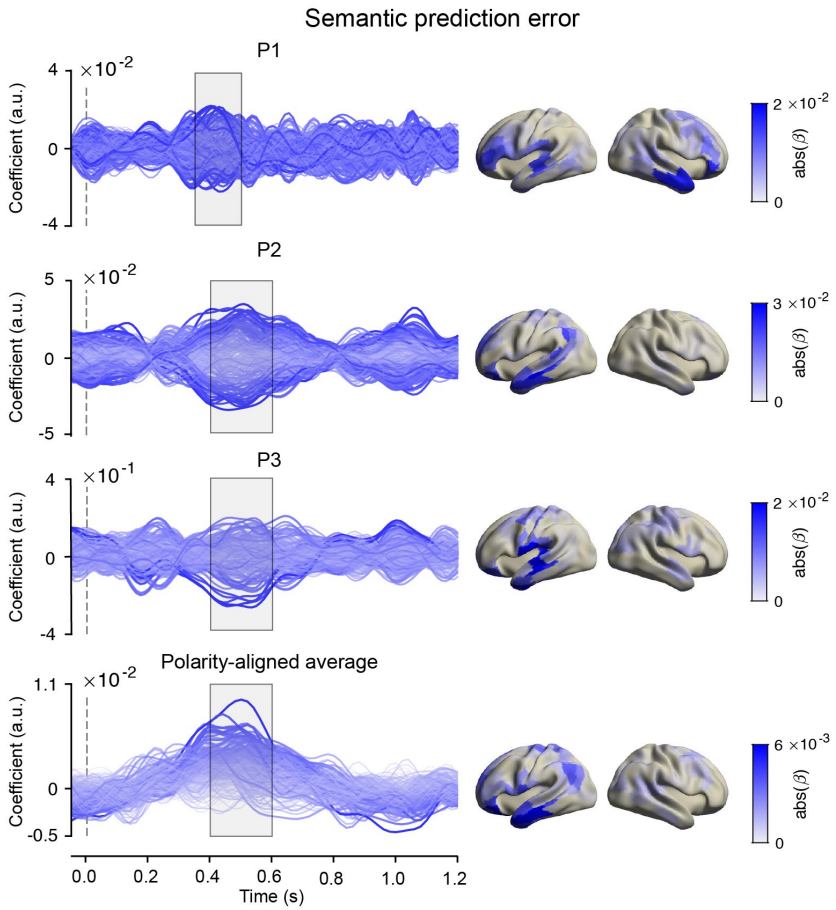

Figure S4.12. Coefficients for semantic prediction error from the integrated model (Figure S4.6) Left column: coefficients for each source for each individual in the MEG experiment, and the polarity-aligned average across participants. Right column: absolute value of the coefficients across the brain, averaged across the highlighted time-period.

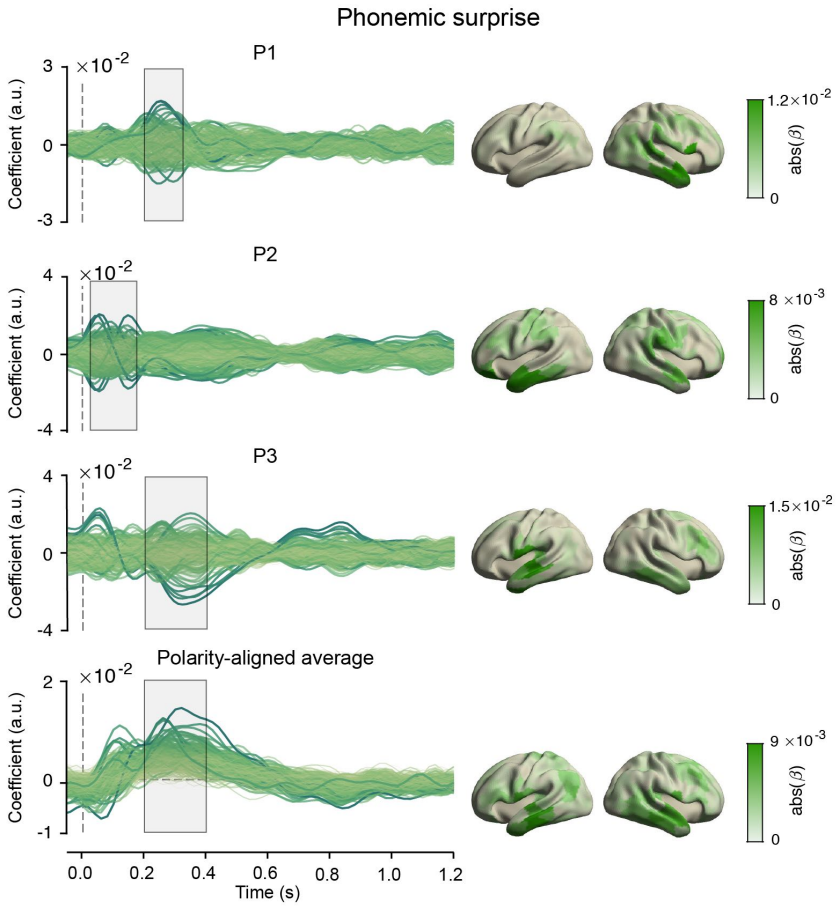

Figure S4.13. Coefficients for phonemic surprise from the integrated model (Figure S4.6) Left column: coefficients for each source for each individual in the MEG experiment, and the polarity-aligned average across participants. Right column: absolute value of the coefficients across the brain, averaged across the highlighted time-period..

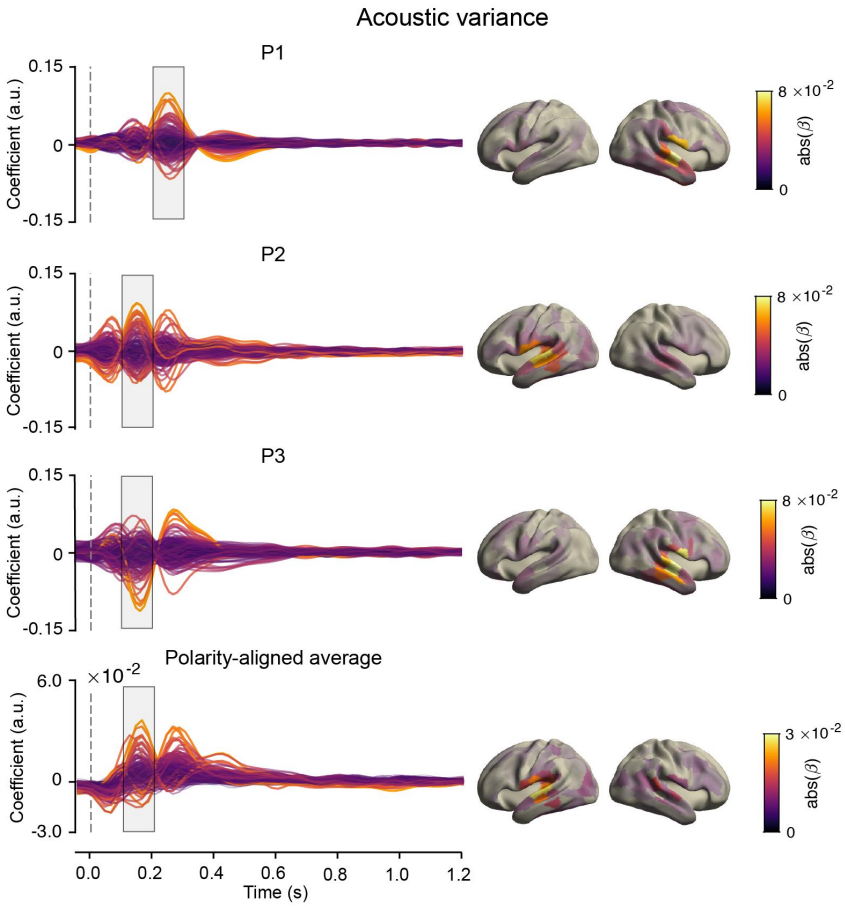

Figure S4.14. Coefficients for envelope variability from the integrated model (Figure S4.6) Left column: coefficients for each source for each individual in the MEG experiment, and the polarity-aligned average across participants. Right column: absolute value of the coefficients across the brain, averaged across the highlighted time-period.

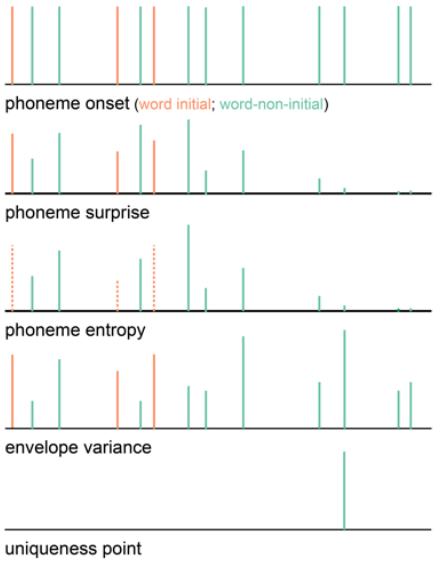

Figure S4.15. Regressors of the phoneme model. As indicated by the different colours, both the constants and covariates were modelled separately for word-initial and word-non-initial phonemes.

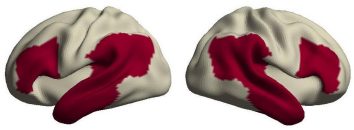

Figure S4.16. Language network definition The language network was defined as temporal cortex plus temporo-parietal junction, and IFG and dorsolateral prefrontal cortex; all bilaterally. In terms of Brodmann areas this corresponded to 20, 21, 22, 38, 39, 40, 41, 42, 44, 45, 46 and 47, amounting to a total of 100 out of 370 cortical parcels.

Chapter 5

Prior uncertainty modulates beta-band activity during the perception of natural speech

Abstract

Models of predictive processing posit that the brain constantly relies on top-down predictions, and that these top-down predictions are signalled via specific frequency bands – in particular the beta (12-28 Hz) and alpha (8-12 Hz) band. Several studies have indeed reported such oscillatory signatures of top-down prediction. However, most of these studies used simple designs with extremely strong regularities, leaving open whether these oscillatory top-down effects are as ubiquitous as predictive processing theory implies. Here, we address this question by testing for oscillatory signatures of predictive processing during naturalistic speech perception, quantifying prediction and surprise on a phoneme-by-phoneme basis. Results revealed that phoneme surprise modulated the theta and delta band amplitude in the temporal lobe, while prior uncertainty about the incoming word modulated activity in the beta band in frontotemporal areas. Investigations of the coefficients tentatively suggest that uncertainty about the incoming word *increases* pre-stimulus beta amplitude. This is in line with prior literature on language processing, but opposite to what we expected based on predictive processing accounts of neural oscillations. However, it is not necessarily inconsistent with predictive processing. Together, these results show that in naturalistic speech perception, prediction confidence about incoming stimuli modulates the ongoing beta amplitude. However, they also highlight how deriving testable hypotheses about the relation between predictive processing and the modulation of specific frequency bands can be less straightforward than it may seem. Methodological opportunities to strengthen the conclusions are discussed.

This chapter is based on:
Heilbron M., Westner, B., Hagoort P, de Lange FP. (2021). Prior uncertainty modulates beta-band activity during the perception of natural speech. (In preparation.)

Introduction

In everyday situations, the brain is confronted with a dizzyingly complex and ambiguous stream of sensory information. And yet tasks like object recognition or speech perception are generally handled with great efficiency, and seemingly without effort.

Theories of predictive processing propose that the brain achieves this feat by relying on internal generative models of the world (Clark, 2013; Friston, 2005; Rao and Ballard, 1999). From these models, the brain generates top-down predictions of bottom-up sensory input which guide the processing of the incoming sensory stream. There is a large and growing body of evidence supporting predictive processing, primarily by demonstrating how predictions enhance perception and modulate brain responses (see de Lange, Heilbron, and Kok, 2018; Keller and Mrsic-Flogel, 2018 for review).

However, one key tenet of predictive processing remains less studied and is not as well-supported: the existence of distinct spectral profiles of prediction and prediction error signals. While not part of original formulations of predictive processing (Friston, 2005; Mumford, 1992; Rao and Ballard, 1999), this assumption was motivated by studies demonstrating distinct frequency channels for bottom-up signals (associated with the gamma band) and top-down signals (using beta and alpha bands Buschman and Miller, 2007; Wang, 2010). Because classic predictive coding theory postulates an asymmetry between top-downs signals (carrying predictions) and bottom-up signals (carrying prediction errors), these two variables should have a distinct spectral profile: predictions should be associated with lower frequency beta (12-28 Hz) and/or alpha (8-12 Hz) bands, while prediction errors should be associated with the higher frequency gamma band (> 30 Hz; Arnal and Giraud, 2012; Bastos et al., 2012).

Beyond the potential of gaining more mechanistic insight (by dissociating top-down and bottom-up signalling), another opportunity of studying the spectral characteristics of predictive processing is that it may facilitate probing of pre-stimulus predictive activity. Such activity may be roughly time-locked but not phase-locked to the onset of the next stimulus and hence be measurable through induced but not evoked response analysis (Siegel, Donner, and Engel, 2012). Indeed, sensory predictions have been linked to pre-stimulus beta and alpha activity (Mayer et al., 2016; Meyniel, 2020; Spitzer and Haegens, 2017).

An important but often overlooked consequence of the predictive processing framework for understanding oscillations, is that beta/alpha band signatures of top-down processing should be ubiquitous. In other words, they should not be limited to overt cases of top-down processing requiring active task engagement, like the working memory and attention tasks in which these signatures have been demonstrated

most clearly (Bastos et al., 2015; Buschman and Miller, 2007; Engel and Fries, 2010; Kerkoerle et al., 2014; Michalareas et al., 2016). After all, predictive processing claims that *all* neural processing strongly relies on top-down predictions. And yet, most studies reporting spectral signatures of top-down prediction used task-induced predictions from extremely simple, highly predictable regularities, often with extended pre-stimulus null-periods – thereby mostly focussing on the special case of explicit, conscious prediction (e.g. Bastos et al., 2020; Chao et al., 2018; Ede, Jensen, and Maris, 2010; Fujioka et al., 2009; Meyniel, 2020; Sedley et al., 2016). This leaves open whether these effects are as ubiquitous as predictive processing implies.

Here, we test this implication in the context of natural language understanding. Language provides a powerful testbed because it is governed by complex and yet relatively transparent regularities. This allows one to study predictions without having to induce artificial (typically extremely simple) regularities and without an extraneous task. Instead, one can simply probe the implicit linguistic expectations that arise naturally when understanding language. Building on earlier work (Armeni et al., 2019; Donhauser and Baillet, 2020) we study oscillations during passive audiobook listening and use computational modelling to quantify linguistic expectations on a moment-by-moment basis.

We focus on two datasets of human electrophysiological recordings of participants simply listening to long segments of natural speech, without distinct pre- and post-stimulus periods and without an online task. As such, the recordings constitute a strong test for the ubiquity of the oscillatory signatures of top-down prediction. We were primarily interested in the beta/alpha bands (expecting a positive relationship with prediction confidence, as proposed by predictive processing accounts Arnal and Giraud, 2012; Bastos et al., 2012; Lewis and Bastiaansen, 2015; Lewis, Wang, and Bastiaansen, 2015) and also in the gamma band (expecting a positive relation with unexpectedness).

To foreshadow the results, we found modulations by prior prediction confidence and unexpectedness in all bands, except the gamma band. As suggested by predictive processing accounts, prior prediction confidence appeared to specifically modulate pre-stimulus beta. Unexpectedly, pre-stimulus beta was *higher* when prior predictions were more *uncertain* – a result that was opposite to our what we hypothesised, but that, on reflection is not necessarily inconsistent with predictive processing. Together, these results demonstrate the feasibility of studying the oscillatory signatures of predictive processing in naturalistic conditions. However, they also show how using predictive processing to derive testable hypotheses about specific frequency bands can sometimes be less straightforward than it may seem.

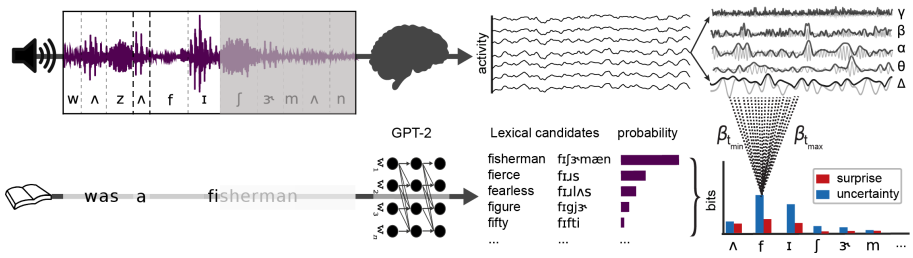

Figure 5.1. Experimental and analytical framework. *Top row:* participants listened to long, continuous segments from audiobooks, extensively annotated to yield onset and offset times for each word and phoneme (light/thick dashed lines indicate phoneme/word boundaries). Brain activity was recorded using either EEG (N=19, 1 hour per participant) or (source-localised) MEG (N=3, 9 hours per participant). Recordings were decomposed using estimates of instantaneous amplitude at different frequency bands – for the MEG data, this was part of the source localisation, in a hilbert-beamformer procedure (see Methods). *Bottom row:* the speech materials were analysed with GPT-2 to generate word-by-word contextual predictions, which were used to calibrate phoneme-level predictions about the incoming word. For example, in the illustration the incoming word is 'fisherman'. A prediction is computed from the 'cohort' of words consistent with the phonemes so far (**f,I**) *and* the contextual probability of each lexical candidate, derived from GPT-2. In general, uncertainty and surprise tend to be higher at the first phonemes and decrease gradually over the course of the word – but this pattern depends on constraint, and may be different when a word is highly (un)expected in context. Uncertainty and surprise were regressed against frequency amplitude using a time-resolved regression.

Results

We analysed continuous electrophysiological recordings from two independent naturalistic experiments in which participants listened to natural, narrative speech from audiobooks, both of which have been analysed before (Broderick et al., 2018; Di Liberto, O'Sullivan, and Lalor, 2015; Heilbron et al., 2021a). The first experiment collected electroencaphaloghaphic (EEG) recordings of 19 participants (1 hour per participant). The second experiment collected magnetoencephalographic (MEG) data; this dataset comprises three participants, who each participated in 10 sessions (1 hour each), wearing individualised head casts to minimise motion so as to allow high-precision localisation of neural activity. We decomposed the recordings into distinct frequency bands, yielding time-resolved instantaneous amplitude (square root of power) in each band of interest (Figure 5.1).

To quantify linguistic predictions, we used a deep neural language model (GPT-2) to estimate, for each word in the stimulus material, a probabilistic prediction about its identity given the preceding words. Because auditory word recognition is incremental (Marslen-Wilson, 1987; McClelland and Elman, 1986) and we were interested in the precise moment (before and after) recognition, we formalised prediction about

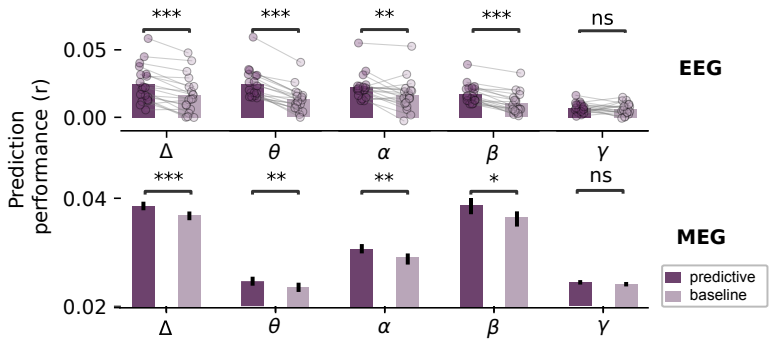

Figure 5.2. Predictions modulate brain activity in the delta trhough beta band. Cross-validated correlation coefficients for EEG (top row) and MEG (bottom-row) of the predictive processing and baseline model. Top row: bars represent mean across participants, dots with connected lines represent individual participants. Bottom row: bars represent grand mean across all cross-validation folds (pooled across participants); error-bars represent the within-fold 95% confidence interval, computed using multi-level non-parametric statistics (hierarchical bootstrap). Significance levels correspond to $p < 0.05(*)$,$p < 0.01(**)$, $p < 0.001(***)$, computed using a bootstrap t-test across participant means (EEG) or a multi-level bootstrap across participants and folds (MEG).

the incoming words on a phoneme-by-phoneme basis (see also Brodbeck, Hong, and Simon, 2018; Donhauser and Baillet, 2020; Gwilliams et al., 2018; Heilbron et al., 2021a). To this end we used the predictions from GPT-2 to calibrate the phonemic probabilistic computations in order to incorporate long-distance linguistic context into the phoneme-by-phoneme predictions (see 5.1 and *Methods*). We extracted two key metrics of interest: the prior uncertainty about the identity of the incoming word (quantified by phoneme-by-phoneme lexical entropy), and the unexpectedness of – or 'surprise' about – each phoneme (quantified by phoneme surprisal). Here, we use uncertainty as metric of (inverse) prior prediction confidence (*before* recognition) and surprisal as an index of the unexpectedness or prediction error (*after* recognition).

With these metrics, we used a regression-based deconvolution approach to estimate the effects of prediction (un)certainty and surprise on the band-limited amplitude in a time-resolved fashion (Fig. 5.1).

Linguistic predictions modulate responses in the delta through beta band

Before testing for specific oscillatory signatures, we first wanted to verify, for each band, that it was sensitive to linguistic predictability in general. To this end, we compared two regression models of the induced responses to the speech material. First, a baseline model which included no prediction-related metrics but only potentially confounding variables: namely the onset of each phoneme, the acoustic (speech en-

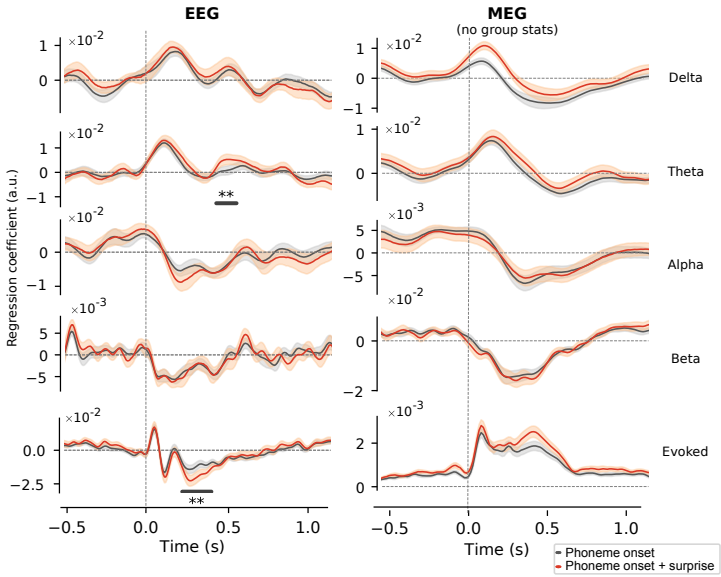

Figure 5.3. Phoneme surprise is associated with increased theta and delta activity Coefficients describing the effects of the unexpectedness of the speech material (phonemic surprisal) on the neural response. Here, the phoneme onset coefficients (black) capture the average response to a phoneme. The surprisal coefficients (red) represent the average to a phoneme, plus the average modulation by surprisal – which is equivalent to the response to a phoneme with a surprisal of 1 SD (note that due to regularisation the coefficients cannot be numerically interpreted as β weights). In the EEG figures, thick lines indicate the mean coefficient in the selected channels, shaded areas represent bootstrapped standard error, both across participants. In the EEG plots, stars indicate significance level using a cluster-based permutation t-test: $p < 0.01(**)$ across the participants. In the MEG plots, coefficients were estimated for each session independently (10 sessions per participant); mean and standard error reflect the average across all sessions. In the MEG data, no population-level inferential statistics were performed, due to the low number of participants. For the evoked coefficients, we took the absolute value before averaging across sources and sessions, to avoid sources with opposite polarity cancelling out.

velope) energy of each phoneme, and the word and sentence boundaries. Second, we considered a 'predictive processing' model, which included the same baseline variables, plus surprise and (un)certainty on a phoneme-by-phoneme basis.

We then evaluated the ability of both models to predict the amplitude in each band, in a cross-validated fashion. This revealed that the predictive processing model performed better than the baseline model for all bands, except for the gamma band. This was the case in both the EEG data (bootstrap t-test; Delta: $p < 10^{-4}$; theta: $p < 10^{-4}$ alpha: $p = 4.4 \times 10^{-3}$; beta: $p < 10^{-4}$ gamma: $p = 0.25$) and in the MEG data (pooling across participants with hierarchical bootstrap t-test; delta: $p < 10^{-4}$ theta: $p = 2.2 \times 10^{-3}$; alpha: $p = 2.7 \times 10^{-3}$; beta: $p = 0.024$; gamma: $p = 0.052$).

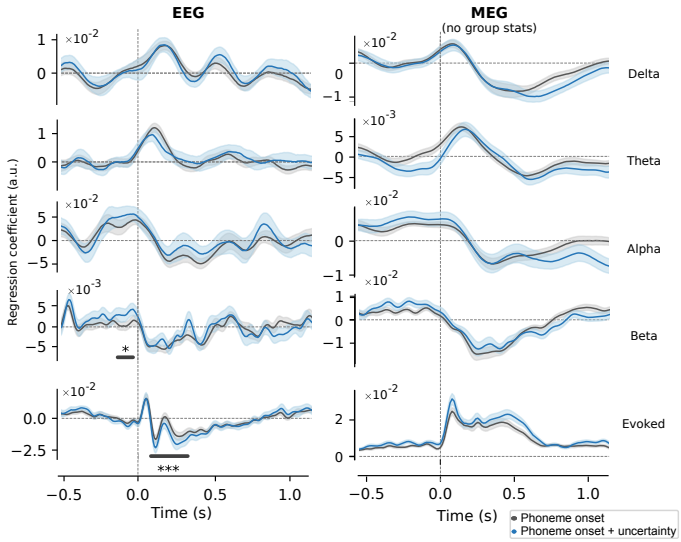

Figure 5.4. Increased pre-stimulus beta is associated with prior prediction uncertainty about the incoming word. Coefficients describing the effects of the prior prediction uncertainty about the incoming word (phoneme-by-phoneme lexical entropy) on the neural response. Response functions indicate the mean across the selected channels or sources (the top 10% most responsive channels for that frequency band, see *methods*), averaged across participants (EEG) or sessions (MEG). Shaded areas represent a bootstrapped standard error, stars indicate significance level using a cluster-based permutation t-test: $p < 0.05(*)$ across participants. As in Figure 5.3, in the MEG evoked coefficients, we took the absolute value before averaging across sources and sessions.

The apparent lack of sensitivity to uncertainty and surprise in the gamma band was unexpected, and might reflects low SNR for gamma-band activity fluctuations (as has been reported previously, see Dalal et al., 2009) rather than the genuine absence of an effect (see *Discussion*). Inspection of the coefficients – which describe *how* the frequency band amplitude changes over time as a function of the speech material – further supported the low SNR hypothesis: there was hardly a temporally consistent gamma response to the stimuli discernible in the EEG data; in the MEG data, it was only observed in 1/3 MEG participants (see Fig. S5.5). For the subsequent analyses, we therefore analysed the other frequency bands.

Phoneme surprise modulates responses in the theta (and delta) band

After establishing that surprise and uncertainty modulated the band-limited response in these frequency bands, we then asked whether these modulations were temporally specific and consistent across participants. To test this, we analysed the coefficients, which capture the time-resolved response function of the amplitude in each band.

Because responses in different bands can have different topographies, we used a functional ROI approach, selecting for each band the 10% of the channels or sources that were most sensitive to the stimulus, in terms of variance explained (by the baseline model, to avoid circularity; see *Methods).*

This revealed a clear stimulus-induced response in all bands in both datasets (Figure 5.3-5.4). The alpha and beta band specifically exhibit the characteristic post-stimulus decreases (typically interpreted as reflecting desynchronisations) which are also observed in traditional, controlled experiments (Engel and Fries, 2010; Spitzer and Haegens, 2017). This confirms that the approach worked, in that it captured characteristic pre- and post-onset induced responses (despite the absence of distinct pre- and post-stimulus periods).

We then tested for temporally specific modulations by surprise in each frequency band. We first focussed on the EEG dataset, because the larger number of participants allows for population-level statistical inference. Here we only found a significant modulation in amplitude of the theta band, based on a positive cluster between 430 and 560 ms (cluster-based permutation t-test: $P = 0.003$). For reference, Figure 5.3 also shows the effect of surprise on the evoked response, an effect that has been reported before (e.g. Donhauser and Baillet, 2020; Heilbron et al., 2021a).

In the MEG data, we observed similar temporal response functions, and again a clear desynchronisation in the alpha and beta band. We did not see as pronounced an effect in the theta band, but rather a strong modulation in the delta band. The structure of the MEG dataset (high with-subject power, low number of subjects) precludes the statistical quantification of this effect using a similar permutation tests at the group level (see *Methods* and *Discussion*) Nonetheless, the modulation in the delta band was highly consistent, and clearly visible in all three individuals (Figure S5.1).

Together, these results show how surprise (a probabilistic metric of prediction error) modulates the lower frequency band (theta and delta) but not other bands. These modulations in the low-frequency bands are broadly consistent with earlier studies on band-limited effects of linguistic predictability (e.g. Donhauser and Baillet, 2020; Piai et al., 2016; Rommers et al., 2017; see Prystauka and Lewis, 2019 for review), and of course with the effect of surprise on the evoked response (see Figures 5.3,S5.1 and *Discussion*).

Prior prediction uncertainty is associated with *increased* pre-stimulus beta

Next we tested how the responses in different bands were modulated by the prior prediction uncertainty about the identity of the incoming word.

For reference, we first analysed the evoked response, finding a clear negative

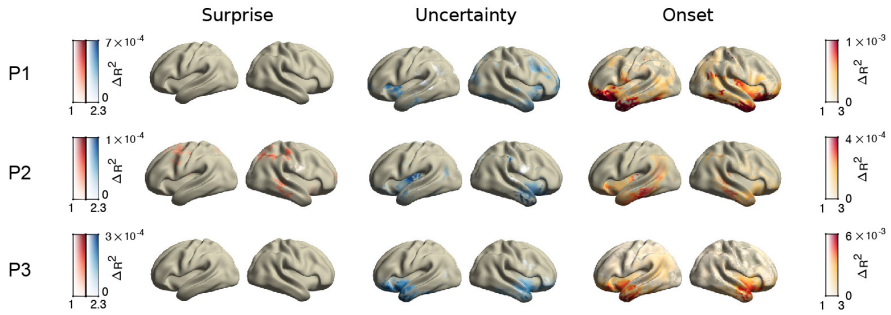

Figure 5.5. Prior prediction uncertainty modulates frontotemporal beta band activity Unique variance of beta band amplitude explained by phoneme surprisal, lexical entropy, and onset across cortical sources in each MEG participant. All plots use a dual coding scheme, in which the colour indicates the amount of additional variance explained (i.e. variance not explained by any other regressor or shared by multiple regressors), and opacity indicates FWE-corrected statistical significance. Note that $p < 0.05$ is equivalent to $-\log 10(p) > 1.3$.

modulation: in both EEG (cluster-based permutation t-test $P < 0.0001$, based on cluster between 85 and 330 ms) and MEG (Fig. 5.4). The effect of prior uncertainty was more pronouncedFig. earlier in the response in both datasets, which is in line with the idea that entropy specifically modulates earlier responses (Donhauser and Baillet, 2020); however, diverging from that proposal, the effect was not confined to the early response.

