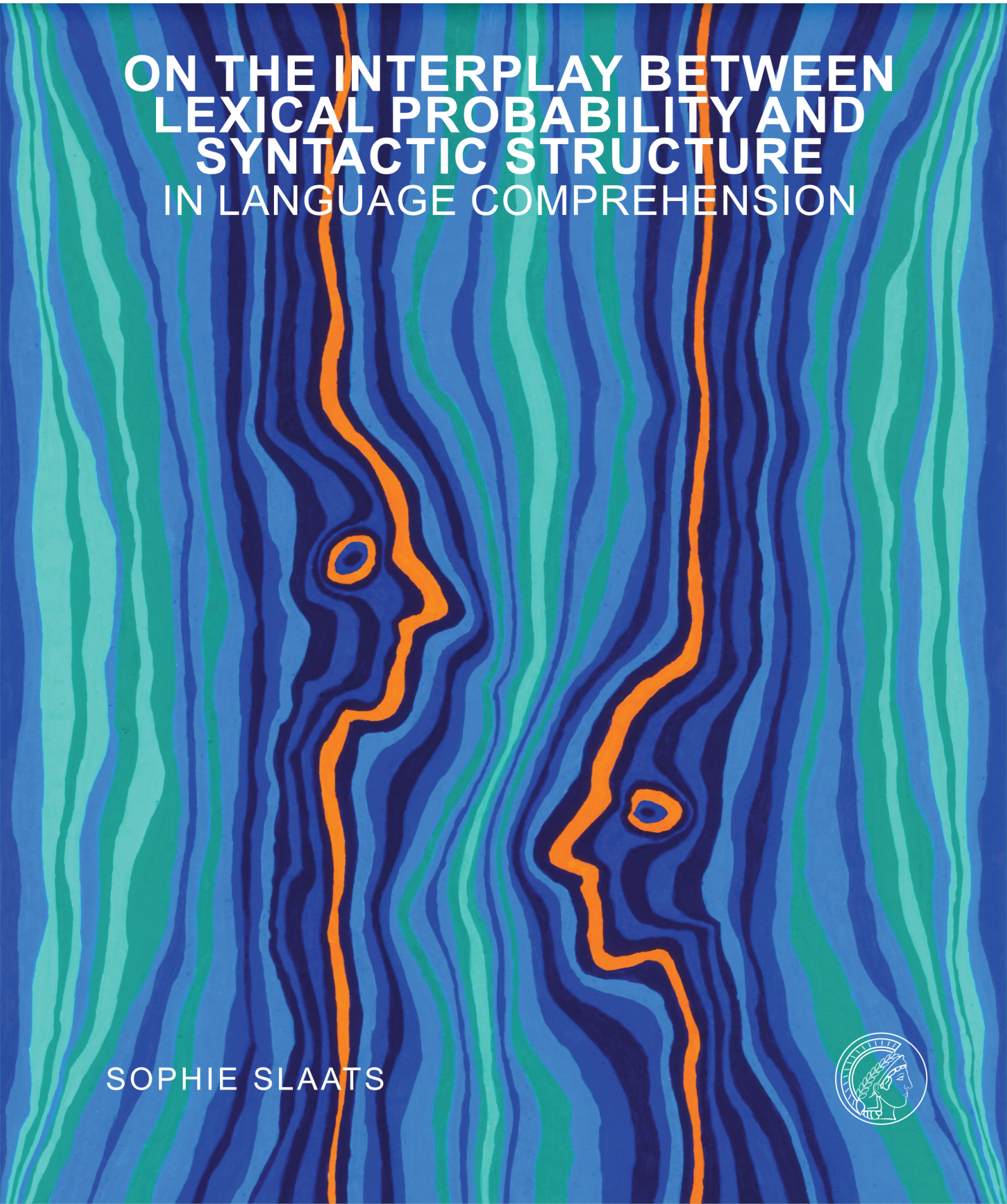
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

ON THE INTERPLAY BETWEEN LEXICAL PROBABILITY AND SYNTACTIC STRUCTURE IN LANGUAGE COMPREHENSION



SOPHIE SLAATS

On the interplay between lexical probability and syntactic structure in language comprehension

Funding body
This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en).