Then we turned to responses in frequency band amplitudes. In the EEG data, we found one statistically significant modulation in the beta band (Figure 5.4; cluster-based permutation t-test; $p = 0.026$, 2-tailed). The fact that this modulation was found in the beta band is in line with what we expected, taking entropy as an (inverse) proxy of top-down predictions strength. The effect occurred pre-stimulus onset, based on a cluster between -130 and -20 ms. Contrary to our expectations, however, the effect was *positive*: increased pre-stimulus beta was associated with more uncertain predictions about the incoming word. In the MEG data, the same pattern was observed: increased pre-stimulus beta for more uncertain predictions (and again a similar pattern of results in the alpha band, Figure 5.4). And again this effect was present in all 3 MEG participants (S5.2).

Interestingly, while the pre-stimulus increase was opposite to our expectations, it is in fact in line with prior literarture on language, and may not be necessarily inconsistent with predictive processing (see *Discussion*). The MEG coefficients also suggested a second pre-stimulus effect: an association between prior uncertainty and pre-stimulus theta amplitude. However, as this pattern was not observed in the EEG data, we refrain from interpreting it.

Prior uncertainty modulates beta amplitude in inferior frontal and anterior temporal cortex

Analysing the coefficients (Figs 5.3,5.4) provided a way to temporally characterise the association between prior uncertainty and beta activity. Next, we sought to spatially localise it.

For this question, we turn to the MEG data, which was source localised for this purpose. To obtain a spatial distribution of the association between beta amplitude and prior uncertainty, we computed for each source the amount of cross-validated variance in the beta amplitude that was *uniquely* explained by prior uncertainty; i.e. not explained by any other regressor. When this analysis is performed for each source independently, it produces a spatial map of the brain areas specifically sensitive to a given regressor (see e.g. Heer et al., 2017).

An additional advantage of analysing unique cross-validated variance, is that it is inherently robust to correlations between regressors. Here, the prior lexical uncertainty and surprisal of the previous phoneme are correlated ($\rho \approx 0.4$), meaning that the estimated coefficients are not guaranteed to correctly disentangle their contributions. This caveat does not apply to unique cross-validated explained variance, so this analysis also provides an additional test to confirm the specific link between prior uncertainty and beta amplitude in the MEG data[1].

Indeed, we found that prior prediction uncertainty explains significant unique variance in the beta band (see Figure 5.5). This aligns with the coefficient analysis, confirming the specific link between prior uncertainty and beta. In Figure 5.5, the effect of entropy on the beta amplitude appears more pronounced than that of surprisal. However, in most participants, the contrast between the contribution of surprise and that of uncertainty was itself not statistically significant when correcting for multiple comparisons. Therefore, the results confirm a specific effect of entropy on the beta band, but do not support a strong dissociation where only uncertainty (and not surprise) is associated with beta amplitude.

The effect of entropy on the beta band is spatially specific to inferior frontal and anterior temporal cortex in all participants (notwithstanding some individual variability in the exact distribution, see Figure 5.5). This spatial distribution seems a more general property of the beta response, rather than being specific to the beta-uncertainty association, as can be seen from the variance explained by the average response (i.e. the onset regressor, see Fig 5.5).

[1]Note that this primarily applies to the MEG data, because in the EEG data, second-level statistics on the coefficients already provides a way to control for the uncertainty in the coefficient estimates induced by the correlations.

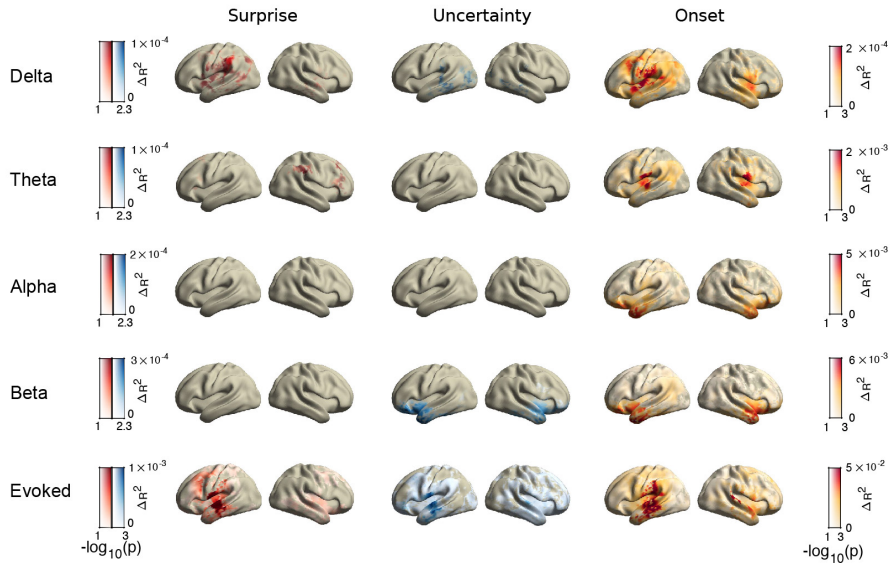

Figure 5.6. Feature importance for all bands and evoked response in a single, example participant. Unique beta band variance explained by phoneme surprisal, lexical entropy, phoneme onset across cortical sources in each MEG participant. In all plots, colour indicates amount of additional variance explained (i.e. variance not explained by any other regressor or shared by multiple regressors). Opacity indicates FWE-corrected statistical significance. Note that $p < 0.05$ is equivalent to $-\log 10(p) > 1.3$.

The specificity of this spatial distribution to the beta band becomes even clearer when comparing the distribution of the onset regressor in the beta band and other bands. As can be seen in Figure 5.6, in the beta (and alpha) bands, variance explained by the stimulus onset (i.e. by the average response to a phoneme) peaks in these anterior temporal and inferior frontal areas. By contrast, in the theta/delta band and evoked response, the variance explained by the onset peaks on more on temporal areas, arguably centred around auditory cortex. Since these anterior-temporal and inferior-frontal areas are considered key higher-order parts of the language network (see e.g. Hagoort, 2005; Hickok and Poeppel, 2007; Matchin and Hickok, 2020) this distribution is well in line with the interpretation of this beta activity as reflecting top-down language processing, but not in accord with an alternative interpretation of these beta fluctuations as reflecting top-down attentional engagement (see *Discussion*).

Discussion

Predictive processing models suggest that the brain continuously relies on top-down predictions, and that these top-down predictions are signalled via the beta and alpha bands. Here, we set out to test this hypothesis, by estimating the putative spectral signatures of prediction in an experiment in which participants simply listened to natural speech, without any form of top-down task engagement. Participants listened to audiobooks, the content of which we extensively analysed to estimate contextual unexpectedness (or surprise) and prior uncertainty in the prediction about the incoming word, on a phoneme-by-phoneme basis. The analyses revealed that surprise modulated the amplitude of the lower frequency bands (theta and delta). Moreover, and in line with predictive processing accounts, prior prediction uncertainty modulated the ongoing amplitude in the beta band. Source reconstruction located this beta modulation in inferior frontal and anterior temporal areas. Analysis of the temporal response function further suggests that this was partly driven by a modulation of pre-stimulus beta. However, the direction of this effect was the opposite of what we expected: beta amplitude was increased when prior predictions were more *uncertain*. Together, the results confirm that beta is associated with prior prediction uncertainty, even during naturalistic speech perception. However, they tentatively suggest that the direction of this beta band modulation is opposite of what one might expect based on predictive processing accounts of oscillations.

For phoneme surprise, we found that it modulated mainly (though not exclusively, see e.g. Fig. S5.4) the lower frequency bands (delta and theta). However, phoneme surprise also strongly modulates the *evoked* response (See Figures 5.3,S5.1; and Brodbeck, Hong, and Simon, 2018; Donhauser and Baillet, 2020; Heilbron et al., 2021a). Since the predominant waveforms of the evoked response are relatively slow and could be spectrally characterised in the delta/theta regime, the low-frequency modulations may simply reflect this evoked response modulation, rather than reflecting an oscillatory effect *per se*. One might argue that these responses are still oscillatory, for instance by casting the event related response as a dampened oscillation, with an early, fast (theta) component, and a later slower (delta) component (see e.g. Donhauser and Baillet, 2020). However, such an exogenous dampened oscillation is quite different from the endogenous oscillations as observed in LFPs, such as theta in the hippocampus (Piai et al., 2016) or in the deep layers of the rodent olfactory system, which both have been linked to top-down processing (Wang, 2010). Our analysis does not allow arbitrating between these options, and we would be reluctant to characterise the low-frequency modulations as endogenous oscillations.

For prior prediction confidence or contextual constraint (indexed through its inverse, lexical uncertainty) we observed effects primarily in the beta band (Figures 5.4,

5.5, 5.6 and S5.3-S5.4). Inspection of the coefficients suggested that this effect was driven by an *increase* of pre-stimulus beta by lexical uncertainty. This pattern was found both in the EEG dataset and in all MEG participants (see Figures 5.4 and S5.2). Although it was observed in both datasets, we should note that the effect can only be formally statistically evaluated in the EEG data (see below for discussion), where it only had modest statistical support. Given this modest statistical support and the unexpected direction, we believe that the effect may benefit from a replication in an additional dataset before drawing definitive conclusions.

Nevertheless, it is worth reflecting for a moment on the direction of this effect (i.e. on the fact that the modulation was *positive*). Our prior expectation was that beta activity should be stronger when prior predictions are stronger (and hence be *negatively* associated with prior uncertainty). This was based on the critical assumption that more confident prior predictions are associated with stronger top-down signalling, an assumption inspired by predictive processing interpretations of neural oscillations, both generally (Arnal and Giraud, 2012; Bastos et al., 2012) and oscillations in the context of linguistic prediction specifically (Lewis and Bastiaansen, 2015; Lewis, Wang, and Bastiaansen, 2015). While this reasoning is appealing, our initial hypothesis rested on a rather strong parallel between 'top-down' in the cognitive sense of having confident prior predictions, and 'top-down' in the functional and anatomical sense of signalling from 'higher' to 'lower' areas. On reflection, this parallel – between the cognitive level and implementational level – may not always hold. For instance, in our case, when prior confidence is high, there is little uncertainty about the identity of the incoming word, and only a small number of 'hypotheses' (lexical candidates) are activated. It is not obvious that in such a case of a 'strong prior prediction', top-down signalling is also necessarily the strongest. Indeed, one can make the opposite argument: perhaps top-down signalling is stronger when there are more potential lexical candidates, resulting in multiple, competing top-down hypotheses. Following this argument, pre-stimulus beta would be higher when there is more uncertainty – which is of course the pattern that we found.

What this illustrates is that going from cognitive-level 'prediction strength' (here operationalised via lexical entropy) to implementational-level top-down signalling (here probed via the beta band amplitude) requires additional assumptions about the processing architecture that do not follow from the predictive processing framework itself. Future work could address this more rigorously by making these assumptions explicit in a neural network model and performing simulations to motivate a specific, quantitative relation between the metric of contextual constraint used (e.g. lexical entropy) and the strength of top-down signalling, which can then be used to motivate empirical predictions (or re-interpret existing findings).

Predictive processing theory aside, it is interesting to note that the *increase* of

pre-stimulus beta by contextual uncertainty that we found, is in fact in line with most neurolinguistic studies on the oscillatory correlates of contextual constraint in language processing (Li et al., 2017; Piai, Roelofs, and Maris, 2014; Piai et al., 2015; Rommers et al., 2017; Wang, Hagoort, and Jensen, 2018; see Prystauka and Lewis, 2019 for review). In these studies, just prior to the presentation of a critical word or picture, alpha and beta band activity was *increased* when contextual constraint was weaker (i.e. when lexical uncertainty was higher). Moreover, source reconstructions by Piai et al. (2015) localised this modulation of pre-stimulus beta to anterior temporal sources that are quite similar to the spatial distribution we observe (Figure 5.5). In that study (see also Piai, Roelofs, and Maris, 2014) participants performed a naming task, and reduction of pre-stimulus beta for highly constraining contexts was interpreted as reflecting articulatory preparation. Other studies have similarly interpreted pre-stimulus alpha/beta band modulations as attentional or task effects (Rommers et al., 2017). Intriguingly, we observe a similar modulation of pre-stimulus beta in participants not engaged in any task, simply listening to continuous speech. This could imply that the beta modulation is not (just) related to task effects, but at least in part to language processing itself. One interpretation is that, in weakly constraining contexts, more lexical candidates are activated and have to compete, which may recruit more top-down signalling – perhaps to accomplish top-down 'lexical selection', in the parlance of Marslen-Wilson (1987). The localisation of beta to anterior temporal and inferior frontal areas (Figs 5.5,5.6) is in line with the modulation originating from higher-order areas in the language network, rather than fronto-parietal areas in the attention network. Of course, we stress that this is a post-hoc interpretation of our findings, and not a conclusion that can be said to be supported by the results of this study.

Another surprising finding was that we did not find modulations by predictability in the gamma band. We believe that this reflects a limitation of our non-invasive data, rather than a property of the neural activity itself. Specifically, the (public) EEG dataset we used was downsampled to 128 Hz, forcing us to focus on a subset of the lowest gamma frequencies. In the MEG dataset, a stimulus-induced gamma response was observed, but only in one of the three participants (Figure S5.5). This individual variability and low SNR for MEG gamma is in line with studies employing simultaneous iEEG and MEG recordings, finding that while gamma was always present in the invasive recordings, it was much weaker in the MEG signals, with considerable variability between participants (Dalal et al., 2009).

A limitation of this study is that while the investigated datasets comprise a large amount of samples per participant, the number of participants was relatively low. While the 19 participants in the EEG dataset are enough for population inference, it only allows for observing effects with a relatively high consistency across par-

ticipants. An obvious and effective way to address this limitation is to extend the analysis to a different dataset with more participants. Another option might be to try to improve the analysis itself – in particular the time-resolved regression. The model-fit on band limited amplitudes was lower than on the original signal (i.e. the real component), marking a potential room for improvement. One way to improve the fit is by changing the forward model – for instance using different basis functions than the impulses employed here, which have many degrees of freedom. This may result in effects that are more consistent, allowing more confident inferences across the 19 participants.

Aside from improving the time-resolved regression, statistical inference provides another avenue for improvement. In the MEG dataset, for simple univariate comparisons we a used multi-level non-parametric procedure (hierarchical bootstrapping) to aggregate across the recordings of all MEG participants (Saravanan, Berman, and Sober, 2020). We are unaware of a similar multi-level approach to the spatiotemporal clustering statistics that we use on brain data. However, extending the method to mass-univariate spatiotemporal clustering would be highly useful. This would not allow for population-level inference from the 3 participants, but it would enable statistical inferences about the 3 individuals combined, by sharing statistical strength across the 30 MEG recording sessions, while appropriately handling individual vs within-subject variance. This would allow unified statistical inferences about the coefficients – and therefore the pre-stimulus beta modulation, see Figure 5.4 – across the full MEG dataset.

A final point regards our focus on phonemes. By analysing processing on a phoneme-by-phoneme basis we do not want to make any theoretical commitments about the psychological reality or ontological status of phonemes as such. One could imagine analysing different units that may be theoretically more satisfying or empirically providing a better fit to the brain data. While this is an interesting question, it is orthogonal to the purpose of this study. Here, analysing phonemes simply provided a convenient and powerful operationalisation of incremental, predictive processing of speech at the sub-lexical level.

In conclusion, we have investigated oscillatory signatures of predictive processing during naturalistic speech perception. Results revealed that phoneme surprise modulated the amplitude of the theta and delta band, while prior uncertainty in the prediction about the incoming word modulated beta band in frontotemporal areas. Preliminary investigations of the coefficients suggest that prior uncertainty may enhance pre-stimulus beta. Together, the study demonstrates the feasibility of studying the oscillatory correlates of linguistic prediction in natural conditions. However, it also highlights how deriving testable hypotheses about the relation between prediction and specific frequency bands is less straightforward than it may seem, and

requires assumptions that do not follow from predictive processing itself. Future work could address this by making these assumptions explicit in computational simulations.

Methods

We analysed EEG and source localised MEG data from two experiments. The EEG data is part of a public dataset that has been published about before (Broderick et al., 2018). The MEG dataset is part of a resource described in detail in Armeni (2021). We have previously analysed prediction signatures in the evoked responses (Heilbron et al., 2021a).

Participants

All participants were native English speakers. In the EEG experiment, 19 subjects (13 male) between 19 and 38 years old participated; in the MEG experiment, 3 subjects participated (2 male) aged 35, 30, and 28. Both experiments were approved by local ethics committees (EEG: ethics committee of the School of Psychology at Trinity College Dublin; MEG: CMO region Arnhem-Nijmegen).

Stimuli and procedure

In both experiments, participants were presented long segments of narrative speech extracted from audiobooks. The EEG experiment used the first chapters of Hemingway's *The Old Man and the Sea*. The MEG experiment used 10 stories from Doyle's *The Adventures of Sherlock Holmes*. In total, EEG subjects listened to 1 hour of speech (containing 11,000 words and 35,000 phonemes); MEG subjects listened to 9 hours of speech (containing 85,000 words and 290,000 phonemes).

In the EEG experiment, each participants performed only a single session, which consisted of 20 runs of about 180s long, amounting to the first hour of the book. Participants were instructed to maintain fixation and minimise movements but were otherwise not engaged in any task. In the MEG experiment, each participant performed a total of 10 sessions, each 1 hour long. Each session was subdivided in 6-7 runs of about ten minutes long (runs were subdivided such that prominent narrative events were not split across runs). Unlike in the EEG experiment, participants in the MEG dataset participants were answer questions in between runs: one multiple choice comprehension question, a question about story appreciation (scale 1-7) and a question about informativeness.

Stimulus annotation

For both datasets, auditory stimuli were analysed using a forced alignment procedure to derive onset and offset times for each word and phoneme. For the EEG dataset, this was performed using the Prosodylab forced aligner; for the MEG dataset, the Penn Forced Aligner Toolkit was used. More details on the procedures is found in original publications, see (Armeni, 2021; Di Liberto, O'Sullivan, and Lalor, 2015).

Data acquisition and pre-processing

The EEG data were acquired using a 128-channel (plus two mastoid channels) using an ActiveTwo system (BioSemi) at a rate of 512 Hz, and downsampled to 128 Hz before being distributed as a public dataset. We visually inspected the raw data to identify bad channels, and performed independent component analysis (ICA) to identify and remove blinks; rejected channels were linearly interpolated with nearest-neighbour approach using MNE-python (Gramfort et al., 2014)

The MEG data were acquired using a 275 axial gradiometer system at 1200 Hz. To minimise head motion, individualised 3D printed headcasts were created for each participant, which enabled average displacement of less than 0.5 mm across the 10 sessions of recordings (see Armeni, 2021, Chapter 3 for details). For the MEG data, preprocessing and source modelling was performed in MATLAB 2018b using fieldtrip (Oostenveld et al., 2011). We applied notch filtering (Butterworh IIR) at the bandwidth of 49–51, 99–101, and 149–151 Hz to remove line noise. To identify and remove eye blink artifacts, independent component decomposition was performed using the FastICA algorithm.

Frequency decomposition

Recordings were decomposed into different frequency bands using the Hilbert transform. This involves performing a band-pass filter (parameters of which detailed below) to preserve the activity in a certain frequency range (e.g. alpha), performing a Hilbert transform on the band-limited signal, and taking the absolute value of the analytic signal to compute the instantaneous amplitude (square root of power) at each point in time.

We folllowed (Jensen, Spaak, and Zumer, 2014) for the definition of the delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta (12-22 Hz) bands. For gamma, we used different definitions for EEG and MEG. Because the EEG dataset was downsampled, we only considered low gamma, defined as 30-45 Hz (upper limit here defined by the line noise). For MEG, we employed a data-driven approach, exploring 4 'bins' of the gamma range: 55-75, 75-95, 105-125, 125-145 Hz. For each band we then fitted

an onset-only model, to identify in which band we could best identify a stimulus-induced response. As can be seen in Fig. S S5.5, this revealed that the stimulus induced response was the most pronounced in the 75-95 Hz range, which then became our definition for the gamma band.

For the filters in the filter-hilbert procedure, we used non-causal, one-pass zero-phase FIR filters, implemented using the time-domain windowed (firwin) method. We used the following filter transition bandwidths. For delta (0.5-4 Hz), we used a lower transition bandwidth of 0.5 Hz (-6 dB cutoff frequency 0.25 Hz) and an upper bandwidth of 2 Hz (cutoff frequency 5 Hz). For theta (4-8 Hz) we used a 2 Hz transition bandwidth. For alpha (8-12 Hz) we used transition bandwidths of 2 and 3 Hz. For beta (12-22 Hz) we used transition bandwidths of 3 and 5.5 Hz. For low-gamma (EEG; 30-45 Hz) we used transition bandwidths of 7.5 and 11 Hz. For high-gamma (MEG; 75-95 Hz) we used transition bandwidths of 18.75 and 23.75 Hz.

Importantly, for the MEG data the frequency decomposition was performed as a part of the source-localization, using a hilbert-beamformer procedure (Westner and Dalal, 2019); see below for details.

Anatomical MRI acquisition and headcast

Headcasts of the MEG participants were based on MRI scans with a 3T MAGNETOM Skyra MR scanner (Siemens AG). For this, a fast low angle shot (FAST) sequence was used with the following image acquisition parameters: slice thickness of 1 mm; field-of-view of 256 × 256 × 208 mm along the phase, read, and partition directions respectively; TE/TR = 1.59/4.5 ms.

Head and source models

As part of the MEG dataset, head models and source models are provided. For these models, MEG sensors were co-registered to the subjects' anatomical MRIs using position information of three localization coils attached to the headcasts. To create source models, FSL's Brain Extraction Tool was used to strip non-brain tissue (Smith et al., 2004). Subject-specific cortical surfaces were reconstructed using Freesurfer (Dale, Fischl, and Sereno, 1999), and post-processing (downsampling and surface-based alignment) of the reconstructed cortical surfaces was performed using the Connectome Workbench command-line tools (v 1.1.1). This resulted in cortically-constrained source models with 7,842 source locations per hemisphere. The lead-fields were provided with the MEG dataset are computed based on a single-shell volume conduction models based on the inner surface of the skull.

Beamformer and parcellation

To estimate the source time series from the MEG data, we used linearly constrained minimum variance (LCMV) beamforming, using Fieldtrip (Oostenveld et al., 2011). Beamforming was performed separately for each session, using a unit-noise-gain weight normalisation, assuming a fixed orientation, and applying a lambda regularisation parameter of 100%. The beamforming procedure was combined with the Hilbert frequency decomposition, into a unified Hilbert-beamformer procedure described earlier by Westner and Dalal (2019)). Specifically, this entails estimating the covariance matrix for each frequency band separately, resulting in different spatial filters for each frequency band. The procedure was performed for each session independently, spatial filters were then averaged across sessions to create one spatial filter per frequency band. To obtain source localised estimates of instantaneous power or amplitude, the Hilbert transform was then performed at the the sensor-level, projecting the analytic signal to source space, and taking the absolute at the source level to obtain instantaneous amplitude at each source.

To perform regression, source amplitude could be derived in two ways. Either beamforming is performed first and regression then performed at the source level directly, or regression can be performed on the sensor-level amplitudes and then evaluated on the source level (see below for details on model estimation). To be able to perform regression at the source level directly, we had to reduce the dimensionality of the source space. For this we used the same parcellation procedure described in Armeni (2021) and Heilbron et al. (2021a), which uses a refined version of the Conte69 atlas, which is based on Brodmann's areas. To this end, we computed, for each session, parcel-based time series by taking the first principal component of the aggregated time series of the dipoles belonging to the same cortical parcel. This resulted in band-limited amplitudes at 370 parcels.

Neural language model: GPT-2

Estimates of contextual predictions were computed using a language model – a model computing the probability of each word given the preceding words. Here, we used GPT-2 (XL) – currently among the best publicly released English language models. GPT-2 is a transformer-based model, that in a single pass turns a sequence of tokens (representing either whole words or word-parts) $U = (u_1, \ldots, u_k)$ into a sequence of conditional probabilities, $(p(u_1), p(u_2|u_1), \ldots, p(u_k \mid u_1, ..., u_{k-1}))$.

Roughly, this happens in three steps: first, an embedding step encodes the sequence of symbolic tokens as a sequence of vectors, which can be seen as the first hidden state h_o. Then, a stack of n transformer blocks each apply a series of operations resulting in a new set of hidden states h_l, for each block l. These blocks consist

of a multi-headed self attention layer, a feedforward layer and normalisation step (see Liu et al., 2018; Radford et al., 2019; Vaswani et al., 2017 for details). Finally, a (log-)softmax layer is applied to compute (log-)probabilities over target tokens.

In other words, the model can be summarised as:

$$h_0 = UW_e + W_p \tag{5.1}$$

$$h_l = \text{transformer\_block}\left(h_{l-1}\right) \forall i \in [1, n] \tag{5.2}$$

$$P(u) = \text{softmax}\left(h_n W_e^T\right), \tag{5.3}$$

where W_e is the token embedding and W_p is the position embedding. In total, GPT-2 (XL) contains $n = 48$ blocks, with 12 heads each; a dimensionality of $d = 1600$ and a context window of $k = 1024$, yielding a total 1.5×10^9 parameters. Note that k refers to the number of Byte-Pair Encoded *tokens*. A token can be either a word or (for less frequent words) a word-part, or punctuation. For words spanning multiple tokens, we computed the word probability as the joint probability of the constituent tokens. We used the PyTorch implementation of GPT-2 provided by HuggingFace's *Transformers* package (Wolf et al., 2020).

Phoneme-by-phoneme linguistic predictions

We used the word-by-word contextual predictions derived from GPT-2 to calibrate phoneme-level predictions about the incoming word. This was done using the same modelling scheme developed and evaluated in Heilbron et al. (2021a). This scheme in turn was is inspired by and is an extension of the approach described by Gwilliams et al. (2018) (see also Ettinger, Linzen, and Marantz, 2014), to compute phoneme-by-phoneme statistics in a non-contextual fashion; i.e. without taking the long term context (here, estimated using GPT-2) into account. In brief, the method involves selecting, for each phoneme coming in, the the 'cohort' of words consistent with the phonemes presented so far. For the first word, this cohort is simply equivalent to the full set of words (i.e. the distribution over the entire vocabulary). To incorporate long-term context, each word is assigned a prior probability in that specific discursive context, which we derive from the contextual word probability estimated by GPT-2. This results in a probability distribution, expressing the probability of each word w in the cohort C, over which the Shannon entropy is computed:

$$-\sum_{w \in C} P(w \mid C) \log P(w \mid C). \tag{5.4}$$

This quantity, lexical entropy, is then used to express the prior uncertainty about the incoming word.

To quantify unexpectedness, phoneme surprisal was computed. To this end, we compute the probability that that a given phoneme at time φ_t is of a specific identity (A), given the prior phonemes within a word:

$$P(\varphi_t = A \mid \varphi_{1:t-1}) = \frac{P(C_{\varphi_t=A})}{P(C_{\varphi_{1:t-1}})}. \tag{5.5}$$

Here, $P(C_{\varphi_t=A})$ denotes the cumulative probability of all words in the remaining cohort of candidate words if the next phoneme were A, and $P(C_{\varphi_{(1:t-1)}})$ denotes the cumulative probability of all words in the prior cohort. To efficiently compute Equations (5.4) and (5.5) for every phoneme, we constructed a statistical phonetic dictionary as a digital pronunciation tree using the vocabulary from the CMU dictionary and the lexical statistiscs from SUBTLEX (Brysbaert and New, 2009; Weide, 1998). Missing words or alternative pronunciations that occurred in the audiobooks but not in the in the CMU pronunciation dictionary, were manually added to the pronunciation tree.

Using the contextual probabilities inside the phoneme model means a new pronunciation tree model has to be constructed for each word in the text. To simplify this process, we used the procedure from Heilbron et al. (2021a), which involves only using the 'nucleus' of the top k predicted words with a cumulative probability of 0.9, and truncated the (less reliable) tail of the distribution. Further, we simply assumed that the rest of the tail was 'flat' and had a uniform probability. We can think of the probabilities in the flat tail as having a (laplacian) 'pseudocount' of 1. If we express the prior probabilities in the nucleus as implied 'pseudofrequencies', the cumulative implied nucleus frequency is complementary to the tail length, which is simply the difference between the vocabulary and nucleus size ($V-k$). This means that for word i in the text, we can express the nucleus as implied frequencies as:

$$\text{freqs}_{\psi} = P_{tr}(w^{(i)}|\text{context})\frac{V-k}{1-\sum_{j=1}^{k} P(w_j^{(i)}|\text{context})}, \tag{5.6}$$

where $P_{tr}(w^{(i)}|\text{context})$ is the trunctated lexical prediction, and $P(w_j^{(i)}|\text{context})$ is the predicted probability that word i in the text is word j in the sorted vocabulary. Note that using this flat tail not only simplifies the computation, but also deals with the fact that the vocabulary of GPT-2 is smaller than that of the pronunciation model. As such, using the flat tail means we can still use the full vocabulary (e.g. to capture phonotactic regularities), while using 90% of the contextual probability density from GPT-2.

Time resolved regression

To quantify the time-resolved modulations of the band-limited amplitude by incoming stimuli, we used a time-resolved regression technique. Simply put, this involves using impulse regressors for both constants and covariates defined at phoneme onsets, and then temporally expanding the design matrix such that each predictor column C becomes a series of columns over a range of temporal lags $C_{t_{min}}^{t_{max}} = (C_{t_{min}}, ..., C_{t_{max}})$. For each predictor one thus estimates a series of weights $\beta_{t_{min}}^{t_{max}}$ (Fig. 5.2) which can be understood as the *modulation function* describing how a given regressor modulates the continuous amplitude response over time. Because we use impulses as a basis function, the procedure mathematically equivalent to FIR deconvolution method in FMRI or rERP (or impulse TRF modelling) in EEG analysis (Goutte, Nielsen, and Hansen, 2000; Lalor et al., 2006; Smith and Kutas, 2015).

Here, we use a temporal lags between -0.55 and 1.4 seconds. All data and regressors were standardised and coefficients were estimated with ℓ_2-norm regularised (Ridge) regression, using the scikit learn sparse matrix implementation (Pedregosa et al., 2011). Regularisaion parameters were set based on leave-one-run-out R^2 comparison.

For regression estimation, different procedures were followed to estimate the coefficients and to evaluate the model predictive performance. Specifically, for estimation of the coefficients regression was performed at the source level directly. This has the distinct advantage that since the regression is done on (always positive) source-level amplitudes, the sign of the regression coefficient can be interpreted as an increase or decrease in amplitude. However, this requires reducing the dimensionality, so for this we used the parcellated source space (307 parcels). For evaluation, we fit the regression on the sensor-level amplitudes (on the training data)

Model comparison

In both datasets, model comparison was based on comparing cross-validated correlation coefficients. Cross-validation was performed in a leave-one-run-out cross-validation scheme, amounting to 19-fold cross-validation in the EEG data and between 63 and 65-fold cross-validation for the MEG data (in some subjects, some runs were discarded due to technical problems).

ROI definition

To reduce the dimensionality of the coefficient analysis we used an ROI approach, testing, in each subject, only the EEG channels our sources that that had the strongest signal, in terms of its sensitivity to the stimulus material. Because responses in different bands and subjects can have different topographies, the ROIs were defined

functionally, selecting for each band the 10% of the channels or sources that were most sensitive to the stimulus, in terms of variance explained by the baseline model (capturing purely the low-level stimulus features).

Regression models

We considered two models. First, a baseline model which functioned as a non-predictive processing baseline. To capture the fluctuation in the envelope (which is the overwhelming driver of neural responses to speech acoustics) we computed acoustic energy (quantified as envelope variance) of every phoneme. This captures the fact that some speech sounds are louder than others (e.g. strong vowels or stressed syllables). To capture significant linguistic events, we included the onset of every phoneme and every word and sentence boundary. The predictive processing model included these same baseline regressors, plus the lexical entropy and phonemic surprisal. All covariates were defined and included for every phoneme.

Statistical comparison

All statistical tests were two-tailed and used an alpha of 0.05. For all simple univariate tests performed to compare model-performance within and between subjects, we used a bootstrap t-test, a robust non-parametric hypothesis test. This involves estimating the observed t-statistic in the data, and comparing this to a null distribution by resampling the same data with zero mean. The p-value is then simply the proportion of samples from the null-distribution that resulted in a t-statistic at least as extreme as the observed t-statistic. For the MEG data, we performed the same procedure, but using a multi-level hierarchical bootstrapping procedure, as described in (Saravanan, Berman, and Sober, 2020).

To perform statistics on the coefficients (Figures 5.3,5.4), we performed temporal cluster permutation tests as implemented in MNE (Gramfort et al., 2014; 10,000 permutations per test). In the MEG, multiple comparison correction for comparison of explained variance across cortical areas was done using Threshold Free Cluster Enhancement (TFCE; Smith and Nichols, 2009). Mass-univariate tests were based on one-sample t-tests plus the 'hat' variance adjustment method with $\sigma = 10^{-3}$ (Ridgway et al., 2012).

Acknowledgements

We are grateful for all the authors of the open source software packages that have made this project possible. We want to specifically thank Kristijan and Jan Mathijs for sharing the MEG dataset and to the Lalor lab for sharing the EEG dataset.

Supplementary information

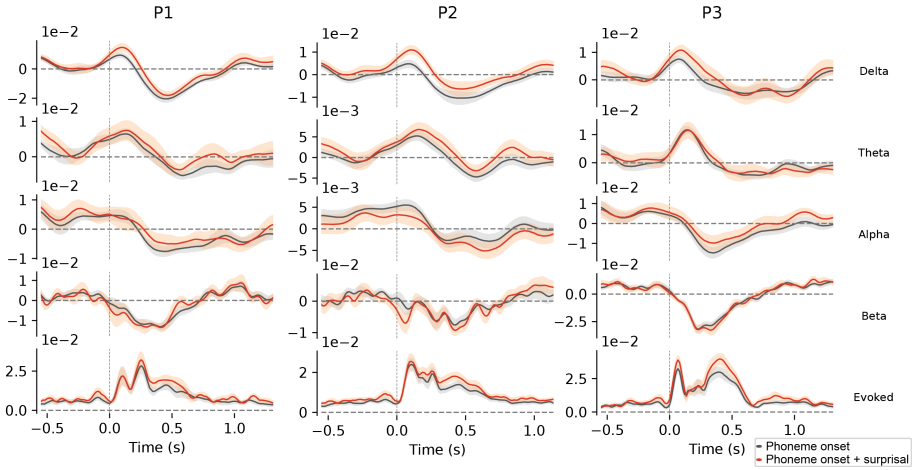

Figure S5.1. All coefficients for phoneme surprise. Same as in Figure 5.3 but for each individual MEG participant. Shaded bars indicate bootstrapped standard error across sessions.

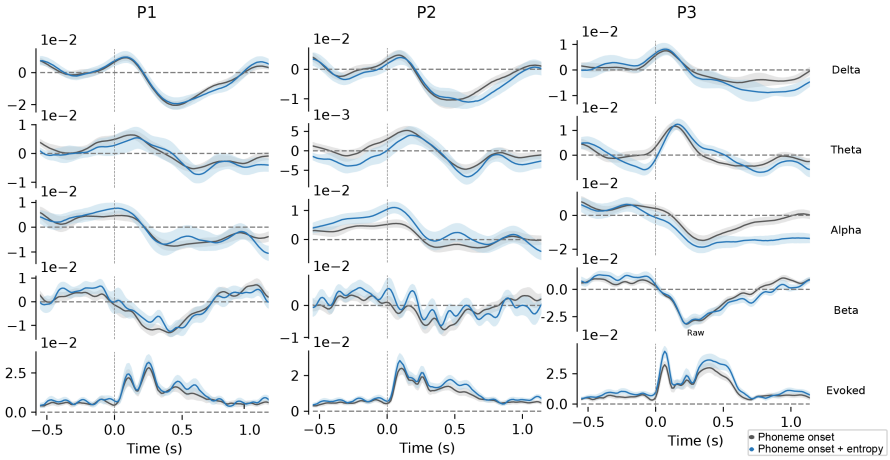

Figure S5.2. All coefficients for prior uncertainty (lexical entropy). Same as in Figure 5.4 but for each individual MEG participant. Shaded bars indicate bootstrapped standard error across sessions.

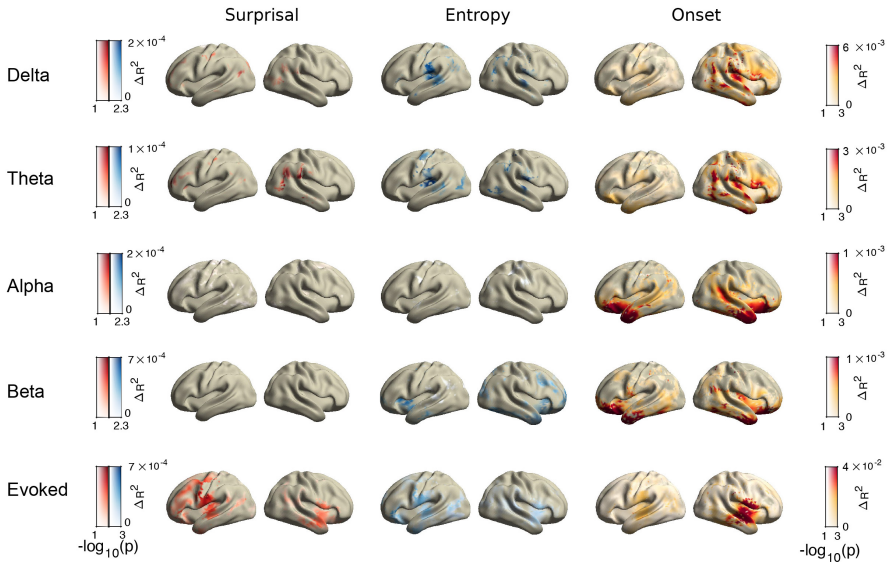

Figure S5.3. Feature importance for all bands and evoked response in participant 1. Feature importance of surprise, uncertainty and stimulus onset in all bands in participant 1. In all plots, colour indicates amount of additional variance explained (i.e. variance not explained by any other regressor or shared by multiple regressors). Opacity indicates FWE-corrected statistical significance. Note that $p < 0.05$ is equivalent to $-\log 10(p) > 1.3$.

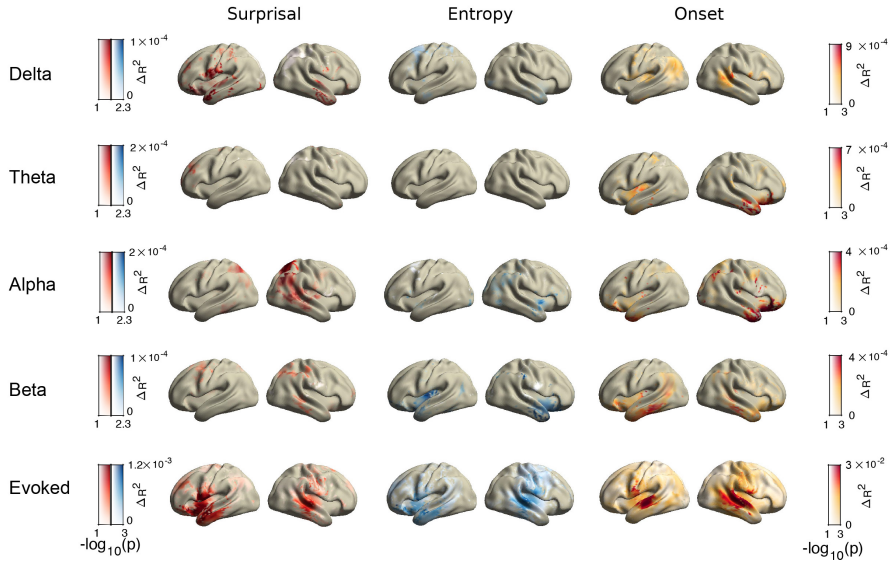

Figure S5.4. Feature importance for all bands and evoked response in participant 2. Unique beta band variance explained by phoneme surprisal, lexical entropy, phoneme onset across cortical sources in each MEG participant. In all plots, colour indicates amount of additional variance explained (i.e. variance not explained by any other regressor or shared by multiple regressors). Opacity indicates FWE-corrected statistical significance. Note that $p < 0.05$ is equivalent to $-\log 10(p) > 1.3$.

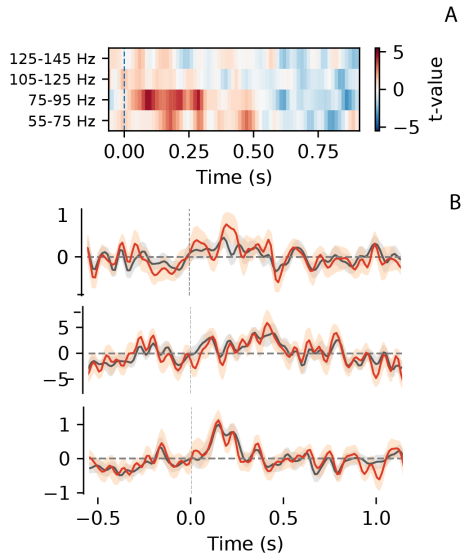

Figure S5.5. Gamma response coefficients in all bands and participants. A) t-scores of phoneme-induced gamma response across all 30 sessions (collapsing across participants). This result clearly shows the response is centered between 75-95 Hz (and hence is more than a broadband expression of the evoked response). **B)** Phoneme induced response (and modulation by phoneme surprisal, red) in the 75-95 Hz in all 3 participants. The black lines (which capture the average response) show that the stimulus induced response is only reliably observed in participant 3.

Chapter 6

Prediction and preview strongly affect reading times but not skipping during natural reading

Abstract

In a typical text, readers look much longer at some words than at others and fixate some words multiple times, while skipping others altogether. Historically, researchers explained this variation via low-level visual or oculomotor factors, but today it is primarily explained via cognitive factors, such as how well words can be predicted from context or discerned from parafoveal preview. While the existence of these effects has been well established in experiments, the relative importance of prediction, preview and low-level factors for eye movement variation in natural reading is unclear. Here, we address this question in three large datasets (n=104, 1.5 million words), using a deep neural network and Bayesian ideal observer to model linguistic prediction and parafoveal preview from moment to moment in natural reading. Strikingly, neither prediction nor preview was important for explaining word skipping – the vast majority of skipping was explained by a simple oculomotor model. For reading times, by contrast, we found clear but independent contributions of both prediction and preview, and effect sizes matching those from controlled experiments. Together, these results challenge dominant models of eye movements in reading by showing that linguistic prediction and parafoveal preview are not important determinants of word skipping.

This chapter is based on:
Heilbron M., van Haren, J. Hagoort P., de Lange F.P. 2021. Prediction and preview strongly affect reading times but not skipping during natural reading. *bioRxiv.*

Introduction

When reading a text, readers move their eyes across the page to bring new information to the centre of the visual field, where perceptual sensitivity is highest. While it may subjectively feel as if the eyes smoothly slide along the text, they in fact traverse the words with rapid jerky movements called *saccades*, followed by brief stationary periods called *fixations*. Across a text, saccades and fixations are highly variable and seemingly erratic: Some fixations last less than 100 ms, others more than 400; and while some words are fixated multiple times, many other words are skipped altogether (Dearborn, 1906; Rayner and Pollatsek, 1987). What explains this striking variation?

Historically, researchers have pointed to low-level non-linguistic factors like word length, oculomotor noise, or the relative position where the eyes happen to land (Bouma and Voogd, 1974; Buswell, 1920; Dearborn, 1906; O'Regan, 1980). Such explanations were motivated by the idea that oculomotor control was largely *autonomous*. In this view, readers can adjust saccade lengths and fixation durations to global characteristics like text difficulty or reading strategy, but not to subtle word-by-word differences in language processing (Bouma and Voogd, 1974; Buswell, 1920; Dearborn, 1906; Morton, 1964).

As reading was studied in more detail, however, it became clear that the link between eye movements and cognition was more direct. For instance, it was found that fixation durations were shorter for words with higher frequency (Inhoff, 1984; Rayner, 1977). Eye movements were even shown to depend on how well a word's identity could be inferred *before* fixation. Specifically, researchers found that words are read faster and skipped more often if they are *predictable* from linguistic context (Balota, Pollatsek, and Rayner, 1985; Ehrlich and Rayner, 1981) or if they are identifiable from a *parafoveal preview* (McConkie and Rayner, 1975; Rayner, 1975; Schotter, Angele, and Rayner, 2012). These demonstrations of a direct link between eye movements and language processing overturned the autonomous view, replacing it by cognitive accounts describing eye movements during reading as largely, if not entirely, controlled by linguistic processing (Clifton et al., 2016; Reichle, Rayner, and Pollatsek, 2003). Today, many studies still build on classic techniques like gaze-contingent displays, but now to ask much more detailed questions, like whether word identification is a distributed or sequential process (Kliegl, Nuthmann, and Engbert, 2006; Kliegl, Risse, and Laubrock, 2007); how many words can be processed in the parafovea (Rayner, Juhasz, and Brown, 2007), at which level they are analysed (Hohenstein and Kliegl, 2014), and how this might differ between writing systems or orthographies (Tiffin-Richards and Schroeder, 2015; Yan et al., 2010).

Here, we ask a different, perhaps more elemental question: how much of the vari-

ation in eye movements do linguistic prediction, parafoveal preview preview, and non-linguistic factors each explain? That is, how important are these factors for determining how the eyes move during reading? Dominant, cognitive models explain eye movement variation primarily as a function of ongoing processing. Skipping, for instance, is modelled as the probability that a word is identified before fixation (Engbert and Kliegl, 2003; Engbert et al., 2005; Reichle, Rayner, and Pollatsek, 2003). Some, however, have questioned this purely cognitive view, suggesting that low-level features like word length might be more important (Drieghe et al., 2004; Reilly and O'Regan, 1998; Vitu et al., 1995). Similarly, one may ask what drives next-word identification: is identifying the next word mostly driven by linguistic predictions or by parafoveal perception? Remarkably, while it is well-established that both linguistic and oculomotor, and both predictive and parafoveal processing, all affect eye-movements (Drieghe et al., 2004; Kliegl et al., 2004; Schotter, Angele, and Rayner, 2012; Staub, 2015), a comprehensive picture of the their relative explanatory power is currently missing, perhaps because they are seldom studied all at the same time.

To arrive at such a comprehensive picture we focus on natural reading, analysing three large datasets of participants reading passages, long articles, and even an entire novel – together encompassing 1.5 million (un)fixated words, across 108 individuals (Cop et al., 2017; Kennedy, 2003; Luke and Christianson, 2018). Instead of manipulating word predictability or perturbing parafoveal perceptibility, we combine deep neural language modelling (Radford et al., 2019) and Bayesian ideal observer analysis (Duan and Bicknell, 2020) to quantify how much information is conveyed by both factors, on moment-by-moment basis. This way, we can probe the effect of both prediction and preview on *each* word during natural reading. Such a broad-coverage approach has been applied to the effects of predictability on reading before (Frank et al., 2013; Goodkind and Bicknell, 2018; Kliegl et al., 2004; Luke and Christianson, 2016; Smith and Levy, 2013), but either without considering preview or only through coarse heuristics such as using frequency as a proxy for parafoveal identifiability (Kennedy et al., 2013; Kliegl, Nuthmann, and Engbert, 2006; Pynte and Kennedy, 2006) (cf. Duan and Bicknell, 2020). By contrast, here we explicitly model both, in addition to low-level explanations like autonomous oculomotor control. To assess explanatory power, we use set theory to derive the unique and shared variation in eye movements explained by each model.

To preview the results, this revealed a striking dissociation between skipping and reading times. For word skipping, the overwhelming majority of variation could be explained – mostly *uniquely* explained – by a non-linguistic oculomotor model. For reading times, by contrast, we found strong effects of both prediction and preview, tightly matching effect sizes from controlled designs. Interestingly, linguistic prediction and parafoveal preview seem to operate independently: we found strong

evidence against Bayes-optimal integration of the two. Together, these results challenge dominant cognitive models of reading, and show that skipping (or the decision of *where* to fixate) and reading times (i.e. *how long* to fixate) are governed by different principles.

Results

We analysed eye movements from three large datasets of participants reading texts ranging from isolated paragraphs to an entire novel. Specifically, we considered three datasets: Dundee Kennedy, 2003 (N=10, 51.502 words per participant), Geco Cop et al., 2017 (N=14, 54.364 words per participant) and Provo Luke and Christianson, 2018 (N=84, 2.689 words per participant). In each corpus, we analysed both skipping and reading times (indexed by gaze duration), as they are thought to reflect separate processes: the decision of *where* vs *how long* to fixate, respectively (Drieghe et al., 2004; Reichle, Rayner, and Pollatsek, 2003).

To estimate the effect of linguistic prediction and parafoveal preview, we quantified the amount of information conveyed by both factors for each word in the corpus (for preview, this was tailored to each individual participant, since each word was previewed at a different eccentricity by each participant). To this end, we formalised both processes as a probabilistic belief about the identity of the next word, given either the preceding words (prediction) or a noisy parafoveal percept (preview; see Figure 6.1a). As such, we could describe these disparate cognitive processes using a common information-theoretic currency. To compute the probability distributions, we used GPT-2 for prediction (Radford et al., 2019) and a Bayesian ideal observer for preview (Duan and Bicknell, 2020) (see Figure 6.1b and *Methods*).

Prediction and preview increase skipping rates and reduce reading times

We first asked whether our formalisations allowed us to observe the expected effects of prediction and preview, while statistically controlling for oculomotor and lexical variables in a multiple regression model. Because the decisions of whether to skip and how long to fixate a word are made at different moments, we modeled each separately with a different set of explanatory variables; but for both, we considered the full model (detailed below).

As expected, we found in all datasets that words were more likely to be skipped if there was more information available from the linguistic prediction (Bootstrap: Dundee, $p = 0.023$; GECO, $p = 0.034$; Provo $p < 10^{-5}$) and/or the parafoveal preview (Bootstrap: Dundee, $p = 4 \times 10^{-5}$; GECO, $p < 10^{-5}$; Provo $p < 10^{-5}$). Similarly, reading times were reduced for words that were more predictable (all $p's < 3.2 \times 10^{-4}$) or more identifiable from the parafovea (all $p's < 4 \times 10^{-5}$).

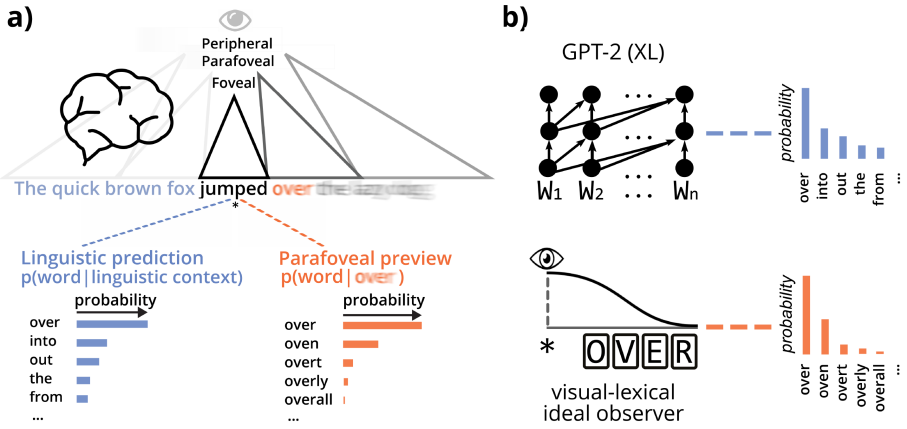

Figure 6.1. Quantifying two types of context during natural reading. a) Readers can infer the identity of the next word before fixation either by predicting it from context or by discerning it from the parafovea. Both can be cast as a probabilistic inference about the next word, either given the preceding words (prediction, blue) or given a parafoveal percept (preview, orange). **b)** To model prediction, we use GPT-2, one of the most powerful publicly available language models (Radford et al., 2019). For preview, we use an ideal observer (Duan and Bicknell, 2020) based on well-established 'Bayesian Reader' models (Bicknell and Levy, 2010; Norris, 2006, 2009). Importantly, we do not use either model as a cognitive model *per se*, but rather as a tool to quantify how much information is *in principle* available from prediction or preview on a moment-by-moment basis.

Together this confirms that our model-based approach can capture the expected effects of both prediction (Clifton et al., 2016) and preview (Schotter, Angele, and Rayner, 2012) in natural reading, while statistically controlling for other variables.

Skipping can be largely explained by non-linguistic oculomotor factors

After confirming that prediction and preview had a statistically significant influence on word skipping and reading times, we went on to assess their relative explanatory power. After confirming the effects of prediction and preview, we then further examined their relative explanatory power. That is, we asked the question how important these factors were, by examining how much variance was explained by each. To this end, we grouped the variables from the full regression model into different types of explanations, and assessed how well each type accounted for the data.

For skipping, we considered three explanations. First, a word might be skipped *purely* because it could be predicted from context – i.e. purely as a function of the amount of information conveyed by the prediction. Secondly, a word might be skipped because its identity could be gleaned from a parafoveal preview – that is, purely as a function of the informativeness of the preview. Finally, a word might be

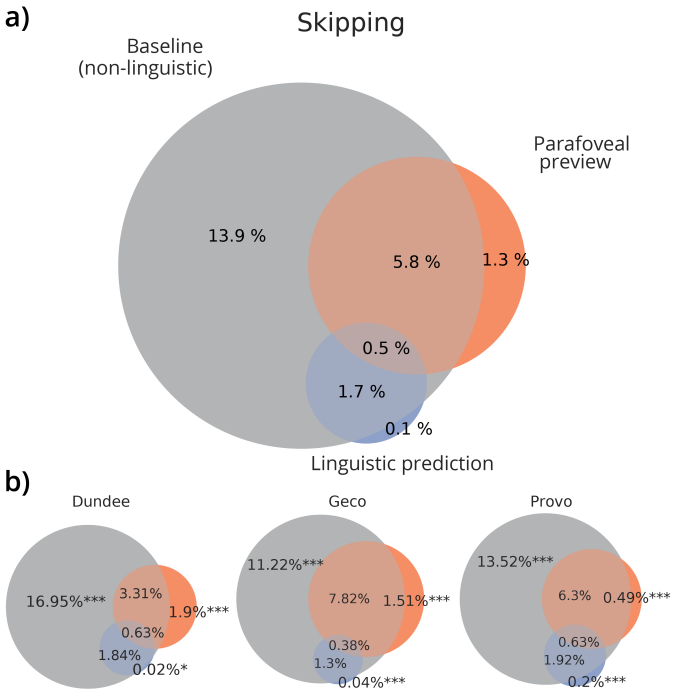

Figure 6.2. Variation in skipping explained by predictive, parafoveal and autonomous oculomotor processing. a) Proportions of cross-validated variation explained by prediction (blue), preview (orange) oculomotor baseline (grey) and their overlap; averaged across datasets (each dataset weighted equally). **b)** Variation partitions for each individual dataset, including statistical significance of variation uniquely explained by predictive, parafoveal or oculomotor processing. Stars indicate significance-levels of the cross-validated unique variation explained (bootstrap t-test against zero): $p < 0.05$ (*), $p < 0.05$ (**), $p < 0.001$ (***) For results of individual participants, and their consistency, see Figure S6.5.

skipped simply because it is so short or so close to the prior fixation location that a saccade of average length will overshoot it, irrespective of its linguistic properties – in other words, purely as a function of length and eccentricity. Note that we did not include often used lexical attributes like frequency to predict skipping, because to the extent that these affect identifiability, this is already captured by parafoveal entropy (see Fig S2; see Methods for more details on the variables used).

For each word, we thus modelled the probability of skipping either as a function of prediction, preview, or oculomotor information, or by any combination of the three. Then we partitioned the unique and shared cross-validated variation explained by each account. As can be seen in Figure 6.2, the overwhelming majority of explained variation (94 %) could be accounted for by the non-linguistic baseline. Moreover, the majority of the variation was *only* explained by the baseline, which ex-

plained 10 times more unique variation than prediction and preview combined. There was a large degree of overlap between preview and the oculomotor baseline, which is unsurprising since a word's identifiability decreases as a function of its eccentricity and length. Interestingly, there was even more overlap between the prediction and baseline model: almost all of the effect of contextual constraint could be equally well explained by the oculomotor baseline factors. Importantly, while the contribution of prediction and preview was small, it was significant both for prediction (bootstrap: Dundee, $p = 0.015$; Geco, $p = 0.0001$; Provo, $p < 10^{-5}$) and preview (all $p's < 5 \times 10^{-5}$), confirming that both factors do affect skipping. Crucially however, the vast majority of skipping that could be explained by either prediction or preview was equally well explained by the more parsimonious oculomotor model – which also explained much more data overall.

Reading times are strongly modulated by prediction and preview

For reading times (operationalised through gaze durations, so considering foveal 'reading' only), we also considered three explanatory factors. First, a word might be read faster because it was predictable from the preceding context, which we formalised via linguistic surprise. Second, a word might be read faster if it could already be partly identified from the parafoveal preview (before fixation). This informativeness of the preview was again formalised via the parafoveal preview entropy. Finally, a word might be read faster due to attributes of the word itself, such as lexical frequency. This last explanatory factor functioned as an aggregate baseline model that captured key non-contextual word attributes, both linguistic and non-linguistic (see Methods).

In all datasets, prediction (all $p's < 7.7 \times 10^{-3}$), preview (all $p's < 1.2 \times 10^{-4}$) and non-contextual woord attributes (all $p's < 1.8 \times 10^{-4}$) again all explained significant unique variation. The non-contextual baseline explained the most variance, which shows – perhaps unsurprisingly – that properties of the word itself are more important than contextual factors in determining how long a word is fixated. Critically however, compared to skipping the *unique* contribution of prediction and preview was more than three times higher (see Fig 6.3). Specifically, while prediction and preview could only uniquely account for 6% of explained word skipping variation, they uniquely accounted for more than 18 % of explained variation in reading times. Importantly, the *non-contextual* baseline used to predict reading times included both linguistic (e.g. lexical frequency) and non-linguistic information (viewing position) of the current word. When we analysed these separately, we found that the unique contribution of non-linguistic factors was small (see S6.7). This shows that contrary to skipping, variation in reading time is heavily influenced by online linguistic pro-

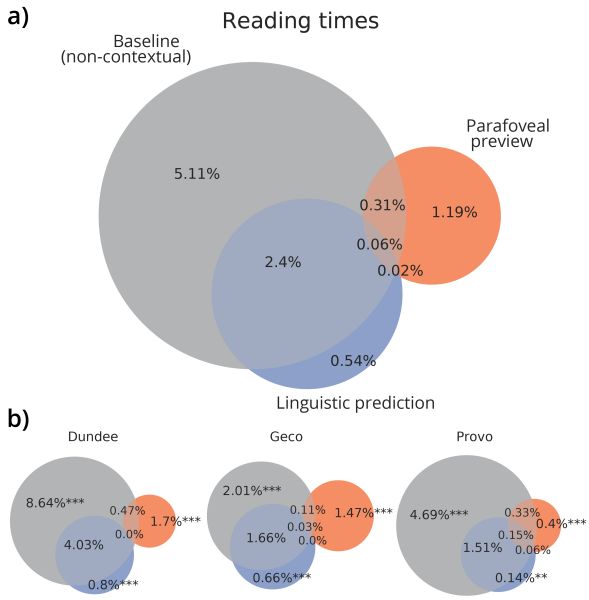

Figure 6.3. Variation in reading times explained by predictive, parafoveal and non-contextual information. **a)** Grand average of partitions of cross-validated variance in reading times (indexed by gaze durations) across datasets (each dataset weighted equally) explained by non-contextual factors (grey), parafoveal preview (orange), and linguistic prediction (blue). **b)** Variance partitions for each individual dataset, including statistical significance of the cross-validated variance explained uniquely by the predictive, parafoveal or non-contextual explanatory variables. Stars indicate significance-levels of the cross-validated unique variance explained (bootstrap t-test against zero): $p < 0.05$ (**), $p < 0.001$ (***). Note that the non-contextual model here included both lexical attributes (e.g. frequency) and oculomotor factors (relative viewing or landing position); assessing these separately reveals that reading time variation uniquely explained by oculomotor factors was small (see Fig S6.7). For results of individual participants, see Figure S6.6.

cessing.

Naturalistic prediction and preview benefit effect match experimental effect sizes

The previous result confirms that reading times (indexed via gaze durations) are highly sensitive to linguistic and parafoveal context, in line with decades of research on eye movements in reading (Rayner, 2009). But how well do our results compare exactly to established findings from the experimental literature?

To directly address this question, we simulated, for each participant the effect size of two well-established effects that would be expected to be obtained if we would conduct a well-controlled factorial experiment. Specifically, because we estimated

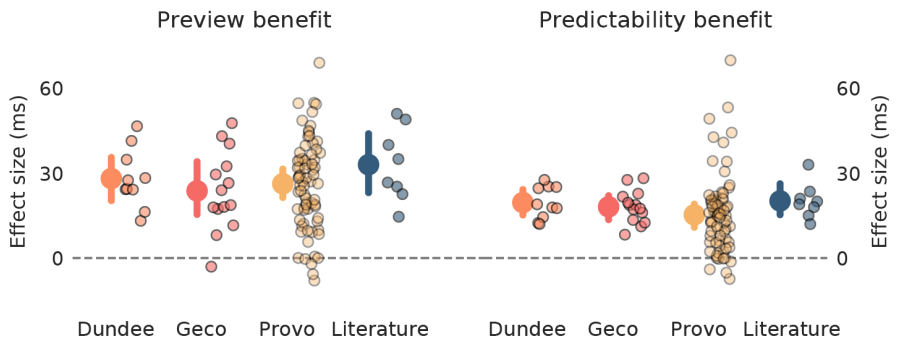

Figure 6.4. Simulated preview and predictability benefits match those reported in experimental literature. Preview (left) and predictability benefits (right) inferred from our analysis of each dataset, and observed in a sample of studies (see Table S6.1). In this analysis, preview benefit was simulated as the expected difference in gaze duration after a preview of average informativeness versus after no preview at all. Predictability benefit was defined as the difference in gaze duration for high versus low probability words; 'high' and 'low' were defined by subdividing the cloze probabilities from provo into equal thirds of 'low', 'medium' and 'high' probability (see Methods). In each plot, small dots with dark edges represent either individual subjects within one dataset or individual studies in the sample of the literature; larger dots with error bars represent the mean effect across individuals or studies, plus the bootstrapped 99% confidence interval.

how much additional information from either prediction or preview (in bits) reduced reading times (in milliseconds) we could predict reading times for words that are expected vs unexpected (predictability benefit (Rayner and Well, 1996; Staub, 2015)) or have valid vs invalid preview (i.e. preview benefit (Schotter, Angele, and Rayner, 2012)).