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-62-2
Cover design and lay-out by Sophie Slaats
Printed and bound by Ipskamp Drukkers, Enschede

 The research reported in this thesis was conducted at the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands

On the interplay between lexical probability and syntactic structure in language comprehension

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

maandag 16 september 2024
om 10.30 uur precies

door
Sophie Slaats
geboren op 29 december 1993
te Nijmegen

Promotor:

Prof. dr. A.S. Meyer

Copromotor:

Dr. A.E. Martin

Manuscriptcommissie:

Prof. dr. C.F. Rowland

Dr. ir. R. Oostenveld

Prof. dr. D. Poeppel (New York University, Verenigde Staten)

Contents

1 | General introduction

Our world is full of regularities. Big ones, such as the rising and setting of the sun, or the transitioning of the seasons, and infinitely many small ones, such as which leg a spider moves first, the ripples in a lake after a drop of rain falls in, and the rhythmic breath of your sleeping pet. Many animals are really good at learning different patterns and probabilities from the things they observe (Santolin & Saffran, 2018). Seals, for example, can recognize different rhythms (Verga, Sroka, Varola, Villanueva, & Ravignani, 2022), zebra finches learn the probability of the next note in their own song (Chen & ten Cate, 2015), and tamarin monkeys can learn a statistical relationship between two vowels, even when there is another sound in between (Newport, Hauser, Spaepen, & Aslin, 2004). Humans are extraordinarily good at learning about their environment. One aspect of human life that clearly displays this, is our ability to communicate our thoughts in a structured way: with language.

To produce and understand language, there are lots of rules and regularities that we must know, ranging from the correct intonation to signal a question, to the correct order of adjectives in a phrase like “the big white car”. The aspect of language that arguably best demonstrates our extraordinary human capacity for learning about our environment, is the syntax: our ability to combine morphemes, words and phrases such that the resulting combination carries a specific meaning. This ability is very powerful. Our words can be combined in infinitely many ways: we can understand very short sentences, and very long ones, and sentences that we have never heard before. At the same time, while we create phrases and sentences out of a sequence of words, the words themselves do not disappear. They remain available to us at all times. The human ability to create and analyze sentence structure is unprecedented. But how do we do it?

There are multiple views of which (neural) mechanisms underlie our capacity to form and understand sentences with a wide array of different structures. In this dissertation, I will focus on two specific ones. The first strand of research has focused on our ability to build syntactic structure as the result of learning and using sequential statistics, such as transitional probabilities between different units (e.g., Frank & Bod, 2011; Frank & Christiansen, 2018; Frost, Armstrong, &

Christiansen, 2019; McCauley & Christiansen, 2019). Upon this view, the use of abstract, hierarchical structure is not (always) necessary to understand what is being said or signed; we can make do with just the sequential statistics between phonemes, words, or phrases. The other has modeled the role of syntactic structure as the necessary inference of a separate level of representation that is hierarchically structured and abstracts away from the lexical items itself (Brennan & Hale, 2019; Everaert, Huybregts, Chomsky, Berwick, & Bolhuis, 2015; Lo, Tung, Ke, & Brennan, 2022; Martin, 2016, 2018; Matchin & Hickok, 2020); in other words, the phrase structure rules. Upon this view, knowledge of syntax is not tied to the specific words or morphemes. The rules operate over syntactic categories such as Noun, Verb, and higher-level categories like Noun Phrase (NP) and Verb Phrase (VP). The interpretation of the input crucially depends on the output of this rule-like system.