The simulated effects tightly corresponded to those from experimental studies (see Fig 6.4). This shows that our analysis does not strongly underfit or otherwise underestimate the effect of either prediction or preview. Moreover, it shows that the effect sizes, which are well-established in controlled designs, generalise to natural reading. This is especially interesting for the preview benefit, because it implies that this effect can be largely explained through parafoveal lexical identification, rather than visual preprocessing or interference effects (see Discussion).

No integration of prediction and preview

So far, we have treated prediction and preview as being independent. However, it might be that these processes, while using different information, are integrated – such that a word is parafoveally more identifiable when it is *also* more predictable in context. Bayesian probability theory proposes an elegant and mathematically op-

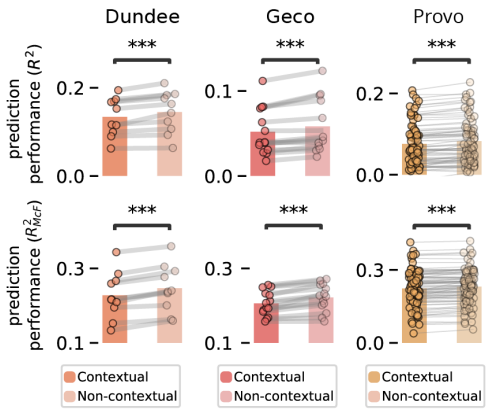

Figure 6.5. Evidence against bayesian integration of linguistic prediction and parafoveal preview. Cross-validated prediction performance of the full reading times (top) and skipping (bottom) model (including all variables), equipped with parafoveal preview information either from the contextual observer or from the non-contextual observer. Dots with connecting lines indicate participants; stars indicate significance: $p < 0.001$ (***).

timal way to integrate these sources of information: the prediction of the next word could be incorporated as a prior in perceptual inference. Such a contextual prior fits into hierarchical Bayesian models of vision (Lee and Mumford, 2003), and has been observed in speech perception, where a contextual prior guides the recognition of words from a partial sequence of phonemes (Heilbron et al., 2021a). Does such a prior also guide word recognition in reading, based on a partial parafoveal percept?

To test this, we recomputed the parafoveal identifiability of each word for each participant, but now with an ideal observer using the prediction from GPT-2 as a prior. As expected, bayesian integration enhanced perceptual inference: on average, the observer using linguistic prediction as a prior extracted more information from the preview (± 6.25 bits) than the observer not taking the prediction into account (± 4.30 bits; $T_{1.39\times10^6} = 1.35 \times 10^{11}, p \approx 0$). Interestingly however, it provided a worse fit to the human reading data. This was established by comparing two versions of the full regression model: one with parafoveal entropy from the (theoretically superior) contextual ideal observer and one from the non-contextual ideal observer. In all datasets both skipping and reading times were better explained by a model including parafoveal identifiability from the non-contextual observer (skipping: all $p's < 10^{-5}$; reading times: $p's < 10^{-5}$; see Figure 6.5).

Together, this suggests linguistic prediction and parafoveal preview are not integrated, but instead operate independently – thereby highlighting a remarkable suboptimality in reading, and potentially an intriguing difference between visual and auditory word recognition.

Discussion

Eye movements during reading are highly variable. Across three large datasets, we have assessed the relative importance of two major cognitive explanations for this variability – linguistic prediction and parafoveal preview – compared to alternative non-linguistic and non-contextual explanations. This revealed a stark dissociation between skipping and reading times. For word skipping, neither prediction nor preview were especially important, as the overwhelming majority of variation could be explained – mostly *uniquely* explained – by an oculomotor baseline model using just word length and eccentricity. For reading times, by contrast, we observed clear contributions of both prediction and preview, and effect sizes matching those obtained in tightly controlled experiments. Interestingly, preview effects were best captured by a non-contextual observer, suggesting that while readers use both linguistic prediction and preview, these do not appear to be integrated on-line. Together, the results underscore the dissociation between skipping and reading times, and show that for word skipping, the link between eye movements and cognition is less direct than commonly thought.

Our results on skipping align well with earlier findings by Drieghe and colleagues (Drieghe et al., 2004). They analysed effect sizes from studies on skipping and found a disproportionately large effect of length, compared to proxies of processing-difficulty like frequency and predictability. We significantly extend their findings by modelling skipping itself (rather than effect sizes from studies) and making a direct link to processing mechanisms. For instance, based on their analysis it was unclear how much of the length effect could be attributed to the decreasing visibility of longer words – i.e. how much of the length effect may be an identifiability effect (Drieghe et al., 2004, p. 19). We show that length and eccentricity alone explained three times as much variation as parafoveal identifiability – and that most of the variation explained by identifiability was equally well explained by length and eccentricity. This demonstrates that length and eccentricity themselves – not just to the extent they reduce parafoveal identifiability – are key drivers of skipping.

This conclusion challenges dominant, cognitive models of eye movements, which describe lexical identification as the primary driver behind skipping (Engbert and Kliegl, 2003; Engbert et al., 2005; Reichle, Rayner, and Pollatsek, 2003). However, it does not challenge the notion of predictive or parafoveal word identification itself. In fact, we believe this happens routinely – after all, most skips are not followed by regressions. Rather, our results challenge the notion that moment-to-moment decisions of whether to skip individual words are primarily determined by the identification of those words. Instead, they support a much simpler strategy, which is primarily sensitive to a word's length and eccentricity.

One such simpler strategy would be a 'blind' random walk: making saccades of some average length, plus oculomotor noise. However, we do not think this is likely, since landing positions are distributed with preferred positions with respect to word boundaries (Drieghe et al., 2004; O'Regan, 1992). Instead, we suggest an alternatie view, in which the decision of where to look next is based on an analysis of the parafovea – but at a very low level, aimed to discern mostly the next word's length and eccentricity (see also Drieghe et al., 2004; Reilly and O'Regan, 1998). This is not the whole story, since preview and prediction explain some unique skipping variation that cannot be reduced to low-level variables (or other variables (Duan and Bicknell, 2020)). Our results may thus support a hybrid account, in which most skipping decisions are made by a low-level 'autopilot', whereas in some limited cases skipping is influenced by high-level contextual information. How the brain arbitrates between these strategies is an interesting question for future research.

A distinctive feature of our approach is that we we focus on a limited number of computationally explicit functional explanations, rather than using lexical attributes as proxies for functional explanations (e.g. a word's frequency as a proxy for its identifiability). For instance, we model parafoveal identifiability using a single variable that should in principle capture all important effects such as that of frequency and orthography (see Figure S6.3 and *Methods*). A limitation of this approach is that an imperfection in the model can cause an underestimation of preview importance. However, a key advantage of using explicit modelling rather than proxies is that it can avoid confound-related misinterpretations. For instance, word frequency is strongly correlated with length, so when using frequency as a proxy for identifiability (e.g. to predict skipping), one may find apparent identifiability effects which are in fact length effects, and strongly overestimate the importance of preview (Brysbaert and Drieghe, 2003). Therefore, we have only used explanatory variables that explicitly relate to the dependent variable (*Methods*). After all, our goal was not to measure as many effects as possible, but to gain a clear picture of the importance of two cognitive explanations for eye movement variation. Based on the effect sizes for gaze duration (Fig 6.4) we do not believe that we strongly underestimate either prediction or preview, and we are optimistic the results provide the comprehensive, interpretable picture we aimed for.

When comparing Figures 6.2, 6.3 and 6.5, the numerical R^2 values of the reading times regression may seem rather small, potentially indicating a poor fit. However, our (cross-validated) R^2's for gaze durations are not lower than R^2's reported by other regression analyses of gaze durations in natural reading (e.g. Kliegl, Nuthmann, and Engbert, 2006); moreover we find effect sizes in line with the experimental literature (Fig 6.4). Therefore, we do not believe we either overfit or underfit the gaze durations. Instead, what the relatively low R^2 values indicate, we suggest, is that gaze

durations are inherently noisy, and that only a limited amount of the word-by-word variation is *systematic* variation, due to e.g. preview or frequency effects. While this is interesting in itself, it is not of primary interest in this study, which focusses on the relative importance of different *explanations*, and hence only on systematic variation. Therefore, what matters is not as much the absolute R^2 values, but rather the relative importance of different explanations – in other words, the relative size of the circles in Figures 6.2, 6.3 and S6.7, their overlap, and the explanations each circle represents. It is on this level of analysis that we find the stark dissociation – that for skipping (but not for reading times) a simple low-level heuristic can account for almost all of the explained variation – and not on the level of numerical values for variation explained.

A remarkable result is that we found preview benefits comparable to effect sizes from controlled designs, despite major methodological differences. Specifically, in controlled designs preview benefits are the difference in reading time for words with preview, versus words where the preview was masked or invalid (i.e. where a different word was previewed). As such, it seemed plausible that a significant portion of this difference may reflect interference or mismatch between preview and fixation, rather than purely the lack of parafoveal identification of the next word. Our analysis modelled the effect purely in terms of lexical identification, and yielded only slightly smaller effect sizes (Fig. 6.4). This suggests that preview benefits are largely the result of lexical identification, and that interference or visual 'preprocessing' may only play a minor role (cf Reichle, Rayner, and Pollatsek, 2003; Schotter, Angele, and Rayner, 2012).

Another notable finding is that preview was best explained by a non-contextual observer – a model which only takes word frequency (and not contextual predictability) into account. This replicates and extends the only study that explicitly compared contextual and non-contextual accounts of parafoveal preview (Duan and Bicknell, 2020). That study only analysed skipping (in the Dundee corpus); the fact that we find the same for reading times (where preview and prediction effects are much stronger) and replicate the result in other corpora, considerably strengthens the conclusion that parafoveal word recognition is not informed by linguistic context. This conclusion seems to contradict experiments finding an interaction between linguistic context and preview, which was interpreted as context constraining preview (Balota, Pollatsek, and Rayner, 1985; McClelland and O'Regan, 1981; Schotter et al., 2015; Veldre and Andrews, 2018). One explanation for this discrepancy stems from how the effect is measured. Experimental studies did not explicitly model contextual and non-contextual recognition, but looked at the effect of context on the difference in reading time after valid versus invalid preview (Schotter et al., 2015; Veldre and Andrews, 2018). This may reveal a context effect not on recognition, but at a later

stage (e.g. priming between the context, preview and fixated word). Arguably, these scenarios yield different predictions: if context affects recognition it may allow identification of otherwise unidentifiable words. However, if the interaction occurs later it might only *amplify* processing of recognisable words. Constructing a model that formally reconciles this discrepancy – and predicts the context-preview interaction using a non-contextual prior – is an interesting challenge for future work.

The lack of influence of contextual constraint on parafoveal preview might be understood through time-constraints imposed by the rapid rate of eye movements. Given that readers on average only look some 250 ms at a word in which they have to both recognise the foveal word and process the parafoveal percept, this perhaps leaves too little time to fully integrate the foveal word and the context inform parafoveal perception. Time-constraints are of course not unique to reading: word recognition based on partial information also happens in speech perception, where it also occurs under significant time-constraints. And yet in auditory word recognition, contextual effects are found (McClelland and Elman, 1986; Zwitserlood, 1989), and a formally highly similar analysis of word recognition based on partial phonemic information recently showed clear support for a contextual prior; i.e. the exact opposite of what we find here (Heilbron et al., 2021a). An alternative, more speculative explanation for the lack of context effect in reading but not speech perception is that this may reflect a difference between visual and auditory word recognition. This could be related to the fact that contrary to auditory word recognition, visual word recognition is an acquired skill and occurs throughout areas in the visual system repurposed for reading (Dehaene, 2009; Yeatman and White, 2021), where perhaps the dynamic sentence context cannot exert as much of an influence as rapidly, allowing for facilitation by lexical or orthographic context (Heilbron et al., 2020; Lupyan, 2017; Reicher, 1969; Wheeler, 1970), but not as much of sentence context.

Given that readers use both prediction and preview, why would they strongly affect reading times but hardly word skipping? To understand this dissociation, it is important to consider that they reflect different decisions, namely *where* versus *how long* to fixate, which are made at different moments. Specifically, the decision of where to fixate – and hence whether to skip the next word – is made early in saccade programming, which can take 100-150 ms (Deubel, O'Regan, and Radach, 2000; Drieghe et al., 2004; Rayner, 2009). Although the exact sequence of operations leading to a saccade remains debated, given that readers on average only look some 250 ms at a word, it is clear that skipping decisions are made under strong time constraints, especially given the lower processing rate of parafoveal information. Our results suggest that the brain meets this constraint by resorting to a computationally frugal policy, largely based on low-level characteristics such as length and eccentricity. *How long* to fixate, by contrast, mostly depends on foveal information, which is

processed more rapidly and may thus directly influence the decision to either keep dwelling and accumulate more information or initiate a saccade (and/or an attention shift).

In conclusion, we have found that two important contextual sources of information in reading, linguistic prediction and parafoveal preview, strongly drive variation in reading times, but hardly affect word skipping, which is largely based on low-level factors. Our results show that as readers, we do not always use all information available to us; and that we are, in a sense, of two minds: consulting complex inferences to decide how long to look at a word, while employing semi-mindless scanning routines to decide where to look next. It is striking that these disparate strategies operate mostly in harmony. Only occasionally they go out of step – then we notice that our eyes have moved too far and we have to look back, back to where our eyes left cognition behind.

Methods

We analysed eye-tracking data from three, big, naturalistic reading corpora, in which native English speakers read texts while eye-movement data was recorded (Cop et al., 2017; Kennedy, 2003; Luke and Christianson, 2016).

Eye-tracking data and stimulus materials

We considered the English-native portions of the Dundee, Geco and Provo corpora. The Dundee corpus comprises eye-movements from 10 native speakers from the UK (Kennedy, 2003), who read a total of 56.212 words across 20 long articles from The Independent newspaper. Secondly, the English portion of the Ghent Eye-tracking Corpus (Geco) (Cop et al., 2017) is a collection of eye movement data from 14 UK English speakers who each read Agathe Cristie's *The Mysterious Affair at Styles* in full (54.364 words per participant). Lastly, the Provo corpus (Luke and Christianson, 2018) is a collection of eye movement data from 84 US English speakers, who each read a total of 55 paragraphs (extracted from diverse sources) for a total of 2.689 words.

Language model

Contextual predictions were formalised using a language model – a model computing the probability of each word given the preceding words. Here, we used GPT-2 (XL) – currently among the best publicly released English language models. GPT-2 is a transformer-based model, that in a single pass turns a sequence of tokens (rep-

resenting either whole words or word-parts) $U = (u_1, \ldots, u_k)$ into a sequence of conditional probabilities, $(p(u_1), p(u_2|u_1), \ldots, p(u_k \mid u_1, ..., u_{k-1}))$.

Roughly, this happens in three steps: first, an embedding step encodes the sequence of symbolic tokens as a sequence of vectors, which can be seen as the first hidden state h_o. Then, a stack of n transformer blocks each apply a series of operations resulting in a new set of hidden states h_l, for each block l. Finally, a (log-)softmax layer is applied to compute (log-)probabilities over target tokens. In other words, the model can be summarised as follows:

$$h_0 = UW_e + W_p \tag{6.1}$$

$$h_l = \text{transformer\_block}\,(h_{l-1})\ \forall i \in [1, n] \tag{6.2}$$

$$P(u) = \text{softmax}\left(h_n W_e^T\right), \tag{6.3}$$

where W_e is the token embedding and W_p is the position embedding.

The key component of the transformer-block is *masked multi-headed self-attention* (Fig S6.1). This transforms a sequence of input vectors $(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k)$ into a sequence of output vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k)$. Fundamentally, each output vector $\mathbf{y}_i$ is simply a weighted average of the input vectors: $\mathbf{y}_i = \sum_{j=1}^{k} w_{ij}\mathbf{x}_j$. Critically, the weight $w_{i,j}$ is not a parameter, but is *derived* from a dot product between the input vectors $\mathbf{x}_i^T\mathbf{x}_j$, passed through a softmax and scaled by a constant determined by the dimensionality d_k: $w_{ij} = \left(\exp \mathbf{x}_i^T\mathbf{x}_j / \sum_{j=1}^{k} \exp \mathbf{x}_i^T\mathbf{x}_j\right)\frac{1}{\sqrt{d_k}}$. Because this is done for each position, each input vector $\mathbf{x}_i$ is used in three ways: first, to derive the weights for its own output, $\mathbf{y}_i$ (as the *query*); second, to derive the weight for any other output $\mathbf{y}_j$ (as the *key*); finally, in it used in the weighted sum (as the *value*). Different linear transformations are applied to the vectors in each cases, resulting in Query, Key and Value matrices (Q, K, V). Putting this all together, we obtain:

$$\text{self\_attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{6.4}$$

To be used as a language model, two elements are added. First, to make the operation position-sensitive, a position embedding W_p is added during the embedding step – see Equation (6.1). Second, to enforce that the model only uses information from the past, attention from future vectors is masked out. To give the model more flexibility, each transformer block contains multiple instances ('heads') of the self-attention mechanisms from Equation (6.4).

In total, GPT-2 (XL) contains $n = 48$ blocks, with 12 heads each; a dimensionality of $d = 1600$ and a context window of $k = 1024$, yielding a total 1.5×10^9 parameters. We used the PyTorch implementation of GPT-2 provided by HuggingFace's

Transformers package (Wolf et al., 2020). For words spanning multiple tokens, we computed their joint probability.

Ideal observer

To compute parafoveal identifiability, we implemented an ideal observer based on the formalism by Duan & Bicknell (Duan and Bicknell, 2020). This model formalises parafoveal word identification using Bayesian inference and builds on previous well-established 'Bayesian Reader' models (Bicknell and Levy, 2010; Norris, 2006, 2009). It computes the probability of the next word given a noisy percept by combining a prior over possible words with a likelihood of the noisy percept, given a word identity:

$$p(w \mid \mathcal{I}) \propto p(w)p(\mathcal{I}|w), \tag{6.5}$$

where $\mathcal{I}$ represents the noisy visual input, and w represents a word identity. We considered two priors (see Fig 6.5): a non-contextual prior (the the overall probability of words in English based on their frequency in Subtlex (Brysbaert and New, 2009), and a contextual prior based on GPT2 (see below). Below we describe how visual information is represented and perceptual inference is performed. For a graphical schematic of the model, see Fig S6.2; for some distinctive simulations showing how the model captures key effects of linguistic and visual characteristics on word recognition, see Fig S6.3.

Sampling visual information

Like in other Bayesian Readers (Bicknell and Levy, 2010; Norris, 2006, 2009), noisy visual input is accumulated by sampling from a multivariate Gaussian which is centered on a one-hot 'true' letter vector – here represented in an uncased 26-dimensional encoding – with a diagional covariance matrix $\Sigma(\varepsilon) = \lambda(\epsilon)^{-1/2}I$. The shape of Σ is thus scaled by the sensory quality $\lambda(\varepsilon)$ for a letter at eccentricity ε. Sensory quality is computed as a function of the perceptual span: this uses using a Gaussian integral based follows the perceptual span or processing rate function from the SWIFT model (Engbert et al., 2005). Specifically, for a letter at eccentricity ε, λ is given by the integral within the bounding box of the letter:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x, \tag{6.6}$$

which, following (Bicknell and Levy, 2010; Duan and Bicknell, 2020), is scaled by a scaling factor Λ. Unlike SWIFT, the Gaussian in Equation 6.6 is symmetric, since we only perform inference on information about the next word. By using one-hot encoding and a diagonal covariance matrix, the ideal observer ignores similarity structure

between letters. This is clearly a simplification, but one with significant computational benefits; moreover, it is a simplification shared by all Bayesian Reader-like models (Bicknell and Levy, 2010; Duan and Bicknell, 2020; Norris, 2006), which can nonetheless capture many important aspects of visual word recognition and reading. To determine parameters Λ and σ, we performed a grid search on a subset of Dundee and Geco (see Fig S6.4), resulting in $\Lambda = 1$ and $\sigma = 3$. Note that this σ value is close to the average σ value of SWIFT and (3.075) and corresponds well to prior literature on the size of the perceptual span (±15 characters; Bicknell and Levy, 2010; Engbert et al., 2005; Schotter, Angele, and Rayner, 2012).

Perceptual inference

Inference is performed over the full vocabulary. This is represented as a matrix which can be seen as a stack of word vectors, $\mathbf{y_1}$, $\mathbf{y_2}$, ..., $\mathbf{y_v}$, obtained by concatenating the letter vectors. The vocabulary is thus a $V \times d$ matrix, with V the number of words in the vocabulary and d the dimensionality of the word vectors (determined by the length of the longest word: $d = 26 \times l_{max}$).

To perform inference, we use the belief-updating scheme from (Duan and Bicknell, 2020), in which the posterior at sample t is expressed as a $(V-1)$ dimensional log-odds vector $\mathbf{x^{(t)}}$, in which each entry $\mathbf{x_i^{(t)}}$ represents the log-odds of $\mathbf{y_i}$ relative to the final word $\mathbf{y_v}$. In this formulation, the initial value of $\mathbf{x}$ is thus simply the prior log odds, $\mathbf{x_i^{(0)}} = \log p(w_i) - \log p(w_v)$, and updating is done by summing prior log-odds and the log-odds likelihood. This procedure is repeated for T samples, each time taking the posterior of the previous timestep as the prior in the current timestep. Note that using log odds in this way avoids renormalization:

$$\begin{aligned}
\mathbf{x}_i^{(t)} &= \log \frac{p\left(w_i \mid \mathcal{I}^{(0,\ldots,t)}\right)}{p\left(w_v \mid \mathcal{I}^{(0,\ldots,t)}\right)} \\
&= \log \frac{p\left(w_i \mid \mathcal{I}^{(0,\ldots,t-1)}\right) p\left(\mathcal{I}^{(t)} \mid w_i\right)}{p\left(w_v \mid \mathcal{I}^{(0,\ldots,t-1)}\right) p\left(\mathcal{I}^{(t)} \mid w_v\right)} \\
&= \log \frac{p\left(w_i \mid \mathcal{I}^{(0,\ldots,t-1)}\right)}{p\left(w_v \mid \mathcal{I}^{(0,\ldots,t-1)}\right)} + \log \frac{p\left(\mathcal{I}^{(t)} \mid w_i\right)}{p\left(\mathcal{I}^{(t)} \mid w_v\right)} \\
&= \mathbf{x}_i^{(t-1)} + \Delta \mathbf{x}_i^{(t)}.
\end{aligned} \tag{6.7}$$

In other words, as visual sample $\mathcal{I}^{(t)}$ comes in, beliefs are updated by summing the prior log odds $\mathbf{x}^{(t-1)}$ and the log-odds likelihood of the new information $\mathbf{x}^{(t)}$.

For a given word w_i, the log-odds likelihood of each new sample is the difference of two multivariate Gaussian log likelihoods, one centred on $\mathbf{y}_i$ and one on the last

vector $\mathbf{y}_v$. This can be formulated as a linear transformation of $\mathcal{I}$:

$$\begin{aligned}\Delta\mathbf{x}_i &= \log p\left(\mathcal{I} \mid w_i\right) - \log p\left(\mathcal{I} \mid w_v\right) \\ &= \log p\left(\mathcal{I} \mid \mathcal{N}\left(\mathbf{y}_i, \Sigma\right)\right) - \log p\left(\mathcal{I} \mid \mathcal{N}\left(\mathbf{y}_v, \Sigma\right)\right) \\ &= \left[-\frac{1}{2}\left(\mathcal{I} - \mathbf{y}_i\right)^T \Sigma^{-1}\left(\mathcal{I} - \mathbf{y}_i\right)\right] - \quad \left[-\frac{1}{2}\left(\mathcal{I} - \mathbf{y}_v\right)^T \Sigma^{-1}\left(\mathcal{I} - \mathbf{y}_v\right)\right] \\ &= \frac{\mathbf{y}_v^T \Sigma^{-1} \mathbf{y}_v - \mathbf{y}_i^T \Sigma^{-1} \mathbf{y}_i}{2} + \left(\mathbf{y}_i - \mathbf{y}_v\right)^T \Sigma^{-1} \mathcal{I},\end{aligned} \tag{6.8}$$

which implies that updating can be implemented by sampling from a multivariate normal. To perform inference on a given word, we performed this sampling scheme until convergence (using $T = 50$), and then transformed the posterior log-odds into the log posterior, from which we computed the Shannon entropy as a metric of parafoveal identifiability.

To compute the parafoveal entropy for each word in the corpus, we make the simplifying assumption that parafoveal preview only occurs during the last fixation prior to a saccade, thus computing the entropy as a function of the word itself and its distance to the last fixation location within the previously fixated word (which is not always the previous word). Because this distance is different for each participant, it was computed separately for each word, for each participant. Moreover, because the inference scheme is based on sampling, we repeated it 3 times, and averaged these to compute the posterior entropy of the word. The amount of information obtained from the preview is then simply the difference between prior and posterior entropy.

The ideal observer was implemented in custom Python code, which can be found in the data sharing collection (see below).

Contextual vs non-contextual prior

We considered two observers: one with a non-contextual prior capturing the overall probability of a word in a language, and with a contextual prior, capturing the contextual probability of a word in a specific context. For the non-contextual prior, we simply used lexical frequencies from which we computed the (log)-odds prior used in equation (6.7). For the contextual prior, we derived the contextual prior from log-probabilities from GPT-2. This effectively involves constructing a new Bayesian model for each word, for each participant, in each dataset. To simplify this process, we did not take the full predicted distribution of GPT-2, but only the 'nucleus' of the top k predicted words with a cumulative probability of 0.95, and truncated the (less reliable) tail of the distribution. Further, we simply assumed that the rest of the tail was 'flat' and had a uniform probability. Since the prior odds can be derived from relative frequencies, we can think of the probabilities in the flat tail as having

a 'pseudocount' of 1. If we similarly express the prior probabilities in the nucleus as implied 'pseudofrequencies', the cumulative implied nucleus frequency is then complementary to the the length of the tail, which is simply the difference between the vocabulary size and nucleus size ($V - k$). As such, for word i in the text, we can express nucleus as implied frequencies as follows:

$$\text{freqs}_{\psi} = P_{tr}(w^{(i)}|\text{context})\frac{V-k}{1-\sum_{j=1}^{k} P(w_j^{(i)}|\text{context})}, \tag{6.9}$$

where $P_{tr}(w^{(i)}|\text{context})$ is the trunctated lexical prediction, and $P(w_j^{(i)}|\text{context})$ is predicted probability that word i in the text is word j in the sorted vocabulary. Note that using this flat tail not only simplifies the computation, but also deals with the fact that the vocabulary of GPT-2 is smaller than of the ideal observer – using this tail we can still use the full vocabulary (e.g. to capture orthographic uniqueness effects), while using 95% of the density from GPT-2.

Data selection

In all our analyses, we focus strictly on first-pass reading, analysing only those fixations or skips when none of the subsequent words have been fixated before. We extensively preprocessed the corpora so that we could include as many words as possible. However, we had to impose some additional restrictions. Specifically we did not include words if they a) contained non-alphabetic characters; b) if they were adjacent to blinks; c) if the distance to the prior fixation location was more than 24 characters (±8°); moreover, for the gaze duration we excluded d) words with implausibly short ($< 70ms$) or long ($> 900ms$) gaze durations. Criterion c) was chosen because some participants occasionally skipped long sequences of words, up to entire lines or more. Such 'skipping' – indicated by saccades much larger than the the perceptual span – is clearly different from the skipping of words during normal reading, and was therefore excluded. Note that these criteria are comparatively mild (cf. Duan and Bicknell, 2020; Smith and Levy, 2013), and leave approximately 1.1 million observations for the skipping analysis, and 593.000 reading times observations.

Regression models: skipping

Skipping was modelled via logistic regression in scikit-learn (Pedregosa et al., 2011), with three sets of explanatory variables (or 'models') each formalising a different explanation for why a word might be skipped.

First, a word might be skipped because it could be confidently predicted from context. We formalise this via *linguistic entropy*, quantifying the information conveyed by the prediction from GPT-2. We used entropy, not (log) probability, because using

the next word's probability directly would presuppose that the word is identified, undermining the dissociation of prediction and preview. By contrast, prior entropy specifically probes the information available from prediction only.

Secondly, a word might be skipped because it could be identified from a parafoveal preview. This was formalised via parafoveal entropy, which quantifies the parafoveal preview uncertainty (or, inversely, the amount of information conveyed by the preview). This is a complex function integrating low-level visual (e.g. decreasing visibility as a function of eccentricity) and higher-level information (e.g. frequency or orthographic effects) and their interaction (see Fig S6.3). Here, too we did not use lexical features (e.g. frequency) of the next word to model skipping directly, as this presupposes that the word is identified; and to the extent that these factors are expected to influence identifiability, this is already captured by the parafoveal entropy (Fig S6.3).

Finally, a word might be skipped simply because it is too short and/or too close to the prior fixation location, such that a fixation of average length would overshoot the word. This autonomous oculomotor account was formalised by modelling skipping probability purely as a function of a word's length and its distance to the previous fixation location.

Note that these explanations are not mutually exclusive, so we also evaluated their combinations (see below).

Regression models: reading time

As an index of reading time, we analysed first-pass *gaze duration*, the sum of a word's first-pass fixation durations. We analyse gaze durations as they arguably most comprehensively reflect how long a word is looked at, and are the focus of similar model-based analyses of contextual effects in reading (Goodkind and Bicknell, 2018; Smith and Levy, 2013). For reading times, we used linear regression, and again considered three sets of explanatory variables, each formalising a different kind of explanation.

First, a word may be read more slowly because it is unexpected in context. We formalised this using surprisal $-\log(p)$, a metric of a word's unexpectedness – or how much information is conveyed by a word's identity in light of a prior expectation about the identity. To capture spillover (R; regpaper; smith) we included not just the surprisal of the current word, but also that of the previous two words.

Secondly, a word might be read more slowly because it was difficult to discern from the parafoveal preview. This was formalised using the parafoveal entropy (see above).

Finally, a word might be read more slowly because of non-contextual factors of the word itself. This is an aggregate baseline explanation, aimed to capture all rele-

vant non-contextual word attributes, which we contrast to the two major contextual sources of information about a word identity that might affect reading times (prediction and preview). We included word class, length, log-frequency, and the relative landing position (quantified as the distance to word centre in characters. For log-frequency we used the UK or US version of SUBTLEX depending on the corpus and included the log-frequency of the past two words to capture spillover effects.

Note that, while for skipping, we used a *non-linguistic* baseline, for reading times we use a *non-contextual* baseline. This is because for skipping the most interesting contrast is between the role of non-linguistic oculomotor control vs next-word identification (either through prediction or preview). For reading times, by contrast, the most interesting comparison is between properties of the word itself versus contextual cues, as a purely non-linguistic account for gaze duration variation seemed less plausible (indeed, see Fig S6.7).

Model evaluation

We compared the ability of each model to account for the variation in the data by probing prediction performance in a 10-fold cross-validation scheme, in which we quantified how much of the observed variation in skipping rates and gaze durations could be explained.

For reading times, we did this using the coefficient of determination, defined via the ratio of residual and total sum of squares: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$. The ratio $\frac{SS_{res}}{SS_{tot}}$ relates the error of the model (SS_{res}) to the error of a 'null' model predicting just the mean (SS_{tot}), and gives the variance explained. For skipping, we use a tightly related metric, the McFadden R^2. Like the R^2 it is computed by comparing the error of the model to the error of a null model with only an intercept: $R^2_{McF} = 1 - \frac{L_M}{L_{null}}$, where L indicates the loss.

While R^2 and R^2_{McF} are not identical, they are formally tightly related – critically, both are zero when the prediction is constant (no variation explained) and go towards one proportionally as the error decreases to zero (i.e. towards all variation explained). Note that in a cross-validated setting, both metrics can become negative when prediction of the model is worse than the prediction of a constant null-model.

Variation partitioning

To assess relative explanatory power, we used variation partitioning to estimate how much of the explained variation could be attributed to each set of explanatory variables. This is also known as *variance* partitioning, as it is originally based on partitioning sums of squares in regression analysis; here we use the more general term 'variation' following (Legendre, 2008).

Variation partitioning builds on the insight that when two (groups of) explanatory variables (A and B) both explain some variation in the data y, and A and B are independent, then variation explained by combining A and B will be approximately additive. By contrast, when A and B are fully redundant – e.g. when B only has an *apparent* effect on y through its correlation with A – then a model combining A and B will not explain more than the two alone.

Following (Heer et al., 2017), we generalise this logic to three (groups of) explanatory variables, by testing each individually and all combinations, and use set theory notation and graphical representation for its simplicity and clarity. For three groups of explanatory variables (A, B and C), we first evaluate each separately and all combinations, resulting in 7 models:

$$A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C.$$

From these 7 models we obtain 7 'empirical' scores (expressing variation explained), from which we derive the 7 'theoretical' variation partitions: 4 overlap partitions and 3 unique partitions. The first overlap partition is the variation explained by all models, which we can derive as:

$$A \cap B \cap C = A \cup B \cup C + A + B + C - A \cup B - A \cup C - B \cup C. \tag{6.10}$$

The next three overlap partitions contain all pairwise intersections of models that did not include the other model:

$$\begin{aligned} (A \cap B) \setminus C &= A + B - A \cup B - A \cap B \cap C \\ (A \cap C) \setminus B &= A + C - A \cup C - A \cap B \cap C \\ (B \cap C) \setminus A &= B + C - B \cup C - A \cap B \cap C. \end{aligned} \tag{6.11}$$

The last three partitions are those explained exclusively by each model. This is the relative complement: the partition unique to A is the relative complement of BC: BC^{RC}. For simplicity we also use a star notation, indicating the unique partition of A as A^*. These are derived as follows:

$$\begin{aligned} A^* &= BC^{RC} = A \cup B \cup C - B \cup C \\ B^* &= AC^{RC} = A \cup B \cup C - A \cup C \\ C^* &= AB^{RC} = A \cup B \cup C - A \cup B. \end{aligned} \tag{6.12}$$

Note that, in the cross-validated setting, the results can become paradoxical and depart from what is possible in classical statistical theory, such as partitioning sums of squares. For instance, due to over-fitting, a model that combines multiple EVs could explain *less* variance than all of the EVs alone, in which case some partitions

would become negative. However, following (Heer et al., 2017), we believe that the advantages of using cross-validation outweigh the risk of potentially paradoxical results in some subjects. Partitioning was carried out for each subject, allowing to statistically assess whether the additional variation explained by a given model was significant. On average, none of the partitions were paradoxical.

Simulating effect sizes

Preview benefits were simulated as the expected difference in gaze duration after a preview of average informativeness versus after no preview at all. This this best corresponds to an experiment in which the preceding preview was masked (e.g. XXXX) rather than invalid (see Discussion). To compute this we compared the took the difference in parafoveal entropy between an average preview and the prior entropy. Because we standardised our explanatory variables, this was transformed to subject-specific z-scores and then multiplied by the regression weights to obtain an expected effect size.

For the predictability benefit, we computed the expected difference in gaze duration between 'high' and 'low' probability words. 'High' and 'low' was empirically defined based on the human-normed cloze probabilities in provo, which we divided into thirds using percentiles. The resulting cutoff points (low < 0.02; high >0.25) were log-transformed, applied to the surprisal values from GPT-2, and multiplied by the weights to predict effect sizes. Note that these definitions of 'low' and 'high' may appear low compared to those in literature – however, most studies collect cloze only for specific 'target' words in relatively predictable contexts, which biases the definition of 'low' vs 'high' probability. By contrast, we analysed cloze probabilities for *all* words, yielding these values.

Statistical testing

Statistical testing was performed across participants within each dataset. Because two of the three corpora had a low number of participants (10 and 14 respectively) we used non-parametric bootstrap t-tests, by creating resampling a null-distribution with zero mean counting how likely a t-value at least as extreme as the true t-value was to occur. Each test used at least 10^4 bootstraps; p-values were computed without assuming symmetry (equal-tail bootstrap).

Data and code availability

Data and code to reproduce all results will be made public at the Donders Data Repository. Unfortunately we cannot share the original texts of the Dundee Corpus because

of copyright restrictions on the newspaper articles. Instead we provide a "scrubbed" version of Dundee without the copyrighted material.

Contributions

Conceptualisation: MH. Data wrangling and preprocessing: JvH. Formal analysis: MH, JvH. Statistical analysis and visualisations: MH, JvH. Supervision: FPdL, PH. Initial draft: MH. Final draft: MH, JvH, PH, FPdL.

Acknowledgements

We thank Maria Barrett, Yunyan Duan, and Benedikt Ehinger for useful input and inspiring discussions during various stages of this project. This work was supported by The Netherlands Organisation for Scientific Research (NWO Research Talent grant to M.H.; NWO Vidi grant to F.P.d.L.; Gravitation Program Grant Language in Interaction no. 024.001.006 to P.H.) and the European Union Horizon 2020 Program (ERC Starting Grant 678286 to F.P.d.L).

Supplementary materials

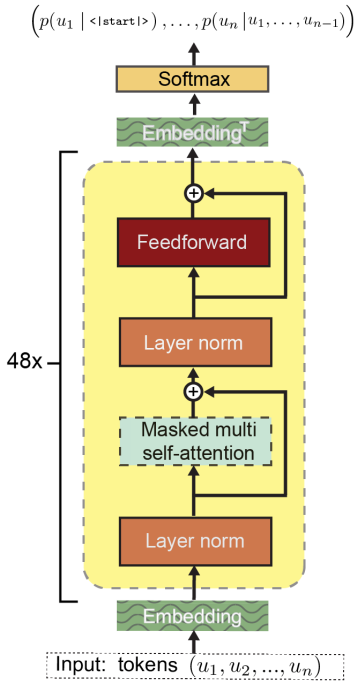

Figure S6.1. GPT-2 Architecture. Note that this panel is based on the original GPT schematic, with some operations modified and re-arranged to reflect the slightly different architecture of GPT-2. The most important and distinctive step of each transformer block is masked multi-headed self-attention (see Methods). Not visualised here is the initial tokenisation, mapping a sequence of characters into a sequence of tokens.

Table S6.1. Literature sample for effect size ranges

Effect type	Publication	Effect size
preview benefit	Inhoff, A. W. (1989). Lexical access during eye fixations in reading: Are word access codes used to integrate lexical information across interword fixations?. Journal of Memory and Language, 28(4), 444-461.	51

Continued on next page

Table S6.1 – *Continued from previous page*

Effect type	**Publication**	**Effect size**
preview benefit	Veldre, A., & Andrews, S. (2018). Parafoveal preview effects depend on both preview plausibility and target predictability. Lexical access during eye fixations in reading: Quarterly Journal of Experimental Psychology, 71(1), 64-74.	49
preview benefit	Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. Perception & psychophysics, 40(6), 431-439.	40
preview benefit	McDonald, S. A. (2006). Parafoveal preview benefit in reading is only obtained from the saccade goal. Vision Research, 46(26), 4416-4424.	35
preview benefit	Williams, C. C., Perea, M., Pollatsek, A., & Rayner, K. (2006). Previewing the neighborhood: The role of orthographic neighbors as parafoveal previews in reading. Journal of Experimental Psychology: Human Perception and Performance, 32(4), 1072.	26.7
preview benefit	Kennison, S. M., & Clifton, C. (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: Implications for eye movement control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(1), 68.	25.25
preview benefit	Blanchard, Harry E., Alexander Pollatsek, and Keith Rayner. "The acquisition of parafoveal word information in reading." Perception & Psychophysics 46.1 (1989): 85-94.	22.6
preview benefit	Schroyens, W., Vitu, F., Brysbaert, M., & d'Ydewalle, G. (1999). Eye movement control during reading: Foveal load and parafoveal processing. The Quarterly Journal of Experimental Psychology Section A, 52(4), 1021-1046.	14.6
prediction benefit	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.	33
prediction benefit	Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. Psychonomic Bulletin & Review, 3(4), 504-509.	20

Continued on next page

Table S6.1 – *Continued from previous page*

Effect type	**Publication**	**Effect size**
prediction benefit	RJ. Altarriba, J. Kroll, A. Sholl, K. Rayner. (1996) The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times Memory & Cognition, 24 (1996), pp. 477-492.	21
prediction benefit	Ashby, J., Rayner, K., & Clifton Jr, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. The Quarterly Journal of Experimental Psychology Section A, 58(6), 1065-1086.	23.5
prediction benefit	Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: implications for the EZ Reader model. Journal of Experimental Psychology: Human Perception and Performance, 30(4), 72	19
prediction benefit	Rayner, K., Binder, K. S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. Vision Research, 41(7), 943-954.	15
prediction benefit	Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. Vision Research, 41(7), 943-954.	18
prediction benefit	Hand, C. J., Miellet, S., O'Donnell, P. J., & Sereno, S. C. (2010). The frequency-predictability interaction in reading: It depends where you're coming from. Journal of Experimental Psychology: Human Perception and Performance, 36(5), 1294–1313.	12

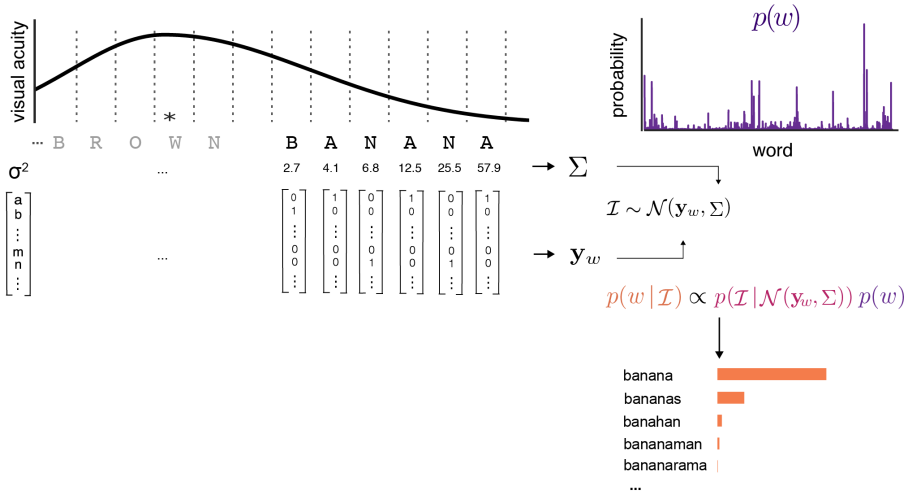

Figure S6.2. Encoding and inference scheme of the ideal observer analysis. A word at a given eccentricity is converted into a noisy visual percept, after which a posterior probability of the identity of the word given the noisy percept was computed using Bayesian inference. The uncertainty of this posterior (expressed in terms of Shannon entropy) was then used to quantify the expected uncertainty in the parafoveal percept – or, inversely, a word's *parafoveal identifiability*.

In this scheme, words are represented as a concatenation of one-hot encoded letter vectors. Visual information ($\mathcal{I}$) is sampled from a multivariate Gaussian centred on the word vector $\mathbf{y}_w$ with a diagonal covariance matrix Σ, the values of which (σ^2) are inversely related to the integral under the visual acuity function around each letter. The posterior is then computed by comining the likelihood of the visual information $\mathcal{I}$ given a particular word, with a prior probability of that word $p(w)$ (e.g. derived from lexical frequency). This computation was performed using a log-odds formulation that exploits the proportionality in Bayes' rule to perform belief-updating without renormalisation (see Methods).

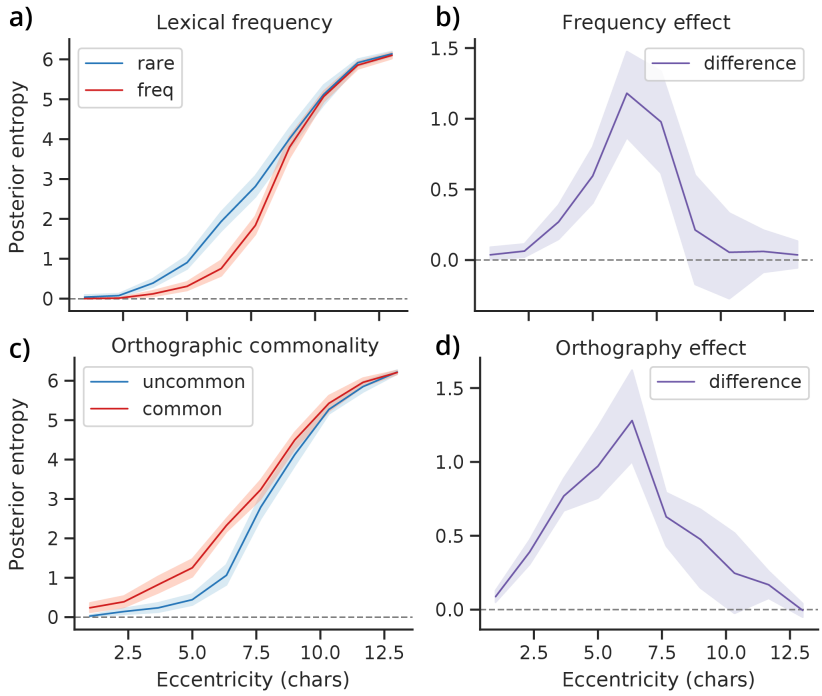

Figure S6.3. Modulation of parafoveal identifiability by visual and linguistic features, and their interaction. The parafoveal entropy for a given word (Fig S6.2) is a complex function that integrates linguistic and visual characteristics, and which can account for various known effects, such as the effect of lexical frequency and orthographic neighbourhood on visual word recognition. To illustrate this, we simulated some characteristic effects of eccentricity, frequency (**a,b**) and orthographic distinctiveness (**c,d**).
For frequency (**a**), we randomly sampled 20 'rare' and 'frequent' 5-letter words (based on a quartile split), and computed the parafoveal identifiability (quantified via posterior entropy) at increasing eccentricities. As can be seen, the percept becomes uncertain at increasing eccentricities more quickly for low-frequency words, showing that lexical frequency boosts parafoveal identifiability.
For orthography (**c**), we similarly sampled 20 7-letter words that were classified as orthographically common or uncommon based on the first three letters. Here, commonality was again defined using a quartile split but now on the number of alternative words starting with the same three letters. For instance, the letters 'awk' in the word 'awkward' are highly uncommon and allow to identify the entire word with high confidence based on just those three letters. As can be seen, the model predicts that orthographic uniqueness boosts parafoveal identifiability – as observed in experiments (see Schotter, Angele, and Rayner, 2012).
Notably, when we consider the difference between the two classes of words (**b,d**), an inverted U shape is apparent: the effects are strongest at intermediate visibility. This demonstrates the well-established fact that the effects of prior (linguistic) knowledge is strongest at intermediate levels of perceptual uncertainty (see Norris, 2006 for discussion). (Note that, while both the orthography and frequency effects are effects of "prior linguistic knowledge", only the frequency effect is technically an effect of the *prior*, since the orthography effect is driven by the generative model.) In all plots, thick lines represent the mean entropy across words; shaded regions indicate bootstrapped 95% confidence intervals.

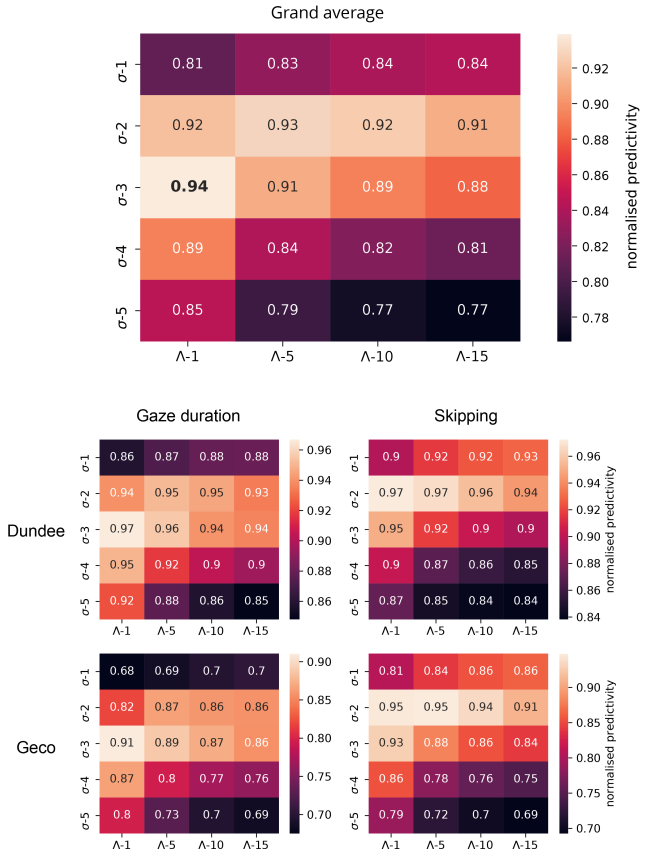

Figure S6.4. Grid search to establish ideal observer parameters. Grid search result grand average (top) and individual results for different corpora and analyses (bottom). To decide on the values for σ and Λ, a grid search was performed on a random subset of 25% of the Dundee and Geco corpus; we did not apply it to PROVO because there was not enough data per participant. In both skipping and reading times, we performed a 10-fold cross-validation with the full model, using parafoveal entropy as computed with different visual acuity parameters σ and Λ (Equation 6.6). To avoid biasing the contextual vs non-contextual model comparison (Figure 6.5), we used both the contextual and non-contextual prior and averaged the results to obtain the results for each analysis in each corpus. To ensure that different analyses and corpora are weighted equally in the grand average, the prediction scores (R^2 or R^2_{McF}) were normalised by dividing the prediction score of each parameter combination by the highest score (i.e. score of the best parameter combination) for each subject, for each analysis. This resulted in $\sigma = 3$ and $\Lambda = 1$, which we have used in all analyses. Note that σ determines the perceptual span (see Figure S6.2) and that $\sigma = 3$ corresponds well to what is known about the size of the perceptual span and is close to default parameters in other models (see Methods).

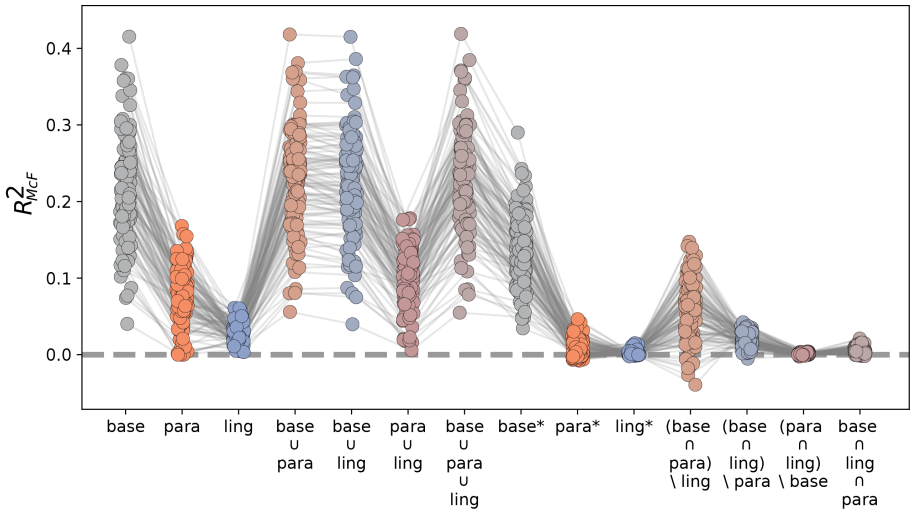

Figure S6.5. Skipping variation partitioning for all participants. Explained cross-validated variation partition for skipping (see Fig 6.2) of each partition, for each participant, for the skipping analysis. Models for the baseline, parafoveal preview and linguistic prediction are indicated by 'base', 'para', and 'ling', respectively. Unions are indicated by ∪, intersections by ∩; for the relative complement we use the asterisk-notation: e.g. 'para*' indicates variation explained uniquely by parafoveal preview. Note that due to cross-validation, the amount of variation explained can become negative in some partitions for individual participants (see Methods).

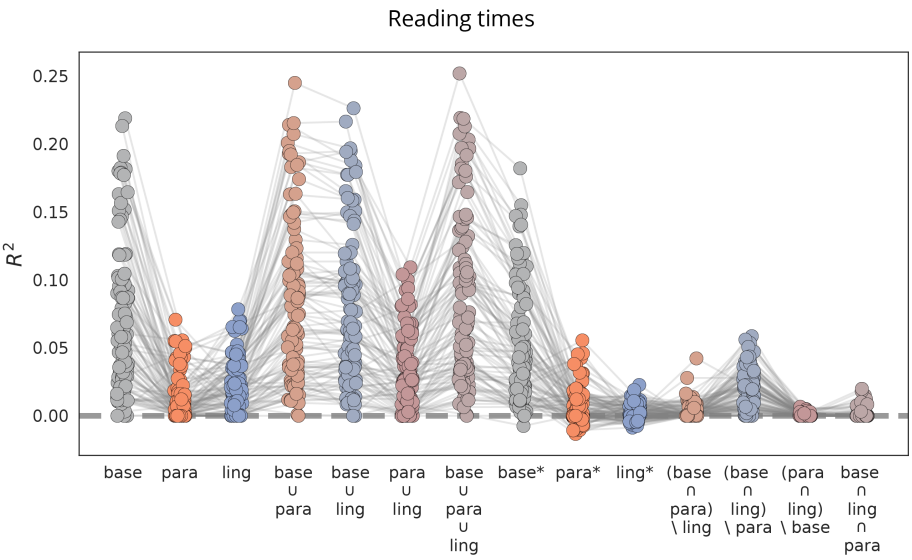

Figure S6.6. Reading times variance partitioning. Explained cross-validated variation partition for skipping (see Fig 6.3) of each partition, for each participant, for the skipping analysis. Models for the baseline, parafoveal preview and linguistic prediction are indicated by 'base', 'para', and 'ling', respectively. Unions are indicated by ∪, intersections by ∩; for the relative complement we use the asterisk-notation: e.g. 'para*' indicates variation explained uniquely by parafoveal preview (see Methods). Note that due to cross-validation, the amount of variation explained can become negative in individual participants (see Methods).

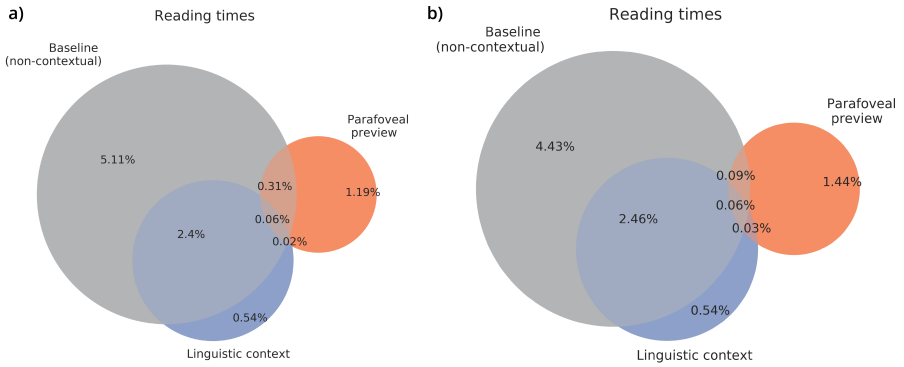

Figure S6.7. Reading times variance partitioning with and without non-linguistic factors Same as in Fig 6.3, but comparing the baseline with (**a**)) or without (**b**)) the primary non-linguistic explanatory factor for reading time variation – viewing position (O'Regan, 1992). Including the viewing position adds 0.7% additional variance explained. This demonstrates while that viewing position affect reading times, the amount of variance uniquely explained by non-linguistic factors is much lower for reading times than for skipping.

Chapter 7

Discussion

The aim of this thesis was to evaluate the predictive processing framework using language as a testbed – and to use predictive processing to understand the role of prediction in language. This was done in five studies on a variety of topics that collectively addressed two key questions about the role of prediction in language processing. *When* does language processing invoke predictions (i.e. under which conditions), and *what* is being predicted (i.e. at which processing level is prediction taking place)?

The predictive processing framework proposes two rather bold answers to these questions. In response to the *when* question, it suggests that language *always* involves prediction: that prediction is an inherent part of language processing and thus not restricted to specific conditions. In response to the *what* question, the framework suggests that prediction occurs at *all* levels of analysis, from the abstract meaning of words in the context of a story, all the way down to the exact shapes and sounds of letters and phonemes within individual words. Importantly, predictions at different levels constrain each other, such that even our abstract expectations about upcoming words in a story would be able to inform the processing of the sounds making up individual words.

Combining computational modelling with fMRI, EEG, MEG and eye tracking, I found broad support for the proposed answers to both questions. As to the *when* question, I found prediction in a wide range of conditions, from participants attentively viewing single words (**Chapter 2**) to simply listening to audiobooks (**Chapters 3-5**) or reading an entire novel (**Chapter 6**). This apparent *ubiquity* of prediction was more explicitly tested in **Chapter 4**. There I found that during story listening, predictability modulations of the brain response (like that of the well known N400 component) are not limited to specific content words in constraining contexts (in which they have been historically studied), but seem to occur for *all* words, exhibiting a sensitivity to very subtle differences in predictability between words that may appear to be equally unpredictable. Together, the results confirm that the brain automatically and inescapably predicts upcoming language, even when passively listening to something as complex and seemingly unpredictable as an audiobook.

As to the *what* question, I found predictions at all levels of processing, from the abstract meaning of words to the exact shapes of individual letters. In **Chapter 2** I found that lexical and orthographic knowledge can enhance the processing of the shapes of expected letters at the earliest visual areas of cortex. In **Chapter 4**, I found that during natural comprehension, the brain is engaged in prediction across many levels of abstraction, revealed by dissociable signatures of syntactic, phonemic and semantic predictions. In that same chapter, I also found that the brain integrates predictions at different levels, such that lower-order predictions about short sequences of within-word phonemes (up to hundreds of milliseconds long) are informed by higher-

order predictions about long sequences of words (up to minutes long). Interestingly, in **Chapter 6** I did not find a similar integration in reading. In a model-based analysis of three large eye movement datasets, I observed clear effects of both linguistic predictability and parafoveal preview, but strong evidence against predictions *constraining* preview. This suggests that in speech perception, word recognition based on partial information is informed by contextual predictions, while in reading it is not. In **Chapter 6** I also found that word skipping (widely believed to be affected by prediction) was hardly influenced by either contextual prediction or parafoveal preview, and instead largely explained by low-level oculomotor factors. This chapter thus shows that the principles of predictive processing cannot be applied indefinitely, and that the brain in some cases resorts to other (simpler) strategies.

Overall though, the results are remarkably well in line with the bold claims made by predictive processing, and roundly support it as a powerful framework for understanding the brain. Compared to other support, the work in this thesis specifically shows that its computational principles that are often studied in tightly controlled and simple experiments, also apply in much more complex, naturalistic conditions such as listening to a story. In other words, that predictive processing can be found not just in the lab but also "in the wild", in line with the idea that it provides universal principles that apply to neural processing at large.

In what follows I will address some of the larger questions that the work in this thesis raises. I will also discuss what I consider to be the key outstanding questions in the field, and the most promising ways forward to address them.

Multi-level prediction – or multi-level integration?

In **Chapters 3-5**, I present a range of findings from which I conclude that the brain is engaged in prediction at multiple levels of analysis. However, most of these findings demonstrate prediction indirectly, via post-stimulus neural signatures of *deviations* from predictions. This raises the notorious prediction vs integration question. Simply put, do unexpected linguistic stimuli evoke different brain responses because they actually violate a prediction, or because they are more difficult to 'integrate' because of a different reason? Essentially, this is a question about confounds: are we measuring what we think we are measuring, or are we capturing something different?

In psycholinguistic research, authors generally agree on what is meant by prediction: the anticipatory pre-activation of a linguistic representation[1] (but see Box 1). By contrast, 'integration' in this context can be more of a moving target; it is sometimes used as a catch-all for any processing mechanism that does not involve 'pre-

[1]Of course, 'a linguistic representation' is deliberately vague here: there will be much less agreement among psycholinguists once we have to specify the exact nature of this representation.

diction' but nonetheless implies that unpredictable words (or other units) are more difficult to process (Pickering and Gambi, 2018). However, thus construed the concept encompasses an infinite set of hypothetical mechanisms, making it untestable. To transform this epistemological angst into a scientific question, we must be explicit about the mechanism we are considering, i.e. about why an unexpected linguistic stimulus would be more difficult to integrate.

Probably the best known 'integration explanation' is that of facilitated semantic integration via intra-lexical priming (see Brown and Hagoort, 1993; Kutas and Hillyard, 1984; Van Berkum et al., 2005). This is based on the fact that content words that are highly unexpected are often also semantically incongruous. For instance, in the famous sentence 'I take my coffee with cream and *dog*' (Kutas and Hillyard, 1980), the expected word ('sugar') is semantically closer to the preceding words than the unexpected word 'dog'. Therefore, it could be easier to process simply due to bottom-up semantic priming. While such priming may involve pre-activation, it is fundamentally different from linguistic prediction. For one, priming also occurs for word lists or sentences with permuted word order. Moreover, while in natural language primed words may often covary with predictions, they are fundamentally distinct, as one can easily imagine a sentence where the most likely word is not the most semantically associated word[2]. I therefore control for such priming in **Chapter 3 and 4** by including the degree of semantic association between each content word and the preceding context in the baseline model. The results clearly demonstrate that the effect of word unexpectedness is not reducible to facilitated integration via intra-lexical priming (see also Nieuwland et al., 2020).