Despite the relative absence of reconciliation between these two views of syntactic structure, work in the fields of psycholinguistics and neurolinguistics has provided evidence for the involvement of both types of knowledge in the process of language comprehension. On the one hand, sequential statistical information such as the probability that one word follows another has been found to be a good predictor of the time needed to read a word (Aurnhammer & Frank, 2019; Lowder, Choi, Ferreira, & Henderson, 2018; Monsalve, Frank, & Vigliocco, 2012), but also of some measures of brain activity (Armeni, Willems, van den Bosch, & Schoffelen, 2019; Gillis, Vanthornhout, Simon, Francart, & Brodbeck, 2021; Nelson, Dehaene, Pallier, & Hale, 2017; Weissbart, Kandylaki, & Reichenbach, 2019). On the other hand, it has been shown that models of these types of data are better when some kind of information about the syntactic structure is included (Monte-Ordoño & Toro, 2017; Nelson, El Karoui, et al., 2017; Roark, Bachrach, Cardenas, & Pallier, 2009; Toro, Sinnett, & Soto-Faraco, 2011; van Schijndel & Schuler, 2015). In addition, changes in the structure of the input have been found to affect brain activity, as well: for example, neural recordings of a person listening to language change depending on whether they are listening to words, phrases or sentences (Bai, Meyer, & Martin, 2022; Coopmans, de Hoop, Hagoort, & Martin, 2022; Kaufeld, Bosker, et al., 2020; Ten Oever, Kaushik, & Martin, 2022).

When we consider all of these findings together, we must conclude that both distributional and abstract, hierarchical syntactic information play a role in language comprehension – and that both shape the way the brain responds to input. An adequate theory of language comprehension must therefore account for both

views: human brains are incredible probabilistic engines, and they are capable of producing hierarchical, abstract representations. In this dissertation, I approach language comprehension from such a perspective. I investigated how lexical distributional information, such as surprisal and word frequency, and syntactic information jointly shape the process of language comprehension. Investigating this question can help us to understand which mechanisms play a role in the brain's capacity to infer hierarchical structure from a highly variable sequential signal. I approach this question various ways.

In **Chapter 2**, I asked two main questions that surround the use of distributional information. Firstly, I asked why lexical surprisal – a metric of word probability given the preceding context - works well as a predictor for human behavioral and neural data. To explore this, I used simulation with a toy grammar and recurrent neural networks (RNN), varying both word frequency values and the grammar of the input language. Secondly, I asked how the results from studies that used lexical surprisal as a predictor can inform mechanistic theories of language comprehension.

In **Chapter 3**, I present results from an analysis project of magnetoencephalography (MEG) data. I investigated whether the presence of syntactic structure affects how low-frequency neural activity represents lexical information. I did this by extracting a purely lexical response from two different conditions: sentences and word lists (scrambled versions of the sentences). Using temporal response functions (TRFs), a multivariate linear regression approach, it is possible to model responses to different aspects of the stimulus simultaneously. This approach allowed me to disentangle signatures of lexical processing from other processes, such as the response to the acoustics of the stimulus. I modeled the response to lexical information with *word frequency*, the unigram probability of a word, and compared these responses between the sentence and word list conditions in sensor space and in source space. The results from this study speak to how top-down knowledge of the structure of a sentence affects lower-level (i.e., closer to the sensory input) processing – in this case, lexical processing.

In **Chapter 4**, I approached the interplay between lexical distributional information and syntactic information from another perspective: instead of investigating whether lexical information is processed differently given the availability of syntactic information, here, I investigated whether the probability of a word in context affects the use of syntactic information. This question is interesting from two perspectives. The first perspective is language comprehension as an instantiation of cue-based inference, in which statistical knowledge and syntactic

knowledge both function as cues. According to this perspective, the statistical probability of a word and grammatical knowledge of the receiver should both affect the process of comprehension. The second perspective is the recent view that surprisal from various statistical language models can capture all sorts of psycholinguistic effects. With this in mind, in this study I investigated whether lexical surprisal affects the computation of subject-verb agreement in an online self-paced reading paradigm. The results of this study provide insight into how distributional and morphosyntactic cues are weighted during language comprehension (specifically, reading).

In **Chapter 5**, I asked again whether lexical probabilistic information affects syntactic processing, but this time using the approach used in Chapter 3. In this study, I analyzed MEG data from a naturalistic listening paradigm: participants were listening to an audiobook in the scanner. This dataset was the result of a joint effort of the research group. We created several annotations of the audiobooks, among which a minimalist syntactic parse. Using TRF-models, I extracted responses to those syntactic annotations for words that were associated with high- or low surprisal values. Like in Chapter 3, I compared the resulting responses to each other. I repeated this procedure for two types of language models: a short-context trigram model, and GPT2, a model that captures long-context variability. The presence of any differences between the high- and low surprisal responses to syntactic annotations suggests that lexical probability affects the process of syntactic structure building.

Chapter 6 presents an overview of several sets of simulations that complement and inform the analyses presented in Chapters 3 and 5. The goal of the simulations was to assess whether any effects found in the analyses from Chapters 3 and 5 could be attributable to properties of either the data or the linear model that were unrelated to the theoretical phenomenon under consideration. These simulations help situate the interpretation of the findings presented in the thesis. In this Chapter, I specifically address the following four questions. (1) How does the interstimulus interval (ISI) affect the reconstruction accuracy of the TRF model? (2) If a feature enhances reconstruction accuracy in one frequency band, but not the other, does that mean that the response is in this frequency band? (3) Are different feature values able to extract the same TRF waveform? (4) Is the TRF suitable to model interactions between features *in time*?

In **Chapter 7**, I summarize the findings of the preceding Chapters, discuss the results from Chapters 3, 4, and 5 given the theoretical interpretation of surprisal in Chapter 2, and interpret these results in different theoretical frameworks that

leverage time in the computations underlying language comprehension. I end with a proposal for a computational model that has the potential to explain how statistical information informs the process of structure building.

2 | What's surprising about surprisal

Abstract

In the computational and experimental psycholinguistic literature, the mechanisms behind syntactic structure building (e.g., combining words into phrases and sentences) are the subject of considerable debate. Much experimental work has shown that surprisal is a good predictor of human behavioral and neural data. These findings have led some authors to model language comprehension in a purely probabilistic way. In this Chapter, we use simulation to exemplify why surprisal works so well to model human data, and to illustrate why exclusive reliance on it can be problematic for the development of mechanistic theories of language comprehension, particularly those with emphasis on meaning composition. Rather than arguing for the importance of structural or distributional information to the exclusion or exhaustion of the other, we argue more emphasis should be placed on understanding how the brain leverages both types of information (viz., statistical and structured). We propose that distributional information is an important *cue* to the structure in the message, but is not a substitute for the structure itself - neither computationally, formally, nor conceptually. Surprisal and other distributional metrics must play a key role as theoretical objects in any explanatory mechanistic theory of language processing, but that role remains in the service of the brain's goal of constructing structured meaning from sensory input.

2.1 Introduction

When we understand language, the task presented to the brain is to transform physical signals in the environment: from a continuous stream of speech or sign we perceive discrete words, and combine them into phrases and sentences to form a structured and meaningful message. How exactly we perceive phrases and sentences from words (and morphemes), is the subject of considerable debate. A wealth of recent experimental work has shown that lexical distributional information is an incredibly good predictor of both behavior (e.g., reading times) and neural activity (e.g., EEG recordings) during language comprehension (e.g., Aurnhammer & Frank, 2019; Frank, 2013; Gillis et al., 2021; Monsalve et al., 2012; Weissbart et al., 2019). At first blush, this state of affairs seems to necessitate that distributional information play a decisive role in comprehension; a conclusion drawn by many (among others, e.g., Armeni et al., 2019; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2022; Kuperberg & Jaeger, 2016). At the same time, there is ample evidence that human linguistic abilities go far beyond what even the largest probabilistic language models can do: we effortlessly interpret sentences that we never heard before – to wit, we understand just fine what was not likely, in fact, this is probably how we receive new information during conversation. As such, distributional information cannot be the whole story to language comprehension.