Surprisal theory (Hale, 2001; Levy, 2008) offers another mechanism that is sometimes cast as an integration explanation. The theory proposes that surprisal is a *causal bottleneck*: a word's unexpectedness *determines* the size of the syntactic update it elicits – thereby *determining* the word's processing difficulty (i.e. the bottleneck). Because this lexical expectancy effect emerges from parsing without requiring explicit lexical expectations, it is sometimes considered an effect of integration rather than prediction (e.g. Kutas, DeLong, and Smith, 2011). However, the theory also assumes an expectation-based syntactic comprehension scheme in which the brain probabilistically activates all potential whole-sentence analyses consistent with the input so far, and where the 'syntactic update' is the Bayesian surprise incurred by updating this expectation (Levy, 2008). Under my definition (Box 1) this would be a syntactic prediction effect. Moreover, the theory focusses on a single bottleneck at the level of lexical surprisal, which cannot explain the additional effects of syn-

[2]For instance, in "After a long day of catching fish, the fisherman went ...", the word 'home' is more likely than 'fishing', although 'fishing' is obviously semantically closer. (Moreover, a function word like 'to' might be even more likely – about ten times as likely, per GPT-2 – despite lacking semantic content.)

tactic, semantic, and phonemic unexpectedness (**Chapter 4**). One could imagine multiple levels of 'causal bottlenecks': perhaps the unexpectedness at each level *determines* the size of the update at that level – and hence the processing difficulty – without requiring explicit predictions about e.g. phoneme or part-of-speech probabilities. However, this would still require expectation-based processing at each level, bringing us back to multi-level predictive processing. Therefore, I do not believe that surprisal theory can explain away the effects from **Chapters 3-5** as integration difficulty *rather than prediction*[3].

Pre-stimulus evidence for prediction: a promising way forward?

Some believe that the 'prediction vs. integration' question poses such a vexing conundrum that future work should avoid it entirely by focussing on pre-stimulus evidence of prediction itself – evidence for pre-activation. However, while this may sound like a promising way forward, it becomes a lot less promising when we consider what that would empirically entail and theoretically imply.

An often-used method to establish pre-stimulus prediction is the visual world paradigm, where participants move their eyes toward the object that the unfolding sentence *could* be referring to (Allopenna, Magnuson, and Tanenhaus, 1998; Altmann and Kamide, 1999). This paradigm has been vital to demonstrate that comprehenders (at least sometimes) spontaneously anticipate not just an upcoming word, but also its linguistic features (Altmann and Mirkovic, 2009; Tanenhaus, 2007). However, the paradigm can only probe predictions of specific, highly predictable words with a concrete referent or associated target that is visually present. It cannot probe the continuous, probabilistic prediction that predictive processing postulates, and hence only provides limited insight into what and especially *when* the brain predicts.

A more modern test for pre-activation is trying to decode words pre-onset, as explored by Goldstein et al. (2021). They report significant decoding (and encoding) of upcoming words up to hundreds of milliseconds pre-onset. While their results are fascinating and impressive, there are limitations to decoding-based tests for pre-activation. Suppose for instance that we can decode whether the next word will be a noun or a verb, well before onset. This could mean the decoder is picking up a pre-activation. However, it could also mean that it is picking up a property of the *previous words* that allow *the decoder* to predict if the next word is a noun – e.g. nouns may often follow article-like activity patterns, while verbs often follow noun-like activity patterns. This issue is not restricted to nouns/verbs[4] but illustrates a fun-

[3] Indeed, surprisal theory to me shows why the dichotomy implied by 'prediction vs integration' is a false one.

[4] Arguably it is worse for decoding semantic vectors, which are by definition similar for neighbouring words.

damental problem: the same regularities that may allow the brain to predict words from context, can allow a decoder to predict upcoming words from brain responses to *preceding* words.[5]

One might try to eliminate the issue by spacing words in time to avoid picking up responses to preceding words. It is unclear how much this would help, since stimulus information is known to fade surprisingly slowly (persisting even after subsequent stimuli, see e.g. King and Wyart, 2021; Nikolić et al., 2009) but in any case this would require unnaturally slow presentation rates, limiting generalisability. Arguably, truly demonstrating pre-activation in *natural* language requires a fully 'white box' approach: first identifying 'the neural representation' of a linguistic unit (whatever that may be) and then establishing pre-activation of that representation. This is a tall order and in my view not a promising way forward.

Personally, I believe that attaching much theoretical significance to pre-activation is misguided and possibly a cultural artefact of the historical definition of prediction as all-or-none pre-activation of specific words. Once we conceive of prediction as inherently probabilistic and multi-faceted, it becomes clear that most if not all mechanisms that can plausibly explain the predictability effects observed in **Chapters 3-5** and elsewhere (see e.g. Hale et al., 2021 for review) will involve pre-activation of *some* kind. Moreover, pre-activation alone does not make a mechanism predictive in any interesting sense. Indeed, even intra-lexical priming can involve pre-activation, via spreading activation. So instead of asking *whether* there is pre-activation, I believe the real way forward is developing a computationally explicit understanding of how the brain generates predictions and how predictability effects arise.

How (explicitly) are predictions implemented?

In this thesis I have addressed *when* linguistic prediction occurs and *what* is being predicted. But so far I have sidestepped *how* predictions might be implemented in the brain. This is a rather large question so I will split it up in two parts – the first being: how *explicitly* are predictions implemented or represented?

When presenting my work, I may have occasionally left the impression to be claiming that the brain actually computes the variables that I use in my analyses. That, for instance, the syntactic prediction from **Chapter 4** would be actually computed in a kind of syntactic softmax layer in the temporal lobe. A critic could counter this by proposing a more implicit scheme that might perform adaptive and prepara-

[5]One might propose the following criterion: decoding only counts as evidence for pre-activation if the *same decoder* can predict upcoming words *better* from the preceding brain activity than from the preceding words themselves. But while this is a solid criterion in principle, it can range from a very low to extremely high bar, depending on how the preceding words are represented to the decoder (e.g. one hot encoding vs contextual embedding).

tory or anticipatory processing which may be functionally equivalent (or approximant) to 'real prediction' but without those complicated probabilistic calculations. To be sure, such an implicit scheme is exactly how I would imagine predictive processing to be implemented in the brain. There is a rich literature on how the brain might *actually perform* probabilistic inference (Ma et al., 2006; Orbán et al., 2016; Pouget et al., 2013) and within predictive processing some have proposed direct mappings between variables or functions in variational algorithms and biological substrates like specific cell-types in specific cortical layers (Bastos et al., 2012; Friston, 2005; Shipp, 2016). But I myself view the framework at a more abstract level, providing principles that the brain might more-or-less-faithfully implement. How faithfully exactly is an interesting question, but lies outside of the scope of this thesis, as it concerns implementation-level computational neuroscience.

A related question that is closer to my work concerns not the low-level implementation, but the computational approximations that the brain might use. This is especially relevant for involved computations, such as hierarchical inference. In **Chapters 4-6** I model context effects in word recognition probabilistically, using a prediction of the upcoming word given the (high-level) global discourse context as a prior for inferring word identity based on (low-level) local features, such as phonemes (**Chapters 4,5**) or partially perceived letters (**Chapter 6**). Formulating this as hierarchical inference – where the posterior or prediction at a higher level is used as a prior at a lower level – provides a powerful tool, not just to model context effects, but also to think about bi-directional information processing in the brain (Friston, 2008; Lee and Mumford, 2003; McClelland, 2013). However, actually performing hierarchical inference can be involved, and hierarchical Bayesian models can sometimes be effectively approximated using surprisingly simple algorithms (e.g. Yu and Cohen, 2008). It would be interesting to explore such approximations to the hierarchical prediction of phonemes from **Chapter 4**. One obvious option is to replace GPT-2 with a more local prediction model, such as an ngram. More interesting would be to try something much simpler, like a unigram prior plus an exponentially decaying frequency count (a leaky integrator) to capture local frequency effects in addition to global word frequency in English. Ideally such a solution could reasonably approximate the 'optimal' solution (at least in natural language), while being substantially simpler to implement (and providing a better fit to the data in **Chapter 4**). I would consider such a hypothetical model an approximation of, and not an alternative to, the explicit approach used in **Chapters 4-6**. After all, such a model would still be hierarchical in the abstract sense of integrating information over multiple scales; it would perhaps only be no longer hierarchical in the technical sense as defined in Bayesian probability theory.

Moving beyond implementation and approximation, the second part of the *how*

question is more algorithmic in nature, and concerns whether the brain uses a prediction-comparison operation. In other words, does the brain compute *prediction errors*?

Does the language system compute prediction errors?

In **Chapters 3-5** I report a range of predictability effects on the evoked response. Many of these modulations are similar to findings from the traditional ERP literature, like N400 modulations (Kutas and Hillyard, 1980), or the PNP (Van Petten and Luka, 2012). However, they also reflect the more general phenomenon of *expectation suppression*: the fact that expected stimuli evoke weaker responses – something also ubiquitously found in other domains, like perception (Keller and Mrsic-Flogel, 2018; Summerfield and de Lange, 2014). A popular explanation of such suppression effects is that the brain compares internal predictions to the incoming signal to compute *prediction errors*. Expected stimuli result in smaller prediction errors, the idea goes, and hence evoke weaker responses.

But there are other potential explanations. For instance, predictive feedback may amplify expected features and suppresses noise, resulting in an enhanced representation but a reduced aggregate response (de Lange, Heilbron, and Kok, 2018; Kok, Jehee, and de Lange, 2012; Lee and Mumford, 2003; see also Aitchison and Lengyel, 2017). In the context of speech perception, this point was nicely illustrated by Luthra et al. (2021). They build on work by Gagnepain, Henson, and Davis (2012) who reported reduced responses to expected phonemes (similar to **Chapters 4 & 5**), and interpreted this as reflecting reduced prediction errors for expected phonemes. Luthra et al. (2021) simulated the same experiment with TRACE (McClelland and Elman, 1986) and found similar reductions in the network, via increased lateral inhibition. TRACE is a *predictive processing* model and the suppression still reflects predictive feedback. However, TRACE does not compute prediction errors so the results should remind us that the suppression effects in **Chapters 3-5** do not necessarily reflect prediction errors.

A more refined method to empirically test for prediction errors was proposed by Blank and Davis (2016). Using simulations, they show that prediction error signals specifically result in an interaction between top-down context and bottom-up signal quality, in the multivariate response. Strikingly, they found this interaction in an fMRI study and a later MEG study (Sohoglu and Davis, 2020), both on the effect of prior knowledge on the perception of noise vocoded speech. Slightly complicating the picture, however, is that their hallmark is based on predictive coding models, which would exhibit the hallmark in just one of the two neuronal subpopulations the models postulate (see also **Chapter 2**). Nevertheless, their use of computational modelling to develop novel empirical hallmarks is an inspiring and promising direction

for future research. For linguistic ERPs, it seems promising to follow this direction and study computational models that have been proposed, for instance for the N400. Some of these explicitly compute prediction errors (Fitz and Chang, 2019; Frank et al., 2015), other compute more implicit prediction errors (Rabovsky, Hansen, and McClelland, 2018; Rabovsky and McRae, 2014) and yet others are engaged in prediction but do not compute errors (Brouwer et al., 2017). Simulating their distinguishing predictions could reveal novel empirical hallmarks that could be used to dissociate error-based and error-free accounts.

Empirically, whether the language system – and the brain at large – computes prediction errors is still an open question. Theoretically, however, I see three arguments for why it would. First, if the brain compares top-down predictions to the input, the error can serve as a 'teaching signal' for self-supervised learning. In a model like TRACE, there is no comparison – but, tellingly, TRACE does not learn: all the connections are hard-coded. Second, prediction errors can improve inference through error-correction. This can avoid a notorious problem of interactive models like TRACE, where top-down feedback simply activates expected features, reinforcing the model's own predictions – a feedback loop which can easily lead to hallucinations (McClelland, 2013; Norris, McQueen, and Cutler, 2000). Having an explicit comparison operation can potentially break this vicious cycle. Finally, we know that the brain uses prediction errors for dopaminergic value-based learning (Schultz, Dayan, and Montague, 1997) and for online error-correction in sensorimotor integration (Keller, Bonhoeffer, and Hübener, 2012; Kitazawa, Kimura, and Yin, 1998; Marr, 1969). Since we know that the brain *can* compute prediction errors, and given the theoretical reasons for why it *should*, it seems likely that errors are indeed computed throughout the brain. Therefore, the *correlates* of errors I find in **Chapters 3-5** may well reflect actual prediction errors – or error-driven updates.

However, this is still largely based on arguments, not facts. And while error-like signals have been observed throughout the brain (Den Ouden, Kok, and De Lange, 2012; Schultz and Dickinson, 2000) especially in the past few years (Fiser et al., 2016; Garrett et al., 2020; Gillon et al., 2021; Hamm and Yuste, 2016; Homann et al., 2017; Jordan and Keller, 2020) the neural basis of a generic prediction error remains elusive. I therefore view the question of prediction errors as one of the most important problems in the field, both for the domain of language, and for predictive processing at large. With increasingly detailed measurements of neural signals, and more ingenious empirical hallmarks like the one explored by Blank and Davis (2016), it is a question on which we can hopefully see some real progress in the years to come.

When are predictions propagated to lower levels?

In my doctoral work I observed signatures of prediction across a range of processing levels, from semantics to early vision, and found that high-level predictions can constrain low-level ones. This suggests that prediction *can* occur across many levels. But does this always happen? And are predictions always passed 'all the way down'?

The enhancement of sensory information in early visual cortex by word knowledge (**Chapter 2**) is a strong indicator of a top-down effect. It is also in line with a large behavioural literature – dating back to the Reicher-Wheeler paradigm (Reicher, 1969; Wheeler, 1970) – suggesting that the effect of word context on letter perception cannot be fully explained by post-perceptual guessing (see Balota, Yap, and Cortese, 2006, for review). It aligns especially well with recent behavioural work showing that readers both subjectively perceive letters in real words as sharper, and are objectively better in detecting subtle perceptual changes in real words than in nonwords (Lupyan, 2017). It is also in line with neurobiological studies, for instance reporting that during phoneme restoration (Warren, 1970) a 'filling-in' of acoustic-phonetic features was found in auditory cortex already (Leonard et al., 2016); and with the broader literature showing top-down recruitment of higher-order areas, both during reading (e.g. Twomey et al., 2011; Woolnough et al., 2021) and speech perception (e.g. Obleser and Kotz, 2011; Park et al., 2015; Sohoglu et al., 2012). Taken together, this suggests the top-down effect from **Chapter 2** reflects a general property of word recognition.

In **Chapter 4**, I report a different kind of top-down effect, finding that predictions based on long timescales (sequences of words in discourse) constrain expectations about short timescales (sequences of phonemes within a word). However, we do not know the processing level at which such integration occurs. In the modulation of the brain response (see Figures 4.5 & S4.13) the effect of surprise seems to start early but extends until hundreds of ms post-phoneme-onset, so presumably this sensitivity reflects at least in part a higher-level surprise-based update and not only the acoustic-phonetic processing of phonemes.

Moreover, **Chapter 6** shows that the logic of hierarchical inference – or 'mutual constraint satisfaction' in connectionist terms (McClelland, Rumelhart, and Group, 1986) – may not always apply. There, we found that readers are clearly sensitive to a word's contextual predictability and its parafoveal recognisability, but that a word's predictability did not influence its parafoveal recognisability. This suggests that linguistic prediction in this case may not be passed 'down' to the level of parafoveal preview (at least not to the extent that it influences eye movements). The lack of this top-down effect may appear at odds with the fMRI evidence for top-down processing in letter perception in **Chapter 2**. But note that the fMRI study only tested

the influence of lexical context – not sentence context – so it could be in line with **Chapter 6**. However, directly connecting the studies seems fraught because of the myriad differences between them, such as that **Chapter 2** focusses on foveal word recognition at fixation, and **Chapter 6** on parafoveal preview in natural reading.

What may explain the insensitivity to linguistic context we find in **Chapter 6**? One possibility is that it reflects a particularity of parafoveal preview, for instance related to time-constraints. On average, readers only look at a word for 250 ms, in which they have to recognise the foveal word and process the parafoveal percept, so perhaps there is too little time to fully integrate the foveal word and let this context inform parafoveal preview[6]. On the other hand, auditory word recognition also unfolds under time-constraints, and in **Chapter 4** we do find an effect of global context in speech perception (line with behavioural work, e.g. Zwitserlood, 1989). Another, more speculative possibility is that the difference between **Chapter 4** and **Chapter 6** may reflect a difference between auditory and visual word recognition. Speech is the natural medium of language, and auditory word recognition occurs in temporal areas tightly connected to the language network more broadly (Hickok and Poeppel, 2007; Yi, Leonard, and Chang, 2019). Reading by contrast is a skill, laboriously acquired by a repurposing of the visual system (Dehaene, 2009; Yeatman and White, 2021), where the high-level linguistic context itself is not processed and perhaps cannot exert as much of an influence as rapidly. This hypothesis could be empirically tested, for instance by systematically comparing the influence of high-level context in auditory vs. visual word recognition.

Reflecting on predictions and lower levels, I will end with the cautionary tale told by **Chapter 5**. In that chapter, I hypothesised that contextual constraint would be positively associated with pre-stimulus beta, reasoning that 'stronger predictions' would imply stronger top-down signalling – an idea inspired by predictive processing interpretations of oscillations (Arnal and Giraud, 2012; Bastos et al., 2012; Lewis and Bastiaansen, 2015). In the end, I observed the opposite pattern and realised that my initial hypothesis did not follow from predictive processing as directly as I thought. For instance, one could also use predictive processing to make the opposite argument: stronger competition between multiple predictions about the incoming word (i.e. more uncertainty) might result in stronger top-down signalling (and thereby enhanced beta). Which one of these hypotheses is 'actually' in line with predictive processing depends on additional assumptions about the processing architecture that do not follow from predictive processing itself. This illustrates that, while the abstract

[6] This issue touches on a larger conundrum: does prediction help us process language rapidly, compensating for the rapidity of language (Christiansen and Chater, 2016) – or does the rapid rate of language impose a limit on the use of prediction? In other words, does the rapid rate of natural language necessitate prediction or hinder prediction?

principles from predictive processing can provide theoretical guidance, effectively applying them to a given domain always requires additional domain-specific knowledge and models.

What are the sources of linguistic predictions?

Throughout this thesis, I have interpreted the what-question mostly in terms of the *content* of predictions. For instance, in **Chapter 4** the syntactic prediction is a prediction *about* (morpho)syntax. A different – arguably more common (e.g. Huettig, 2015) – interpretation of the what-question is to consider not the content but the *source* of the prediction. Viewed this way, a syntactic prediction is a prediction *based on* syntax.

This approach is notably taken by a series of recent papers on syntactic prediction (Brennan et al., 2020; Brennan and Hale, 2019; Hale et al., 2018; Shain et al., 2020; see also Henderson et al., 2016; Nelson et al., 2017; see Hale et al., 2021 for review). These studies mostly take the brain's sensitivity to surprisal as a given (as an index for expectation-based processing difficulty), and then compute surprisal using multiple (linguistically-informed) models and compare which one best fits the 'neural surprisal' observed in brain signals. This has for instance revealed that the brain is specifically sensitive to surprisal computed from probabilistic context free grammars (PCFGs) – suggesting it uses hierarchical syntax to guide its predictions (Brennan and Hale, 2019; Shain et al., 2020, c.f. Frank et al., 2015). Hale et al. (2018) take this approach further by combining a generative model of phrase structure grammar with an explicit (cognitively interpretable) parsing strategy. They show that the surprisal from the explicit syntactic model not only fits the EEG data better than that from an LSTM, but also that different aspects of the parsing operation can be linked to different syntax-related components in the EEG response (see also Brennan et al., 2020).

Asking what predictions are based on is interesting because it can reveal what information the brain is using. Indeed, I use the same approach in **Chapter 4** & **6** when I ask whether the brain uses global context to make local predictions. What I find inspiring about these specific studies, is that they demonstrate that this approach – 'using prediction as a window into language' – can be used to ask questions not just about predictive processing but also about (psycho)linguistic theory proper. As such, it shows how the unifying framework of predictive processing does not replace existing domain-specific theories, but integrates them into a broader picture of the brain as a prediction machine – a perspective that can then be used to answer other, domain-specific questions (such as questions about syntax or parsing).

This line of work embodies an exciting research programme, with many potential

directions left to be explored. To name just one, it can be interesting to take the same approach to the phoneme level, and compare the theory-agnostic lexical-statistical approach I take to compute phoneme surprisal in **Chapter 4-5** (see also Brodbeck, Hong, and Simon, 2018; Ettinger, Linzen, and Marantz, 2014; Gwilliams et al., 2018) with models that draw more on linguistic theory, such those of probabilistic phonotactics (Di Liberto et al., 2019). This way, by studying what the language system *expects* (or more precisely, what it did not expect), we may eventually uncover the underlying principles by which it operates.

Concluding remarks

In this thesis, I have tried to evaluate predictive processing as a framework for understanding the brain, using language as a testbed. The results overall broadly support the framework: they underscore that language processing is inherently predictive, and show how predictions can be used as a window into how the brain processes information. Inevitably, the work in this thesis raises or leaves unaddressed more questions than it answers. Does the success of predictive language models in AI tell us anything about prediction in human language processing? What is the role of prediction in language acquisition? Does linguistic prediction make use of the production system – or do they merely rely on the same linguistic generative models? And are some functions of linguistic prediction more important than others?

Taking a step back, the work also shows the potential of using something as complex as language for the study of predictive processing. While it may seem appealing to focus on something 'simple' like early visual processing, language actually has distinct advantages such as allowing to study predictive processing in naturalistic conditions. This is true now more than ever, since we have powerful computational techniques and generative models to approximate – for any arbitrary stimulus – the linguistic statistics that a truly predictive brain should be sensitive to. Similar generative models for other domains like vision are rapidly improving, and I am optimistic that technological advances like these will eventually lead to *conceptual* advances – not just on how language works, but also on how the brain works, and, ultimately, on what it means to be human.

Nederlandse samenvatting

Het omzetten van talige signalen – trillingen in de lucht, vormpjes op papier – naar woorden en *gedachten* is een van de meest verbluffende prestaties van het menselijk brein. Een recente theorie beschrijft het brein als een voorspellingsmachine, die zintuigelijke informatie voortdurend vergelijkt met interne voorspellingen. In dit proefschrift toets ik deze theorie en gebruik ik taalverwerking als mijn proeftuin. Twee vragen staan centraal: wanneer doet het brein talige voorspellingen? En, als het voorspellingen doet, wat voorspelt het dan precies?

In **hoofdstuk 1** introduceer ik de theorie en plaats ik deze in een historische context.

In **hoofdstuk 2** richt ik me op het waarnemen van letters. Letters kun je makkelijker herkennen in een context (zoals op een verkeersbord) dan zonder context (zoals op een kenteken). In een fMRI-experiment vond ik dat talige context al vroeg in het visuele systeem de informatie over de waargenomen letters versterkt. Dit betekent dat ons brein onze woordenkennis gebruikt om te voorspellen welke vormen we zien. En het impliceert dat we letters in een context niet alleen beter kunnen raden – maar ook letterlijk beter kunnen *zien*.

In **hoofdstuk 3 tot en met 5** analyseer ik hersensignalen van mensen die luisteren naar luisterboeken. Daarnaast analyseer ik de tekst van de luisterboeken zelf, om de voorspelbaarheid van ieder woord en elke klank in het boek te schatten. Het brein blijkt hier zeer gevoelig voor: het reageert sterker op woorden als ze onvoorspelbaarder zijn. Dit geldt niet alleen voor sommige woorden, maar voor schijnbaar *alle* woorden in het verhaal. Bovendien reageert het brein verschillend wanneer de betekenis, de klank of de grammatica onverwacht is. Al met al suggereren de resultaten uit deze hoofdstukken dat ons brein *voortdurend* voorspellingen doet – en op uiteenlopende niveaus (klank, betekenis, grammatica).

In **hoofdstuk 6** analyseer ik oogbewegingen van mensen die lange teksten – en zelfs een volledige roman – lezen. Hieruit blijkt dat voorspellingen een sterk effect hebben op hoe lang proefpersonen naar een woord keken, maar niet naar welke woorden proefpersonen keken. Daarnaast blijkt dat woordherkenning tijdens lezen meer lokale (dat wil zeggen, minder contextuele) voorspellingen raadpleegt dan woordherkenning tijdens spraakperceptie (vergelijk **hoofdstuk 4**).

Alles overziend (**hoofdstuk 7**) ondersteunen de studies de theorie en schetsen een beeld van ons taalsysteem als inherent voorspellend. Een systeem dat taal kan 'volgen' door het voortdurend een stap voor te zijn en te voorspellen – een beetje zoals *autocomplete* op je telefoon. Maar, anders dan *autocomplete*, voorspelt ons brein op uiteenlopende niveaus: van de letters die we zien tot de betekenis van het volgende woord.

Dankwoord

De afgelopen jaren heb ik mogen werken met slimme, inspirerende collega's en was ik omringd door lieve mensen vol geduld en begrip. Zonder hen was dit onderzoek niet mogelijk geweest.

Veel heb ik te danken aan m'n begeleiders. Floris, wat een geluk om met jou te hebben kunnen werken. Je hebt me veel kansen gegeven en was altijd beschikbaar en ontvankelijk voor nieuwe ideeën. Dat was al zo toen ik in 2016 uit het niets vanuit Parijs contact opnam. Ik stelde voor om samen een PhD-beursaanvraag te schrijven. Het was half augustus maar je antwoordde direct. Goed idee, zei je, maar wist ik wel dat de deadline volgende maand was? (Ik had geen idee natuurlijk.) Binnen enkele dagen spraken we af en een paar weken later was het ingediend. Zo flexibel was je. Ook later tijdens mijn PhD, naarmate mijn werk verder afdreef van het plan, bleef je grenzeloos nieuwsgierig naar elke nieuwe wending. Je stelde scherpzinnige vragen die mijn onderzoek verbeterden en me telkens uit een impasse trokken als ik vastzat. Je feedback was trefzeker maar nooit dwingend of zelfs maar sturend – je liet me alles altijd zelf ontdekken. Mocht ik later ooit promovendi begeleiden dan hoop ik net zo'n goede mentor te worden als jij. Of half zo goed – dat zou al fantastisch zijn.

Peter, ik had me geen betere tweede begeleider kunnen wensen. Als levende legende op het gebied van brein en taal, en wandelende encyclopedie van de cognitiewetenschap, heb je me zoveel bijgebracht. Meer dan eens zette je me weer met beide benen op de grond als ik meende iets nieuws te hebben bedacht of gevonden. Fijntjes wees je me dan op een reeks artikelen, vaak al decennia oud (en soms zelfs van jouw hand) die hetzelfde hadden bedacht of aangetoond – en dan vaak zonder de hulp van nieuwerwetse technieken. Pijnlijke realisaties waren dat, die vormend zijn geweest en m'n proefschrift uiteindelijk zoveel sterker maakten. Behalve over taal, brein en geest, leerde je me over de regels van het academisch spel. En voor zover mijn onderzoek enig historisch bewustzijn heeft, heb ik dit aan jou te danken.

I feel indebted to all my colleagues in the Predictive Brain Lab. You have been supportive and kind; gave great feedback and made lab meetings something to look forward to. I'm especially grateful to my direct collaborators in my PhD. First, David and Matthias, who taught me all I know about fMRI analysis. And who convinced me, right at the start, to transition from MATLAB to python – easily one of the

best decisions of my PhD. I am grateful to Benedikt, who taught me about EEG, deconvolution, advanced statistics and dancing. And to Britta, queen of the beamformer, wizard of brainwaves, for pulling me through the challenges of chapter 5. En natuurlijk Jorie, tovenaar achter het toetsenbord, polyglot der programmeertalen. Je stage werd door corona overhoop gegooid, maar wat heb je goed kunnen schakelen. In een fractie van de tijd heb je een indrukwekkend project neergezet. Hoewel het niet altijd duidelijk was wie nu wie begeleidde, ben ik dankbaar een jaar lang zo intensief samen te hebben gewerkt. Hoofdstuk zes was zonder jou nooit zo mooi en succesvol geweest. Kim, your stay in Nijmegen was horrendously disrupted by covid, and we ended up speaking more on Zoom than in real life. But you managed to pull through and put together a terrific project that I'm happy to be a part of. Pius, I wasn't your primary supervisor but it was a real pleasure to work with you. I'm sure you will do brilliant (and societally more meaningful) work in your PhD. En Peter, onderdeel van PBL by legacy, het was het een eer om samen aan het overzichtsartikel te schrijven.

I'm also grateful to my colleagues at the MPI and the NBL group in particular. You were kind, warm colleagues and together form such an interesting, diverse group. Each Friday morning, I was curious what this week's lab meeting would bring. Fresh field work on emerging sign languages in Turkey? New methodological insights for laminar fMRI? Pragmatic inferences in gesture? It was all possible. Thanks for the NBL PhD meetings – organized by Sophie and Laura – discussing doctoral life, the (im)possiblity of syntax in neural nets, the philosophy of the brain, language, intelligence, and beyond. A special word of thanks to Kristijan and Jan-Mathijs, who agreed to collaborate in chapter 4. Without the MEG data you painstakingly collected and all your input, it wouldn't have turned out the way it did. Mante, onze gesprekken en jouw scherpe vragen hielpen me een beter zicht te krijgen op de literatuur en de kern van het debat over voorspellingen in taal. En Andrea, ook al is die samenwerking er nooit meer van gekomen, ik heb veel van onze stimulerende gesprekken geleerd.

Beyond my 'own groups', I want to thank other colleagues at the Donders (or Radboud or MPI) that I interacted with, that made Nijmegen the inspiring environment that it is. In no particular order, I should at least mention Sushrut, Giacomo, Surya, Johannes, Nadine, Roel, Stefan, Jeroen. En uiteraard Paul, master van de MRI, keizer van de kelder. Daarnaast de technische groep: Mike, Marek, Hong, Erik, en consorten. De computationele mogelijkheden op het Donders hebben mijn leven de afgelopen jaren een stuk makkelijker gemaakt. I also want to thank Betty e Mora, for offering the finest canteen food (and vibes) in Nijmegen.

Only one out of the five empirical chapters in this thesis made use of data I collected myself. Most of what I did would have been impossible without the generos-

ity of others. This has been immensely useful during the pandemic, which impacted some 40% of my time as a PhD. While I had to change some plans, I could continue working when the labs closed – all thanks to open data. I am similarly grateful to those making open toolboxes and software. And to those making openly available teaching materials. I should specifically mention Stanford University (CS231n, CS224n, CS224u, have been transformative); Mike X Cohen, Peter Bloem, and Steve Brunton. The open videos, lectures, blogs and demos allowed me to develop the math and computer science skills that were instrumental in my PhD.