In this Chapter, we will outline how the current focus on distributional information can sometimes be at cross purposes with uncovering the mechanisms that explain how we understand. Starting from central assumptions in theoretical linguistics, here we use simulation to show that distributional metrics on their own (1) do not form a mechanistic account for the capacity we seek to explain, and (2) cannot straightforwardly be compared to theory-driven predictors in analysis due to distributional metrics effectively being the data, itself, through a filter. Nevertheless, we know that both structure and statistics matter to the brain, both writ large and in the minutiae of language processing (e.g., Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Nelson, El Karoui, et al., 2017; Weissbart et al., 2019). Thus, rather than arguing for the importance of one to the exclusion or exhaustion of the other, we argue that more emphasis should be placed on how the brain leverages both to reach understanding, and on how both shape processing. We propose that distributional information is an important *cue* to the structure in the message, but is not a substitute, functionally or otherwise, for the structure itself. Using surprisal and other distributional met-

rics as theoretical objects, and especially as *explanans*, lies at cross purposes with the goal of an explanatory and mechanistic theory of language comprehension.

2.2 How to describe a sequence

Humans, like other organisms, strive to reduce uncertainty in their environment by learning, and by anticipating incoming sensory input (e.g., Friston, 2012; Hasson, 2017). In the case of quasi-sequential sensory input, which language is (both signed and spoken), there are multiple ways to learn about what was just perceived and to anticipate what comes next. One way to learn about our environment is by counting occurrences of the sensory events, remembering their sequential order, and tracking how often a given event follows another (e.g., Aslin & Newport, 2014; Linzen, Siegelman, & Bogaerts, 2017; Saffran, Newport, & Aslin, 1996). This is *distributional information*.

2.2.1 Characteristics of probabilities (an introduction to information theory)

Distributional information for language comes in at least two flavors: *surprisal* and *entropy*. The two metrics are related, but differ in their predictive value, and have very distinct functional interpretations. Entropy is a measure of uncertainty about potential outcomes of future events, while surprisal is a post-hoc measure of event expectancy. In language research, surprisal and entropy are typically calculated over sequences of words.[1]

Surprisal is calculated by taking the negative log-transformation of probability information of the word. In many cases, this is the conditional probability of the word: the probability of word given the N previous words. If a word is very likely to appear given the context, surprisal is low; it is high when the word does not often appear in the given context. In information theory, surprisal is also called *(Shannon) Information* (I) (Shannon, 1948). This is directly tied to the term 'surprisal': if a word is very likely to appear, the amount of information gained is low.[2] Here, we will on occasion use I to denote surprisal. The equation of surprisal is shown in 2.1 below. As is shown in 2.2 below, *entropy* is the weighted sum over the surprisal values of all the words that could appear in the

[1] Other measures of surprisal and entropy exist, but the lexical estimate plays a prominent role in recent research.

[2] In psycholinguistics, this measure is often called *transitional probability* (TP). This term usually does not refer to surprisal obtained from neural network models.

position of the word in question. In other words, it is the *average surprisal* of all possible continuations; the expected amount of information for a continuation. Entropy depends both on the number of optional words, and on their probability distribution. If there are a lot of options, or if they all have the same probability of appearing, entropy will be high. On the other hand, if there are few options, or one of them has a much higher probability than the others (= lower surprisal!), entropy will be low. As such, entropy is a quantification of the uncertainty about what the transitional probability to the next word will be.

$$I\left(w_i|w_{i-n}...w_{i-1}\right) = -\log\left(p\left(w_i|w_{i-n}...w_{i-1}\right)\right) \tag{2.1}$$

$$H\left(w_i|w_{i-n}...w_{i-1}\right) = -\sum p\left(w_i|w_{i-n}...w_{i-1}\right)\log\left(p\left(w_i|w_{i-n}...w_{i-1}\right)\right) \tag{2.2}$$

2.2.2 The power of surprisal

Distributional measures like (lexical) surprisal and entropy have been shown to be unequivocally robust predictors of brain activity and behavior. For example, higher surprisal values and a larger decrease in entropy both tend to lead to slower reading times (Aurnhammer & Frank, 2019; Frank, 2013; Hale, 2006; Levy & Gibson, 2013; Linzen & Jaeger, 2016; Pimentel, Meister, Wilcox, Levy, & Cotterell, 2022; Smith & Levy, 2013). Corpus studies and computational models suggest that (backward) surprisal contains information about phrase structure (McCauley & Christiansen, 2019; Thompson & Newport, 2007). More recently, advances in neuroimaging and computational modeling alike have shown that oscillations in the delta, beta, and gamma bands show sensitivity to lexical surprisal (Gillis et al., 2021; Weissbart et al., 2019), that entropy reduction correlates with temporal lobe activity (Nelson, Dehaene, et al., 2017), and that surprisal and word frequency are tracked over and above acoustic and speech segmentation representations (Gillis et al., 2021). Furthermore, gamma power has been observed to increase when a word is highly predictable, but not to increase when it is not so predictable (Molinaro, Barraza, & Carreiras, 2013; Wang, Zhu, & Bastiaansen, 2012). In other words, surprisal and entropy are good predictors for behavioral and neurophysiological measurements – and the number of findings increases every day (see Figure 2.1 below).