Then there are those who contributed to my academic development prior to my PhD. Surya, Stefan and Chris, who hired me as an RA in Utrecht, where I gained research experience. In London I was lucky enough to meet Maria, who gave me the opportunity to collaboratively turn my literature thesis into a review paper, which is now my best known first-author paper. Richard, who supervised me at the FIL and guided and me in the sometimes overwhelming Theoretical Neurobiology and Methods groups. Finally, in Paris, to Florent, who taught me much about math, statistics, and computational neuroscience. All this experience made me feel scientifically quite mature from the start of my PhD already. It gave me the confidence to work rather independently, and develop my own research style and agenda.

I want to thank the dissertation committee and opponents for taking time to assess my doctoral work. En Jessy en Stefan, voor de briljante cover.

Wetenschap gebeurt niet in een vacuüm en ik mag van geluk spreken dat ik was omringd door lieve vrienden en (schoon)familie. Door mijn ouders en m'n zus, die er altijd voor me waren en me adviseerden over levenskeuzes en de academische wereld. Door al m'n vrienden, in Amsterdam en elders, die bijna allemaal iets anders doen en me in de hoognodige afleiding van het academische voorzien. Door m'n paranimfen, Miel en Roman, die in dit proces aan m'n zijde staan. En natuurlijk door Dorien. Mijn lieve Dorien, die me nog elke dag uitdaagt, prikkelt, en me zoveel leert over het leven. Die me steunde als ik er doorheen zat, me motiveerde als ik het niet meer zag zitten, en me uit m'n sleur trok als ik teveel werkte. Die zo ongelofelijk veel geduld heeft gehad. Het is een groot geluk mijn leven met jou te mogen delen en ik hoop dat nog vele jaren te kunnen doen.

Publication list

- Richter, D., **Heilbron, M.** & de Lange, F.P. (2022). Dampened sensory representations for expected input across the ventral visual stream. (under review)
- **Heilbron, M.**, van Haren, J., Hagoort, P., & de Lange, F. P. (2021). Prediction and preview strongly affect reading times but not skipping during natural reading. *bioRxiv*. (under review)
- **Heilbron, M.**, Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2021). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*. (under review)
- **Heilbron, M.**, Richter, D., Ekman, M., Hagoort, P., & De Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature communications*, 11(1), 1-11.
- **Heilbron, M.**, Ehinger, B., Hagoort, P., de Lange, F.P (2019). Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. *Conference on Cognitive Computational Neuroscience, 424-427.*
- **Heilbron, M.**, & Meyniel, F. (2019). Confidence resets reveal hierarchical adaptive learning in humans. PLoS *computational biology, 15(4), e1006972.*
- De Lange, F. P., **Heilbron, M.**, & Kok, P. (2018). How do expectations shape perception?. *Trends in cognitive sciences, 22(9), 764-779.*
- **Heilbron, M.**, & Chait, M. (2018). Great expectations: is there evidence for predictive coding in auditory cortex?. *Neuroscience, 389, 54-73.*
- Paffen, C. L., Gayet, S., **Heilbron, M.**, & Van der Stigchel, S. (2018). Attention-based perceptual learning does not affect access to awareness. Journal of Vision, 18(3), 7-7.
- Gayet, S., van Maanen, L., **Heilbron, M.**, Paffen, C. L., & Van der Stigchel, S. (2016). Visual input that matches the content of visual working memory requires less (not faster) evidence sampling to reach conscious access. *Journal of Vision, 16(11), 26-26.*

Curriculum vitae

Micha Heilbron was born in Amsterdam in 1992. After graduating from high school in 2010, he pursued undergraduate degrees in Psychobiology and in Natural and Social Sciences (major: Theoretical Philosophy) at the University of Amsterdam. During his studies, he organised symposia and debates on societal topics and ocassionally wrote essays and book reviews for various online outlets. Between his bachelor's and master's, Micha worked as a teaching assistant at the University of Amsterdam (Faculty of Sciences, Mathematics and Informatics) and as a research assistent at Utrecht University (Department of Cognitive Psychology) in the lab of Dr. Stefan van der Stigchel. He then completed a Dual Master Programme in Brain and Mind Sciences: first at University College London (with distinction), where he did a research internship at the Wellcome Trust Centre for Neuroimaging on modelling EEG abnormalities in epilepsy, supervised by Dr. Richard Rosch in the Theoretical Neurobiology and Methods group of Prof. Karl Friston. His second (M2) degree he obtained at École normale supérieure and Université Pierre et Marie Curie (*mention très bien*) in Paris, where he did internship research on hierarchical Bayesian models of adaptive learning, supervised by Dr. Florent Meyniel in the Computational Brain Team of the Unicog Department of Neurospin, lead by Prof. Stanislas Dehaene. In 2017 he obtained a NWO Research Talent Grant for pursuing his doctoral research, under the supervision of Prof. Floris de Lange and Prof. Peter Hagoort. Currently, he is a postdoctoral researcher in the Predictive Brain Lab at Radboud University.

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/phd/

Research data management

Ethics

This thesis is based on the results of human studies, which were conducted in accordance with the principles of the Declaration of Helsinki. All studies followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands; Ethics board Trinity College Dublin, Ireland; Ghent University, Belgium; Southampton University, United Kingdom; University of Dundee, United Kingdom; Brigham Young University, United States), including informed consent of all participants.

Data availability

The data that was required originally for this thesis (Chapter 2) has already been published on the Donders Data Repository and is accessible through the DOI listed below. It is available under a data use agreement for identifiable human data (Version RU-DI-HD-1.0) and will remain accessible online for at least 10 years after termination. For the chapters that are not yet published as a journal article, I provide DSC (Data Sharing Collection) identifiers that will become active once the final journal article is published, plus DAC (Data Acquisition Collection) identifiers those studies using data internally acquired at the Donders that is not yet publicly available. For the chapters using publicly available data, the DSC does not re-publish the already published data, but links to the repositories (e.g Datadryad, OSF) where the original data is available, under its original licence. In addition the DSC collections contain code and computational results (e.g. of numerical simulations) required for reproducing the results. Chapter 3 has appeared as a conference paper and will not be published as a journal article; its analyses are a sub-set of Chapter 4, and the associated code and computational results will be part of the publication of that collection.

Chapter 2

Data and code available at: https://doi.org/10.34973/t894-sz74

Chapter 3

This chapter only makes use of public data
Data sharing collection identifier: DSC_3018000.00_752
DOI (active upon publication): https://doi.org/10.34973/kwsh-cb29

Chapter 4

Data sharing collection identifier: DSC_3018000.00_752
DOI (active upon publication): https://doi.org/10.34973/dfkm-h813
Data acquisition collection identifier (DAC): di.dccn.DAC_3011085.05_985

Chapter 5

Data sharing collection identifier: DSC_3018000.00_000
DOI (active upon publication): https://doi.org/10.34973/kwsh-cb29
Data acquisition collection identifier (DAC): di.dccn.DAC_3011085.05_985

Chapter 6

This chapter only makes use of public data
Data sharing collection identifier: DSC_3018000.00_215.
DOI (active upon publication): https://doi.org/10.34973/wnc3-0m87

Bibliography

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:ISSN: 1662-5196.

Abrams D. A., Nicol T., Zecker S., Kraus N. (2008). Right-Hemisphere Auditory Cortex Is Dominant for Coding Syllable Patterns in Speech. *The Journal of Neuroscience* 28(15):3958–3965. ISSN: 0270-6474.

Adams M. J. (1979). Models of word recognition. *Cognitive Psychology* 11(2):133–176. ISSN: 0010-0285.

Aitchison L., Lengyel M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*. Computational Neuroscience 46:219–227. ISSN: 0959-4388.

Ali A., Ahmad N., Groot E. d., Gerven M. A. J. v., Kietzmann T. C. (2021). Predictive coding is a consequence of energy efficiency in recurrent neural networks. *bioRxiv*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 2021.02.16.430904.

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. ISSN: 0749-596X.

Altmann G. T. M., Kamide Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73(3):247–264. ISSN: 0010-0277.

Altmann G. T. M., Mirkovic J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science* 33(4):583–609. ISSN: 1551-6709.

Armeni K. (2021). On model-based neurobiology of language comprehension: Neural oscillations, processing memory, and prediction. *Radboud University Nijmegen (PhD Thesis)*.

Armeni K., Willems R. M., Bosch A. van den, Schoffelen J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*:ISSN: 1053-8119.

Arnal L. H., Giraud A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences* 16(7):390–398. ISSN: 1364-6613.

Arnal L. H., Wyart V., Giraud A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience* 14(6):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational neuroscience;Cortex;Sensory systems Subject_term_id: computational-neuroscience;cortex;sensory-systems, 797–801. ISSN: 1546-1726.

Atal B. S., Schroeder M. R. (1970). Adaptive Predictive Coding of Speech Signals. *Bell System Technical Journal* 49(8):_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1970.tb04297.x, 1973–1986. ISSN: 1538-7305.

Baevski A., Zhou H., Mohamed A., Auli M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*:arXiv: 2006.11477.

Bahdanau D., Chorowski J., Serdyuk D., Brakel P., Bengio Y. (2016). End-to-end attention-based large vocabulary speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 4945–4949.

Bahl L. R., Jelinek F., Mercer R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2):Publisher: IEEE, 179–190.

Baker J. (1975). The DRAGON system–An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23(1):Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing, 24–29. ISSN: 0096-3518.

Bakhtiari S., Mineault P., Lillicrap T., Pack C. C., Richards B. A. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *bioRxiv*:Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article, 2021.06.18.448989.

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. ISSN: 0010-0285.

Balota D. A., Yap M. J., Cortese M. J. (2006). Visual Word Recognition: The Journey from Features to Meaning (A Travel Update). *Handbook of Psycholinguistics (Second Edition)*. Ed. by Traxler M. J., Gernsbacher M. A. London: Academic Press, pp. 285–375. ISBN: 978-0-12-369374-7.

Bar M. (2004). Visual objects in context. *Nature Reviews Neuroscience* 5(8):617–629. ISSN: 1471-0048.

— (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences* 11(7):280–289. ISSN: 13646613.

Bar M. et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences* 103(2):449–454. ISSN: 0027-8424, 1091-6490.

Barlow H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*. Ed. by Rosenblith W. MIT Press, pp. 217–234.

Baron J., Thurston I. (1973). An analysis of the word-superiority effect. *Cognitive Psychology* 4(2):207–228. ISSN: 0010-0285.

Bastos A. M., Lundqvist M., Waite A. S., Kopell N., Miller E. K. (2020). Layer and rhythm specificity for predictive routing. *bioRxiv*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.01.27.921783.

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. ISSN: 1097-4199.

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. ISSN: 1097-4199.

Bever T. G., Poeppel D. (2010). Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. *BIOLINGUISTICS* 4(2-3):Number: 2-3, 174–200. ISSN: 1450-3417.

Bialek W., Nemenman I., Tishby N. (2001). Predictability, complexity, and learning. *Neural Computation* 13(11):2409–2463. ISSN: 0899-7667.

Bicknell K., Levy R. (2010). A Rational Model of Eye Movement Control in Reading. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 1168–1178.

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):Publisher: Oxford Academic, 2767–2796. ISSN: 1047-3211.

Blank H., Davis M. H. (2016). Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. *PLoS biology* 14(11):e1002577. ISSN: 1545-7885.

Bod R., Scha R. J. H., Sima'an K. (2003). *Data Oriented Parsing*. Stanford University. ISBN: 978-1-57586-435-8.

Bouma H., Voogd A. H. d. (1974). On the control of eye saccades in reading. *Vision Research* 14(4):Publisher: Elsevier, 273–284. ISSN: 0042-6989.

Brandman T., Peelen M. V. (2017). Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *Journal of Neuroscience* 37(32):7700–7710. ISSN: 0270-6474, 1529-2401.

Brennan J. R., Dyer C., Kuncoro A., Hale J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*:107479. ISSN: 0028-3932.

Brennan J. R., Hale J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE* 14(1):Publisher: Public Library of Science, e0207741. ISSN: 1932-6203.

Brink D. van den, Brown C. M., Hagoort P. (2001). Electrophysiological Evidence for Early Contextual Influences during Spoken-Word Recognition: N200 Versus N400 Effects. *Journal of Cognitive Neuroscience* 13(7):Publisher: MIT Press, 967–985. ISSN: 0898-929X.

Brink D. van den, Hagoort P. (2004). The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. *Journal of Cognitive Neuroscience* 16(6):1068–1084. ISSN: 0898-929X.

Brodbeck C., Hong L. E., Simon J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current biology: CB* 28(24):3976–3983.e5. ISSN: 1879-0445.

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.

Broderick M. P., Anderson A. J., Lalor E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*:0584–19. ISSN: 0270-6474, 1529-2401.

Brouwer H., Crocker M. W., Venhuizen N. J., Hoeks J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science* 41(Suppl Suppl 6):1318–1352. ISSN: 0364-0213.

Brown C., Hagoort P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience* 5(1):Publisher: MIT Press, 34–44. ISSN: 0898-929X.

Brown T. B. et al. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*:arXiv: 2005.14165.

Brysbaert M., Drieghe D. (2003). Please stop using word frequency data that are likely to be word length effects in disguise. *Behavioral and Brain Sciences* 26(4):Publisher: [New York]: Cambridge University Press, 1978-, 479.

Brysbaert M., New B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4):977–990. ISSN: 1554-3528.

Burton M. W., Baum S. R., Blumstein S. E. (1989). Lexical effects on the phonetic categorization of speech: the role of acoustic structure. *Journal of Experimental Psychology. Human Perception and Performance* 15(3):567–575. ISSN: 0096-1523.

Buschman T. J., Miller E. K. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science* 315(5820):Publisher: American Association for the Advancement of Science Section: Report, 1860–1862. ISSN: 0036-8075, 1095-9203.

Buswell G. T. (1920). *An experimental study of the eye-voice span in reading*. 17. University of Chicago.

Cattell J. M. (1886). The Time Taken up by Cerebral Operations. *Mind* 11(43):377–392. ISSN: 0026-4423.

Caucheteux C., King J.-R. (2020). Language processing in brains and deep neural networks: computational convergence and its limits. *bioRxiv*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.07.03.186288.

Chalk M., Marre O., Tkacik G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences* 115(1):Publisher: National Academy of Sciences Section: Biological Sciences, 186–191. ISSN: 0027-8424, 1091-6490.

Chan W., Jaitly N., Le Q., Vinyals O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 4960–4964.

Chang F., Dell G. S., Bock K. (2006). Becoming syntactic. *Psychological Review* 113(2):234–272. ISSN: 0033-295X.

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. ISSN: 1097-4199.

Charniak E. (1997). Statistical parsing with a context-free grammar and word statistics. *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*. AAAI'97/IAAI'97. Providence, Rhode Island: AAAI Press, pp. 598–603. ISBN: 978-0-262-51095-0.

Chater N., Manning C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10(7):335–344. ISSN: 1364-6613.

Chomsky N. (1957). *Syntactic Structures*. Publication Title: Syntactic Structures. De Gruyter Mouton. ISBN: 978-3-11-021832-9.

Christiansen M. H., Chater N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39:Publisher: Cambridge University Press. ISSN: 0140-525X, 1469-1825.

Clark A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3):Publisher: Cambridge University Press, 181–204. ISSN: 0140-525X, 1469-1825.

Clifton C., Ferreira F., Henderson J. M., Inhoff A. W., Liversedge S. P., Reichle E. D., Schotter E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language* 86:1–19. ISSN: 0749-596X.

Cohen L., Dehaene S., Naccache L., Lehéricy S., Dehaene-Lambertz G., Hénaff M. A., Michel F. (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain: A Journal of Neurology* 123 (Pt 2):291–307. ISSN: 0006-8950.

Coltheart M., Rastle K., Perry C., Langdon R., Ziegler J. (2001). DRC a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108(1):204–256. ISSN: 0033-295X.

Cop U., Dirix N., Drieghe D., Duyck W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods* 49(2):602–615. ISSN: 1554-3528.

Dalal S. S., Baillet S., Adam C., Ducorps A., Schwartz D., Jerbi K., Bertrand O., Garnero L., Martinerie J., Lachaux J.-P. (2009). Simultaneous MEG and intracranial EEG recordings during attentive reading. *NeuroImage* 45(4):1289–1304. ISSN: 1053-8119.

Dale A. M., Fischl B., Sereno M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* 9(2):179–194. ISSN: 1053-8119.

Daube C., Ince R. A. A., Gross J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology* 29(12):1924–1937.e9. ISSN: 0960-9822.

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. ISSN: 1053-8119.

Davis M. H., Johnsrude I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research* 229(1-2):132–147. ISSN: 0378-5955.

Dayan P., Hinton G. E., Neal R. M., Zemel R. S. (1995). The Helmholtz machine. *Neural Computation* 7(5):889–904. ISSN: 0899-7667.

de Lange F. P., Heilbron M., Kok P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences* 22(9):764–779. ISSN: 1364-6613.

Dearborn W. F. (1906). *The psychology of reading: an experimental study of the reading pauses and movements of the eye ...* Archives of philosophy, psychology and scientific methods, no. 4. n. p.: Archives of philosophy, psychology and scientific methods, no. 4.

Dehaene S. (2009). *Reading in the brain: The new science of how we read.* Penguin.

Dehaene S., Cohen L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences* 15(6):254–262. ISSN: 1879-307X.

Dehaene S., Cohen L., Sigman M., Vinckier F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences* 9(7):335–341. ISSN: 1364-6613.

Dell G. S., Chang F. (2014). The P-chain. Relating sentence production and its disorders to comprehension and acquisition. 369(1634):20120394. ISSN: 1471-2970.

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):Number: 8 Publisher: Nature Publishing Group, 1117–1121. ISSN: 1546-1726.

Den Ouden H., Kok P., De Lange F. (2012). How Prediction Errors Shape Perception, Attention, and Motivation. *Frontiers in Psychology* 3:548. ISSN: 1664-1078.

Desikan R. S. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31(3):968–980. ISSN: 1053-8119.

Deubel H., O'Regan J. K., Radach R. (2000). Commentary on Section 2 - Attention, Information Processing, and Eye Movement Control. *Reading as a Perceptual Process.* Ed. by Kennedy A., Radach R., Heller D., Pynte J. Oxford: North-Holland, pp. 355–374. ISBN: 978-0-08-043642-5.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT (1).*

Di Liberto G. M., O'Sullivan J. A., Lalor E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology* 25(19):2457–2465. ISSN: 0960-9822.

Di Liberto G. M., Wong D., Melnik G. A., Cheveigne A. de (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *NeuroImage* 196:237–247. ISSN: 1053-8119.

Dikker S., Rabagliati H., Farmer T. A., Pylkkänen L. (2010). Early Occipital Sensitivity to Syntactic Category Is Based on Form Typicality. *Psychological Science* 21(5):Publisher: SAGE Publications Inc, 629–634. ISSN: 0956-7976.

Ding N., Simon J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* 109(29):Publisher: National Academy of Sciences Section: Biological Sciences, 11854–11859. ISSN: 0027-8424, 1091-6490.

Donhauser P. W., Baillet S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron* 105(2):385–393.e9. ISSN: 0896-6273.

Drieghe D., Brysbaert M., Desmet T., De Baecke C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology* 16(1-2):79–103. ISSN: 0954-1446, 1464-0635.

Duan Y., Bicknell K. (2020). A Rational Model of Word Skipping in Reading: Ideal Integration of Visual and Linguistic Information. *Topics in Cognitive Science* 12(1):387–401. ISSN: 1756-8757, 1756-8765.

Dupoux E., Kakehi K., Hirose Y., Pallier C., Mehler J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25(6):Place: US Publisher: American Psychological Association, 1568–1578. ISSN: 1939-1277.

Ede F. van, Jensen O., Maris E. (2010). Tactile expectation modulates pre-stimulus beta-band oscillations in human sensorimotor cortex. *NeuroImage* 51(2):867–876. ISSN: 1053-8119.

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. ISSN: 0022-5371.

Elias P. (1955). Predictive coding–I. *IRE Transactions on Information Theory* 1(1):Conference Name: IRE Transactions on Information Theory, 16–24. ISSN: 2168-2712.

Elman J. L. (1990). Finding structure in time. *Cognitive Science* 14(2):179–211. ISSN: 0364-0213.

— (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7(2):195–225. ISSN: 1573-0565.

Engbert R., Kliegl R. (2003). The game of word skipping: Who are the competitors? *Behavioral and Brain Sciences* 26(4):Publisher: Cambridge University Press, 481.

Engbert R., Nuthmann A., Richter E. M., Kliegl R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review* 112(4):777–813. ISSN: 0033-295X.

Engel A. K., Fries P. (2010). Beta-band oscillations–signalling the status quo? *Current Opinion in Neurobiology* 20(2):156–165. ISSN: 1873-6882.

Ettinger A., Linzen T., Marantz A. (2014). The role of morphology in phoneme prediction: evidence from MEG. *Brain and Language* 129:14–23. ISSN: 1090-2155.

Federmeier K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44(4):491–505. ISSN: 0048-5772.

Fiser A., Mahringer D., Oyibo H. K., Petersen A. V., Leinweber M., Keller G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience* 19(12):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group

Subject_term: Learning and memory;Sensory processing Subject_term_id: learning-and-memory;sensory-processing, 1658–1664. ISSN: 1546-1726.

Fitz H., Chang F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology* 111:15–52. ISSN: 0010-0285.

Fleur D. S., Flecken M., Rommers J., Nieuwland M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition* 204:104335. ISSN: 0010-0277.

Fodor J. A. (1983). *The Modularity of Mind*. Cambridge, MA, USA: A Bradford Book. ISBN: 978-0-262-06084-4.

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 Section A* 33(4):Publisher: SAGE Publications, 465–495. ISSN: 0272-4987.

Forster K. I. (1989). Basic issues in lexical processing. *Lexical representation and process*. Cambridge, MA, US: The MIT Press, pp. 75–107. ISBN: 978-0-262-13240-4.

Frank S. L., Fernandez Monsalve I., Thompson R. L., Vigliocco G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods* 45(4):1182–1190. ISSN: 1554-3528.

Frank S. L., Otten L. J., Galli G., Vigliocco G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language* 140:1–11. ISSN: 0093-934X.

Friederici A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences* 6(2):78–84. ISSN: 1364-6613.

Friston K. (2008). Hierarchical Models in the Brain. *PLOS Computational Biology* 4(11):Publisher: Public Library of Science, e1000211. ISSN: 1553-7358.

— (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11(2):127–138. ISSN: 1471-003X, 1471-0048.

Friston K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360(1456):815–836. ISSN: 0962-8436.

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. ISSN: 1053-8119.

Friston K. J., Moran R., Seth A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology* 23(2):172–178. ISSN: 1873-6882.

Friston K. J., Williams S., Howard R., Frackowiak R. S., Turner R. (1996). Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine* 35(3):346–355. ISSN: 0740-3194.

Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., Pezzulo G. (2017). Active Inference: A Process Theory. *Neural Computation* 29(1):1–49. ISSN: 0899-7667.

Fujioka T., Trainor L. J., Large E. W., Ross B. (2009). Beta and gamma rhythms in human auditory cortex during musical beat processing. *Annals of the New York Academy of Sciences* 1169:89–92. ISSN: 1749-6632.

Gagnepain P., Henson R. N., Davis M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology* 22(7):Publisher: Elsevier, 615–621. ISSN: 0960-9822.

Garrett M. et al. (2020). Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife* 9:e50340. ISSN: 2050-084X.

Gill P., Woolley S. M. N., Fremouw T., Theunissen F. E. (2008). What's That Sound? Auditory Area CLM Encodes Stimulus Surprise, Not Intensity or Intensity Changes. *Journal of Neurophysiology* 99(6):Publisher: American Physiological Society, 2809–2820. ISSN: 0022-3077.

Gillon C. J. et al. (2021). Learning from unexpected events in the neocortical microcircuit. *bioRxiv*:Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article, 2021.01.15.426915.

Glezer L. S., Jiang X., Riesenhuber M. (2009). Evidence for highly selective neuronal tuning to whole words in the "Visual Word Form Area". *Neuron* 62(2):199–204. ISSN: 0896-6273.

Glezer L. S., Riesenhuber M. (2013). Individual variability in location impacts orthographic selectivity in the "visual word form area". *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33(27):11221–11226. ISSN: 1529-2401.

Goldstein A. et al. (2021). Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*:Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Goodkind A., Bicknell K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*. Salt Lake City, Utah: Association for Computational Linguistics, pp. 10–18.

Goodman K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist* 6(4):Place: United Kingdom Publisher: Taylor & Francis, 126–135.

Gorgolewski K. J. et al. (2017). Nipype a flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.13.1.

Goutte C., Nielsen F. A., Hansen L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE transactions on medical imaging* 19(12):1188–1201. ISSN: 0278-0062.

Gramfort A., Luessi M., Larson E., Engemann D. A., Strohmeier D., Brodbeck C., Parkkonen L., Hämäläinen M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage* 86:446–460. ISSN: 1053-8119.

Graves A., Mohamed A.-r., Hinton G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 6645–6649.

Gray R. M. (2010). *Linear Predictive Coding and the Internet Protocol*. Now Publishers Inc.

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*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.04.04.025684.

Gwilliams L., Poeppel D., Marantz A., Linzen T. (2018). Phonological (un)certainty weights lexical activation. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*. Salt Lake City, Utah: Association for Computational Linguistics, pp. 29–34.

Hagoort P., Brown C. M. (2000). ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia* 38(11):1531–1549. ISSN: 0028-3932.

Hagoort P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences* 9(9):416–423. ISSN: 1364-6613.

Hagoort P., Brown C., Groothusen J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes* 8(4):Place: United Kingdom Publisher: Taylor & Francis, 439–483. ISSN: 1464-0732(Electronic),0169-0965(Print).

Hale J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hale J. T., Campanelli L., Li J., Bhattasali S., Pallier C., Brennan J. R. (2021). Neuro-computational models of language processing. *Annual Review of Linguistics*:Publisher: Annual Reviews.

Hale J., Dyer C., Kuncoro A., Brennan J. (2018). Finding syntax in human encephalography with beam search. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2727–2736.

Halle M., Stevens K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory* 8(2):Conference Name: IRE Transactions on Information Theory, 155–159. ISSN: 2168-2712.

Hamm J. P., Yuste R. (2016). Somatostatin Interneurons Control a Key Component of Mismatch Negativity in Mouse Visual Cortex. *Cell Reports* 16(3):Publisher: Elsevier, 597–604. ISSN: 2211-1247.

Heeger D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*:201619788. ISSN: 0027-8424, 1091-6490.

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

Heilbron M., Armeni K., Schoffelen J.-M., Hagoort P., de Lange F. P. (2021a). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*.

Heilbron M., Chait M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*:ISSN: 0306-4522.

Heilbron M., Ehinger B., Hagoort P., Lange F. P. de (2019). Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. *2019 Conference on Cognitive Computational Neuroscience*:arXiv: 1909.04400.

Heilbron M., Richter D., Ekman M., Hagoort P., Lange F. P. de (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications* 11(1):Number: 1 Publisher: Nature Publishing Group, 321. ISSN: 2041-1723.

Heilbron M., van Haren J., Hagoort P., de Lange F. P. (2021b). Prediction and preview strongly affect reading times but not skipping during natural reading. *bioRxiv*.

Henaff O. J., Srinivas A., Fauw J. D., Razavi A., Doersch C., Eslami S. M. A., Oord A. v. d. (2019). Data-Efficient Image Recognition with Contrastive Predictive Coding.

Henderson J. M., Choi W., Lowder M. W., Ferreira F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132:293–300. ISSN: 1095-9572.

Herten M. van, Kolk H. H. J., Chwilla D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Brain Research. Cognitive Brain Research* 22(2):241–255. ISSN: 0926-6410.

Hickok G., Poeppel D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience* 8(5):Number: 5 Publisher: Nature Publishing Group, 393–402. ISSN: 1471-0048.

Hindle D., Rooth M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics* 19(1):103–120. ISSN: 0891-2017.

Hinton G. E., Dayan P., Frey B. J., Neal R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science* 268(5214):Publisher: American Association for the Advancement of Science Section: Reports, 1158–1161. ISSN: 0036-8075, 1095-9203.

Hohenstein S., Kliegl R. (2014). Semantic preview benefit during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 40(1):166–190. ISSN: 1939-1285.

Hohwy J. (2013). *The Predictive Mind*. Oxford: Oxford University Press. ISBN: 978-0-19-968273-7.

Homann J., Koay S. A., Glidden A. M., Tank D. W., Berry M. J. (2017). Predictive Coding of Novel versus Familiar Stimuli in the Primary Visual Cortex. *bioRxiv*:197608.

Hosoya T., Baccus S. A., Meister M. (2005). Dynamic predictive coding by the retina. *Nature* 436(7047):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7047 Primary_atype: Research Publisher: Nature Publishing Group, 71–77. ISSN: 1476-4687.

Huang Y., Rao R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(5):Publisher: Wiley Online Library, 580–593.

Huettig F. (2015). Four central questions about prediction in language processing. *Brain Research*. Predictive and Attentive Processing in Perception and Action 1626:118–135. ISSN: 0006-8993.

Huettig F., Mani N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience* 31(1):19–31.

Huettig F., Rommers J., Meyer A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*. Visual search and visual world: Interactions among visual attention, language, and working memory 137(2):151–171. ISSN: 0001-6918.

Huth A. G., Heer W. A. de, Griffiths T. L., Theunissen F. E., Gallant J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600):Number: 7600 Publisher: Nature Publishing Group, 453–458. ISSN: 1476-4687.

Inhoff A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior* 23(5):612–624. ISSN: 0022-5371.

Issa E. B., Cadieu C. F., DiCarlo J. J. (2018). Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *eLife* 7:ed. by Connor E., Marder E., Connor E. Publisher: eLife Sciences Publications, Ltd, e42870. ISSN: 2050-084X.

Jackendoff R. (2003). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford, New York: Oxford University Press. ISBN: 978-0-19-926437-7.

Jaeger F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology* 61(1):23–62. ISSN: 0010-0285.

Jain S., Huth A. G. (2018). Incorporating Context into Language Encoding Models for fMRI. *bioRxiv*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 327601.

Jehee J. F. M., Rothkopf C., Beck J. M., Ballard D. H. (2006). Learning receptive fields using predictive feedback. *Journal of Physiology-Paris*. Theoretical and Computational Neuroscience: Understanding Brain Functions 100(1):125–132. ISSN: 0928-4257.

Jelinek F., Bahl L., Mercer R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory* 21(3):Conference Name: IEEE Transactions on Information Theory, 250–256. ISSN: 1557-9654.

Jelinek F. (1998). *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press. ISBN: 978-0-262-10066-3.

Jensen O., Spaak E., Zumer J. M. (2014). Human brain oscillations: From physiological mechanisms to analysis and cognition. *Magnetoencephalography*. Vol. 9783642330452. Springer, pp. 359–403. ISBN: 978-3-642-33044-5.

Jordan R., Keller G. B. (2020). Opposing Influence of Top-down and Bottom-up Input on Excitatory Layer 2/3 Neurons in Mouse Primary Visual Cortex. *Neuron* 108(6):1194–1206.e5. ISSN: 0896-6273.

Jurafsky D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic linguistics* 21:Publisher: Citeseer.

Jurafsky D., Martin J. H. (2014). *Speech and language processing*. Vol. 3. Pearson London.

— (2021). Speech and Language Processing (3rd draft ed.)

Kay K. N., Yeatman J. D. (2017). Bottom-up and top-down computations in word- and face-selective cortex. *eLife* 6:ed. by Gold J. I., e22341. ISSN: 2050-084X.

Keller G. B., Bonhoeffer T., Hübener M. (2012). Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron* 74(5):809–815. ISSN: 0896-6273.

Keller G. B., Mrsic-Flogel T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron* 100(2):424–435. ISSN: 0896-6273.

Kennedy A. (2003). The dundee corpus [cd-rom].

Kennedy A., Pynte J., Murray W. S., Paul S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: an eye movement analysis. *Quarterly Journal of Experimental Psychology (2006)* 66(3):601–618. ISSN: 1747-0226.

Kerkoerle T. v., Self M. W., Dagnino B., Gariel-Mathis M.-A., Poort J., Togt C. v. d., Roelfsema P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National*

Academy of Sciences 111(40):Publisher: National Academy of Sciences Section: Feature Article, 14332–14341. ISSN: 0027-8424, 1091-6490.

Kersten D., Mamassian P., Yuille A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology* 55:271–304. ISSN: 0066-4308.

Khan A. G., Hofer S. B. (2018). Contextual signals in visual cortex. *Current Opinion in Neurobiology* 52:131–138. ISSN: 1873-6882.

Kiebel S. J., Daunizeau J., Friston K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLOS Computational Biology* 4(11):Publisher: Public Library of Science, e1000209. ISSN: 1553-7358.

King J.-R., Wyart V. (2021). The Human Brain Encodes a Chronicle of Visual Events at Each Instant of Time Through the Multiplexing of Traveling Waves. *Journal of Neuroscience* 41(34):Publisher: Society for Neuroscience Section: Research Articles, 7224–7233. ISSN: 0270-6474, 1529-2401.

Kitazawa S., Kimura T., Yin P.-B. (1998). Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* 392(6675):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6675 Primary_atype: Research Publisher: Nature Publishing Group, 494–497. ISSN: 1476-4687.

Kleiner M., Brainard D., Pelli D., Ingling A., Murray R., Broussard C. (2007). What's new in psychtoolbox-3. *Perception* 36(14):1–16. ISSN: 0301-0066.

Kliegl R., Grabner E., Rolfs M., Engbert R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16(1-2):262–284. ISSN: 0954-1446, 1464-0635.

Kliegl R., Nuthmann A., Engbert R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135(1):Place: US Publisher: American Psychological Association, 12–35. ISSN: 1939-2222(Electronic),0096-3445(Print).

Kliegl R., Risse S., Laubrock J. (2007). Preview benefit and parafoveal-on-foveal effects from word n + 2. *Journal of Experimental Psychology: Human Perception and Performance* 33(5):Place: US Publisher: American Psychological Association, 1250–1255. ISSN: 1939-1277(Electronic),0096-1523(Print).

Knill D. C., Pouget A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27(12):712–719. ISSN: 0166-2236.

Koen V., Skak J. S., Vandborg S. K. (2010). Silk speech codec. *The Internet Engineering Task Force (IETF)*.

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. ISSN: 0896-6273.

Kuperberg G. R., Jaeger T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience* 31(1):32–59. ISSN: 2327-3798.

Kutas M., DeLong K. A., Smith N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*. New York, NY, US: Oxford University Press, pp. 190–207. ISBN: 978-0-19-539551-8.

Kutas M., Hillyard S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207(4427):Publisher: American Association for the Advancement of Science Section: Reports, 203–205. ISSN: 0036-8075, 1095-9203.

Kutas M., Hillyard S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307(5947):Place: United Kingdom Publisher: Nature Publishing Group, 161–163. ISSN: 1476-4687(Electronic),0028-0836(Print).

Lalor E. C., Pearlmutter B. A., Reilly R. B., McDarby G., Foxe J. J. (2006). The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage* 32(4):1549–1561. ISSN: 1053-8119.

Lawrence S. J. D., Formisano E., Muckli L., Lange F. P. de (2017). Laminar fMRI: Applications for cognitive neuroscience. *NeuroImage*:ISSN: 1053-8119.

LeCun Y. (2016). Predictive Learning.

LeCun Y., Bengio Y., Hinton G. (2015). Deep learning. *Nature* 521(7553):436–444. ISSN: 1476-4687.

Lee T. S., Mumford D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 20(7):1434–1448. ISSN: 1084-7529.

Legendre P. (2008). Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology* 1(1):3–8. ISSN: 1752-9921.

Leonard M. K., Baud M. O., Sjerps M. J., Chang E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications* 7:13619. ISSN: 2041-1723.

Levy R. (2008). Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177. ISSN: 0010-0277.

Lewis A. G., Bastiaansen M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*. Special issue: Prediction in speech and language processing 68:155–168. ISSN: 0010-9452.

Lewis A. G., Wang L., Bastiaansen M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*. The electrophysiology of speech, language, and its precursors 148:51–63. ISSN: 0093-934X.

Li X., Zhang Y., Xia J., Swaab T. Y. (2017). Internal mechanisms underlying anticipatory language processing: Evidence from event-related-potentials and neural oscillations. *Neuropsychologia* 102:70–81. ISSN: 0028-3932.

Liberman A. M., Cooper F. S., Shankweiler D. P., Studdert-Kennedy M. (1967). Perception of the speech code. *Psychological Review* 74(6):Place: US Publisher: American Psychological Association, 431–461. ISSN: 1939-1471.

Linzen T., Baroni M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linguistics* 7(1):_eprint: https://doi.org/10.1146/annurev-linguistics-032020-051035, 195–212.

Liu P. J., Saleh M., Pot E., Goodrich B., Sepassi R., Kaiser L., Shazeer N. (2018). Generating Wikipedia by Summarizing Long Sequences. *International Conference on Learning Representations*.

Liu X., Zhang F., Hou Z., Mian L., Wang Z., Zhang J., Tang J. (2021). Self-supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*:Conference Name: IEEE Transactions on Knowledge and Data Engineering, 1–1. ISSN: 1558-2191.

Lopopolo A., Frank S. L., Bosch A. v. d., Willems R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE* 12(5):Publisher: Public Library of Science, e0177794. ISSN: 1932-6203.

Lotter W., Kreiman G., Cox D. (2017). Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. *arXiv:1605.08104 [cs, q-bio]*:arXiv: 1605.08104.

Luke S. G., Christianson K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology* 88:22–60. ISSN: 0010-0285.

— (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods* 50(2):826–833. ISSN: 1554-3528.

Lupyan G. (2017). Objective effects of knowledge on visual perception. *Journal of Experimental Psychology. Human Perception and Performance* 43(4):794–806. ISSN: 1939-1277.

Luthra S., Li M. Y. C., You H., Brodbeck C., Magnuson J. S. (2021). Does signal reduction imply predictive coding in models of spoken word recognition? *Psychonomic Bulletin & Review* 28(4):1381–1389. ISSN: 1531-5320.

Ma W. J., Beck J. M., Latham P. E., Pouget A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9(11):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 11 Primary_atype: Research Publisher: Nature Publishing Group, 1432–1438. ISSN: 1546-1726.

Manning C. D., Clark K., Hewitt J., Khandelwal U., Levy O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision.

Proceedings of the National Academy of Sciences:Publisher: National Academy of Sciences Section: Physical Sciences. ISSN: 0027-8424, 1091-6490.

Manning C., Schutze H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mantegna F., Hintz F., Ostarek M., Alday P. M., Huettig F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia* 134:107199. ISSN: 0028-3932.

Maris E., Oostenveld R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164(1):177–190. ISSN: 0165-0270.

Marr D. (1969). A theory of cerebellar cortex. *The Journal of Physiology* 202(2):437–470.1. ISSN: 0022-3751.

— (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA, USA: MIT Press. ISBN: 978-0-262-51462-0.

Marslen-Wilson W. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature* 244(5417):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5417 Primary_atype: Research Publisher: Nature Publishing Group, 522–523. ISSN: 1476-4687.

— (1989). Access and integration: Projecting sound onto meaning. *Lexical representation and process*. Cambridge, MA, US: The MIT Press, pp. 3–24. ISBN: 978-0-262-13240-4.

Marslen-Wilson W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*. Special Issue Spoken Word Recognition 25(1):71–102. ISSN: 0010-0277.

Martelli M., Majaj N. J., Pelli D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision* 5(1):58–70. ISSN: 1534-7362.

Matchin W., Brodbeck C., Hammerly C., Lau E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping* 40(2):_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24403, 663–678. ISSN: 1097-0193.

Matchin W., Hickok G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex (New York, N.Y.: 1991)* 30(3):1481–1498. ISSN: 1460-2199.

Mayer A., Schwiedrzik C. M., Wibral M., Singer W., Melloni L. (2016). Expecting to See a Letter: Alpha Oscillations as Carriers of Top-Down Sensory Predictions. *Cerebral Cortex* 26(7):3146–3160. ISSN: 1047-3211.

McClelland J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Frontiers in Psychology* 4:Publisher: Frontiers. ISSN: 1664-1078.

McClelland J. L., Elman J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18(1):1–86. ISSN: 0010-0285.

McClelland J. L., Hill F., Rudolph M., Baldridge J., Schutze H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*:Publisher: National Academy of Sciences Section: Perspective. ISSN: 0027-8424, 1091-6490.

McClelland J. L., Mirman D., Holt L. L. (2006). Are there interactive processes in speech perception? *Trends in cognitive sciences* 10(8):363–369. ISSN: 1364-6613.

McClelland J. L., O'Regan J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance* 7(3):Place: US Publisher: American Psychological Association, 634–644. ISSN: 1939-1277(Electronic),0096-1523(Print).

McClelland J. L., Rumelhart D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review* 88(5):375–407. ISSN: 1939-1471(Electronic),0033-295X(Print).

McClelland J. L., Rumelhart D. E., Group P. R. (1986). *Parallel distributed processing*. Vol. 2. MIT press Cambridge, MA.

McConkie G. W., Rayner K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics* 17(6):Publisher: Springer, 578–586.

McQueen J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance* 17(2):Place: US Publisher: American Psychological Association, 433–443. ISSN: 1939-1277.

Meyniel F. (2020). Brain dynamics for confidence-weighted learning. *PLOS Computational Biology* 16(6):Publisher: Public Library of Science, e1007935. ISSN: 1553-7358.

Michalareas G., Vezoli J., Pelt S. van, Schoffelen J.-M., Kennedy H., Fries P. (2016). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* 89(2):384–397. ISSN: 1097-4199.

Miller G. A., Heise G. A., Lichten W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology* 41(5):Place: US Publisher: American Psychological Association, 329–335. ISSN: 0022-1015.

Miller G. A., Isard S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior* 2(3):217–228. ISSN: 0022-5371.

Millidge B., Tschantz A., Buckley C. L. (2020). Predictive Coding Approximates Backprop along Arbitrary Computation Graphs. *arXiv:2006.04182 [cs]*:arXiv: 2006.04182.

Mooney C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue canadienne de psychologie* 11(4):Place: Canada Publisher: University of Toronto Press, 219–226. ISSN: 0008-4255.

Morton J. (1964). The Effects of Context upon Speed of Reading, Eye Movements and Eye-voice Span. *Quarterly Journal of Experimental Psychology* 16(4):Publisher: SAGE Publications, 340–354. ISSN: 0033-555X.

Mumford D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics* 66(3):241–251. ISSN: 0340-1200.

Neisser U. (1967). *Cognitive Psychology: Classic Edition*. New York: Psychology Press. ISBN: 978-1-315-73617-4.

Nelson M. J. et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences* 114(18):Publisher: National Academy of Sciences Section: PNAS Plus, E3669–E3678. ISSN: 0027-8424, 1091-6490.

Nieuwland M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews* 96:367–400. ISSN: 1873-7528.

Nieuwland M. S., Arkhipova Y., Rodriguez-Gomez P. (2020). Anticipating words during spoken discourse comprehension. A large-scale, pre-registered replication study using brain potentials. *Cortex: a Journal Devoted to the Study of the Nervous System and Behavior* 133:1–36. ISSN: 1973-8102.

Nieuwland M. S. et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7:ed. by Shinn-Cunningham B. G. Publisher: eLife Sciences Publications, Ltd, e33468. ISSN: 2050-084X.

Nieuwland M. S. et al. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1791):Publisher: Royal Society, 20180522.

Nikolić D., Häusler S., Singer W., Maass W. (2009). Distributed Fading Memory for Stimulus Properties in the Primary Visual Cortex. *PLOS Biology* 7(12):Publisher: Public Library of Science, e1000260. ISSN: 1545-7885.

Norris D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52(3):189–234. ISSN: 0010-0277.

— (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological Review* 113(2):327–357. ISSN: 0033-295X.

— (2009). Putting it all together: a unified account of word recognition and reaction-time distributions. *Psychological Review* 116(1):207–219. ISSN: 0033-295X.

Norris D., McQueen J. M., Cutler A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23(3):Publisher: Cambridge University Press, 299–325. ISSN: 1469-1825, 0140-525X.

Obleser J., Kotz S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage* 55(2):713–723. ISSN: 1053-8119.

Olshausen B. A., Field D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6583 Primary_atype: Research Publisher: Nature Publishing Group, 607–609. ISSN: 1476-4687.

Oord A. v. d., Li Y., Vinyals O. (2019). Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*:arXiv: 1807.03748.

Oostenveld R., Fries P., Maris E., Schoffelen J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience* 2011:156869. ISSN: 1687-5273.

Orbán G., Berkes P., Fiser J., Lengyel M. (2016). Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron* 92(2):530–543. ISSN: 1097-4199.

O'Regan J. K. (1980). The control of saccade size and fixation duration in reading: The limits of linguistic control. *Perception & Psychophysics* 28(2):112–117. ISSN: 1532-5962.

— (1992). Optimal Viewing Position in Words and the Strategy-Tactics Theory of Eye Movements in Reading. *Eye Movements and Visual Cognition: Scene Perception and Reading*. Ed. by Rayner K. Springer Series in Neuropsychology. New York, NY: Springer, pp. 333–354. ISBN: 978-1-4612-2852-3.

Osterhout L., Holcomb P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language* 31(6):785–806. ISSN: 0749-596X.

Paap K. R., Newsome S. L., McDonald J. E., Schvaneveldt R. W. (1982). An activation–verification model for letter and word recognition: The word-superiority effect. *Psychological Review* 89(5):573–594. ISSN: 1939-1471(Electronic),0033-295X(Print).

Park H., Ince R. A. A., Schyns P. G., Thut G., Gross J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Current Biology* 25(12):1649–1653. ISSN: 0960-9822.

Pedregosa F. et al. (2011). Scikit-learn Machine Learning in Python. *Journal of Machine Learning Research* 12(Oct):2825–2830. ISSN: ISSN 1533-7928.

Pelli D. G., Farell B., Moore D. C. (2003). The remarkable inefficiency of word recognition. *Nature* 423(6941):752–756. ISSN: 1476-4687.

Pereira F. (2000). Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358(1769):Publisher: Royal Society, 1239–1253.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. (2018). Deep contextualized word representations. *Proceedings of NAACL-HLT*, pp. 2227–2237.

Piai V., Anderson K. L., Lin J. J., Dewar C., Parvizi J., Dronkers N. F., Knight R. T. (2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proceedings of the National Academy of Sciences* 113(40):Publisher: National Academy of Sciences Section: Biological Sciences, 11366–11371. ISSN: 0027-8424, 1091-6490.

Piai V., Roelofs A., Maris E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia* 53:146–156. ISSN: 0028-3932.

Piai V., Roelofs A., Rommers J., Maris E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping* 36(7):2767–2780. ISSN: 1097-0193.

Piantadosi S. T., Tily H., Gibson E. (2012). The communicative function of ambiguity in language. *Cognition* 122(3):280–291. ISSN: 0010-0277.

Pickering M. J., Gambi C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin* 144(10):Place: US Publisher: American Psychological Association, 1002–1044. ISSN: 1939-1455(Electronic),0033-2909(Print).

Pickering M. J., Garrod S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36(4):Publisher: Cambridge University Press, 329–347. ISSN: 0140-525X, 1469-1825.

Pouget A., Beck J. M., Ma W. J., Latham P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience* 16(9):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Neural encoding Subject_term_id: neural-encoding, 1170–1178. ISSN: 1546-1726.

Prystauka Y., Lewis A. G. (2019). The power of neural oscillations to inform sentence comprehension: A linguistic perspective. *Language and Linguistics Compass* 13(9):_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12347, e12347. ISSN: 1749-818X.

Pynte J., Kennedy A. (2006). An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research* 46(22):3786–3801. ISSN: 0042-6989.

Rabovsky M., Hansen S. S., McClelland J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour* 2(9):693. ISSN: 2397-3374.

Rabovsky M., McRae K. (2014). Simulating the N400 ERP component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition* 132(1):68–89. ISSN: 1873-7838.

Radford A., Narasimhan K., Salimans T., Sutskever I. (2018). *Improving language understanding by generative pre-training*.

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* 1:8.

Ramachandran P., Varoquaux G. (2011). Mayavi 3D visualization of scientific data. *Computing in Science & Engineering* 13(2):40–51.

Rao R. P., 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. ISSN: 1097-6256.

Rayner K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology* 7(1):65–81. ISSN: 0010-0285.

— (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition* 5(4):443–448. ISSN: 1532-5946.

— (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology (2006)* 62(8):1457–1506. ISSN: 1747-0226.

Rayner K., Juhasz B. J., Brown S. J. (2007). Do readers obtain preview benefit from word N + 2? A test of serial attention shift versus distributed lexical processing models of eye movement control in reading. *Journal of Experimental Psychology. Human Perception and Performance* 33(1):230–245. ISSN: 0096-1523.

Rayner K., Pollatsek A. (1987). Eye movements in reading: A tutorial review. *Attention and performance 12: The psychology of reading*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, pp. 327–362. ISBN: 978-0-86377-083-8 978-0-86377-084-5.

Rayner K., Well A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review* 3(4):504–509. ISSN: 1531-5320.

Reicher G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology* 81(2):275–280. ISSN: 0022-1015.

Reichle E. D., Rayner K., Pollatsek A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *The Behavioral and Brain Sciences* 26(4):445–476, 445–476. ISSN: 0140-525X.

Reilly R. G., O'Regan J. K. (1998). Eye movement control during reading: A simulation of some word-targeting strategies. *Vision Research* 38(2):303–317. ISSN: 0042-6989.

Remez R. E., Rubin P. E., Pisoni D. B., Carrell T. D. (1981). Speech Perception Without Traditional Speech Cues. *Science*:Publisher: American Association for the Advancement of Science.

Richter D., Ekman Matthias de Lange F. P. (2018). Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *Journal of Neuroscience* 38(34):7452–7461. ISSN: 0270-6474, 1529-2401.

Ridgway G. R., Litvak V., Flandin G., Friston K. J., Penny W. D. (2012). The problem of low variance voxels in statistical parametric mapping; a new hat avoids a 'haircut'. *Neuroimage* 59(3):2131–2141. ISSN: 1053-8119.

Riesenhuber M., Poggio T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2(11):1019–1025. ISSN: 1097-6256.

Rommers J., Dickson D. S., Norton J. J. S., Wlotko E. W., Federmeier K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience* 32(5):576–589. ISSN: 2327-3798.

Rubin J., Ulanovsky N., Nelken I., Tishby N. (2016). The Representation of Prediction Error in Auditory Cortex. *PLoS computational biology* 12(8):e1005058. ISSN: 1553-7358.

Ruder S., Peters M. E., Swayamdipta S., Wolf T. (2019). Transfer Learning in Natural Language Processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 15–18.

Rumelhart D. E., McClelland J. L. (1982). An Interactive Activation Model of Context Effects in Letter Perception: II. The Contextual Enhancement Effect and Some Tests and Extensions of the Model. *Psychological Review* 89(1):60–94.

Rumelhart D. E., Siple P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review* 81(2):99–118. ISSN: 1939-1471(Electronic),0033-295X(Print).

Ryskin R., Levy R. P., Fedorenko E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia* 136:107258. ISSN: 0028-3932.

Saravanan V., Berman G. J., Sober S. J. (2020). Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neurons, behavior, data analysis and theory* 3(5).

Schotter E. R., Angele B., Rayner K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics* 74(1):5–35. ISSN: 1943-393X.

Schotter E. R., Lee M., Reiderman M., Rayner K. (2015). The effect of contextual constraint on parafoveal processing in reading. *Journal of Memory and Language* 83:118–139. ISSN: 0749-596X.

Schrimpf M., Blank I., Tuckute G., Kauf C., Hosseini E. A., Kanwisher N., Tenenbaum J., Fedorenko E. (2020). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*:Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.06.26.174482.

Schroeder M. R. (2004). *Computer Speech: Recognition, Compression, Synthesis*. 2nd ed. Springer Series in Information Sciences. Berlin Heidelberg: Springer-Verlag. ISBN: 978-3-540-21267-6.

Schultz W., Dayan P., Montague P. R. (1997). A Neural Substrate of Prediction and Reward. *Science* 275(5306):Publisher: American Association for the Advancement of Science, 1593–1599.

Schultz W., Dickinson A. (2000). Neuronal Coding of Prediction Errors. *Annual Review of Neuroscience* 23(1):_eprint: https://doi.org/10.1146/annurev.neuro.23.1.473, 473–500.

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. ISSN: 1097-4199.

Sedley W., Gander P. E., Kumar S., Kovach C. K., Oya H., Kawasaki H., Howard III M. A., Griffiths T. D. (2016). Neural signatures of perceptual inference. *eLife* 5:ed. by King A. J. Publisher: eLife Sciences Publications, Ltd, e11476. ISSN: 2050-084X.

Seidenberg M. S., McClelland J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review* 96(4):523–568. ISSN: 0033-295X.

Shain C., Blank I. A., Schijndel M. van, Schuler W., Fedorenko E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* 138:107307. ISSN: 0028-3932.

Shannon C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal* 30(1):_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1951.tb01366.x, 50–64. ISSN: 1538-7305.

Shipp S. (2016). Neural Elements for Predictive Coding. *Frontiers in Psychology* 7:ISSN: 1664-1078.

Siegel M., Donner T. H., Engel A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience* 13(2):Bandiera_abtest: a

Cg_type: Nature Research Journals Number: 2 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cognitive neuroscience;Computational neuroscience;Neuronal physiology;Sensorimotor processing Subject_term_id: cognitive-neuroscience;computational-neuroscience;neuronal-physiology;sensorimotor-processing, 121–134. ISSN: 1471-0048.

Smith A. T., Kosillo P., Williams A. L. (2011). The confounding effect of response amplitude on MVPA performance measures. *NeuroImage* 56(2):525–530. ISSN: 1095-9572.

Smith E. C., Lewicki M. S. (2006). Efficient auditory coding. *Nature* 439(7079):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7079 Primary_atype: Research Publisher: Nature Publishing Group, 978–982. ISSN: 1476-4687.

Smith N. J., Kutas M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology* 52(2):157–168. ISSN: 0048-5772.

Smith N. J., Levy R. (2008). Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

— (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319. ISSN: 0010-0277.

Smith S. M., Nichols T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44(1):Publisher: Elsevier, 83–98.

Smith S. M. et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23 Suppl 1:S208–219. ISSN: 1053-8119.

Sohoglu E., Davis M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *eLife* 9:ed. by King A. J., Kok P., Kok P., Press C., Lalor E. C. Publisher: eLife Sciences Publications, Ltd, e58077. ISSN: 2050-084X.

Sohoglu E., Peelle J. E., Carlyon R. P., Davis M. H. (2012). Predictive Top-Down Integration of Prior Knowledge during Speech Perception. *Journal of Neuroscience* 32(25):Publisher: Society for Neuroscience Section: Articles, 8443–8453. ISSN: 0270-6474, 1529-2401.

Spitzer B., Haegens S. (2017). Beyond the Status Quo: A Role for Beta Oscillations in Endogenous Content (Re)Activation. *eNeuro* 4(4):Publisher: Society for Neuroscience Section: Review. ISSN: 2373-2822.

Spratling M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing* 17(3):279–305. ISSN: 1612-4790.

Srinivasan M. V., Laughlin S. B., Dubs A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci* 216(1205):427–59. ISSN: 0080-4649 (Print).

Staub A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass* 9(8):_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12151, 311–327. ISSN: 1749-818X.

Stevens K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America* 32:Place: US Publisher: Acoustical Society of American, 47–55. ISSN: 0001-4966.

Summerfield C., de Lange F. P. d. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience* 15(11):nrn3838. ISSN: 1471-0048.

Suzuki S., Cavanagh P. (1995). Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance* 21(4):901–913. ISSN: 1939-1277(Electronic),0096-1523(Print).

Szewczyk J. M., Schriefers H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience* 33(6):Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2017.1401101, 665–686. ISSN: 2327-3798.

Tanenhaus M. K. (2007). Eye movements and spoken language processing. *Eye movements: A window on mind and brain*. Amsterdam, Netherlands: Elsevier, pp. 443–469. ISBN: 978-0-08-044980-7.

Thesen T. et al. (2012). Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nature Communications* 3:1284. ISSN: 2041-1723.

Thompson M. C., Massaro D. W. (1973). Visual information and redundancy in reading. *Journal of Experimental Psychology* 98(1):49–54. ISSN: 0022-1015(Print).

Tiffin-Richards S. P., Schroeder S. (2015). Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology* 27(5):Publisher: Routledge _eprint: https://doi.org/10.1080/20445911.2014.999076, 531–548. ISSN: 2044-5911.

Toneva M., Wehbe L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems* 32:14954–14964.

Toshniwal S., Kannan A., Chiu C.-C., Wu Y., Sainath T. N., Livescu K. (2018). A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition. *arXiv:1807.10857 [cs, eess]*:arXiv: 1807.10857.

Turken A. U., Dronkers N. F. (2011). The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses. *Frontiers in Systems Neuroscience* 5:ISSN: 1662-5137.

Twomey T., Kawabata Duncan K. J., Price C. J., Devlin J. T. (2011). Top-down modulation of ventral occipito-temporal responses during visual word recognition. *NeuroImage* 55(3):1242–1251. ISSN: 1053-8119.

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. ISSN: 0278-7393.

Van Essen D. C., Dierker D. L. (2007). Surface-Based and Probabilistic Atlases of Primate Cerebral Cortex. *Neuron* 56(2):209–225. ISSN: 0896-6273.

Van Petten C., Luka B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*. Predictive information processing in the brain: Principles, neural mechanisms and models 83(2):176–190. ISSN: 0167-8760.

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

Veldre A., Andrews S. (2018). Parafoveal preview effects depend on both preview plausibility and target predictability. *Quarterly Journal of Experimental Psychology* 71(1):Publisher: SAGE Publications, 64–74. ISSN: 1747-0218.

Vinckier F., Dehaene S., Jobert A., Dubus J. P., Sigman M., Cohen L. (2007). Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron* 55(1):143–156. ISSN: 0896-6273.

Vitu F., O'Regan J. K., Inhoff A. W., Topolski R. (1995). Mindless reading: Eye-movement characteristics are similar in scanning letter strings and reading texts. *Perception & Psychophysics* 57(3):352–364. ISSN: 1532-5962.

Wacongne C., Labyt E., Wassenhove V. van, 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 of the United States of America* 108(51):20754–20759. ISSN: 1091-6490.

Wang L., Hagoort P., Jensen O. (2018). Language Prediction Is Reflected by Coupling between Frontal Gamma and Posterior Alpha Oscillations. *Journal of Cognitive Neuroscience* 30(3):432–447. ISSN: 1530-8898.

Wang X.-J. (2010). Neurophysiological and Computational Principles of Cortical Rhythms in Cognition. *Physiological reviews* 90(3):1195–1268. ISSN: 0031-9333.

Warren R. M. (1970). Perceptual restoration of missing speech sounds. *Science (New York, N.Y.)* 167(3917):392–393. ISSN: 0036-8075.

Weide R. (1998). *The CMU pronunciation dictionary*. Carnegie Mellon University Pittsburgh, PA.

Weiner K. S. et al. (2018). Defining the most probable location of the parahippocampal place area using cortex-based alignment and cross-validation. *NeuroImage*. Segmenting the Brain 170:373–384. ISSN: 1053-8119.

Weissbart H., Kandylaki K. D., Reichenbach T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience* 32(1):155–166. ISSN: 1530-8898.

Wenger M. J., Townsend J. T. (2006). On the costs and benefits of faces and words: Process characteristics of feature search in highly meaningful stimuli. *Journal of Experimental Psychology: Human Perception and Performance* 32(3):755–779. ISSN: 1939-1277(Electronic),0096-1523(Print).

Westner B. U., Dalal S. S. (2019). Faster than the brain's speed of light: Retinocortical interactions differ in high frequency activity when processing darks and lights. *bioRxiv*:153551.

Wheeler D. D. (1970). Processes in word recognition. *Cognitive Psychology* 1(1):59–85. ISSN: 00100285.

Whittington J. C. R., Bogacz R. (2017). An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Computation* 29(5):1229–1262. ISSN: 1530-888X.

— (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences* 23(3):Publisher: Elsevier, 235–250. ISSN: 1364-6613, 1879-307X.

Willems R. M., Frank S. L., Nijhof A. D., Hagoort P., Bosch A. van den (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex* 26(6):2506–2516. ISSN: 1047-3211.

Wolf T. et al. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*:arXiv: 1910.03771.

Woolnough O., Donos C., Rollo P. S., Forseth K. J., Lakretz Y., Crone N. E., Fischer-Baum S., Dehaene S., Tandon N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour* 5(3):Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Language;Reading Subject_term_id: language;reading, 389–398. ISSN: 2397-3374.

Yan M., Kliegl R., Shu H., Pan J., Zhou X. (2010). Parafoveal load of word N+1 modulates preprocessing effectiveness of word N+2 in Chinese reading. *Journal of Experimental Psychology. Human Perception and Performance* 36(6):1669–1676. ISSN: 1939-1277.

Yarkoni T., Poldrack R. A., Nichols T. E., Van Essen D. C., Wager T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods* 8(8):665–670. ISSN: 1548-7091.

Yeatman J. D., White A. L. (2021). Reading: The Confluence of Vision and Language. *Annual Review of Vision Science* 7(1):_eprint: https://doi.org/10.1146/annurev-vision-093019-113509, null.

Yi H. G., Leonard M. K., Chang E. F. (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron* 102(6):1096–1110. ISSN: 0896-6273.

Yu A. J., Cohen J. D. (2008). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems* 21:1873–1880. ISSN: 1049-5258.

Yuille A., Kersten D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in cognitive sciences* 10(7):301–308. ISSN: 1364-6613.

Zhou H., Friedman H. S., Heydt R. v. d. (2000). Coding of Border Ownership in Monkey Visual Cortex. *Journal of Neuroscience* 20(17):6594–6611. ISSN: 0270-6474, 1529-2401.

Zipser K., Lamme V. A. F., Schiller P. H. (1996). Contextual Modulation in Primary Visual Cortex. *Journal of Neuroscience* 16(22):7376–7389. ISSN: 0270-6474, 1529-2401.

Zwitserlood P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition* 32(1):25–64. ISSN: 0010-0277.