But the power of distributional information does not stop there. Some effects that are attributed to linguistic structure can be evoked by statistical regularities

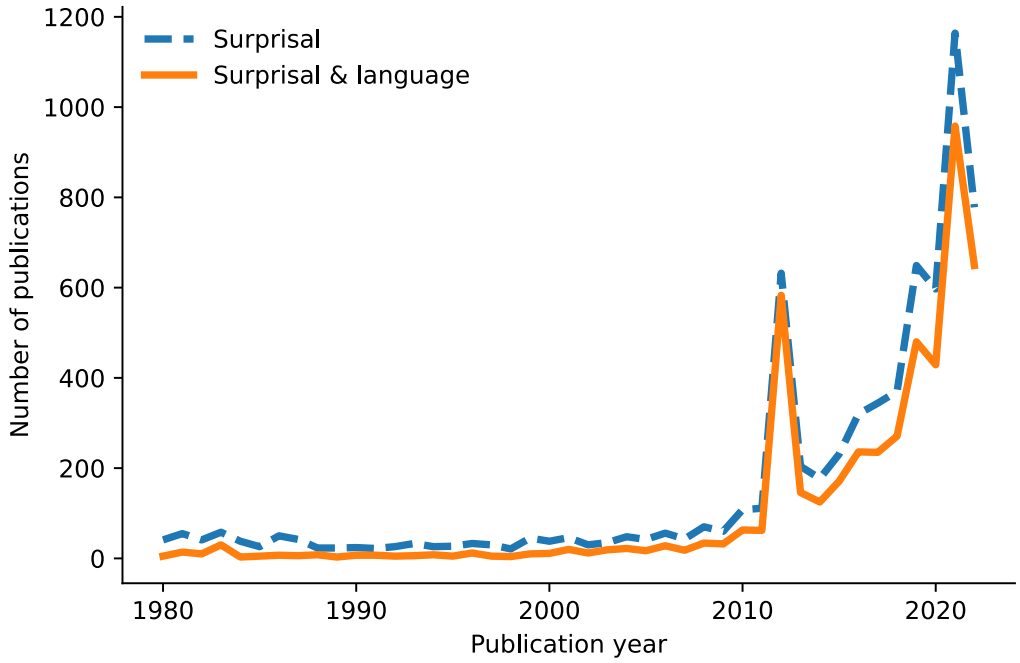


Figure 2.1: Number of publications between 1980 and 2022 that mention the keywords 'Surprisal' or 'Surprisal & Language', obtained from Dimensions.ai on 15/05/2023.

as well. In a seminal paper, Ding et al. (2016) showed that the occurrence rate of linguistic structures (syllables, phrases and sentences) in speech are reflected in power in the neural signal at the corresponding frequencies. This effect was originally suggested to reflect the construction of linguistic units: the brain encodes abstract linguistic information. Nevertheless, since its publication several studies have shown that the low-frequency frequency tagging effects can be induced by transitional probability information alone (Bai et al., 2022; Batterink & Paller, 2017).

Evidence for the importance of distributional patterns in language (prior to the surge of large language models, e.g., Aslin & Newport, 2012; Elman, 1991; Monsalve et al., 2012; Saffran, Aslin, & Newport, 1996) has led to accounts that model (aspects of) language comprehension using distributional information, such as surprisal theory (Hale, 2001, 2016; Levy, 2008a; Levy & Gibson, 2013) and entropy reduction theory (Hale, 2006). In these theories, surprisal and/or entropy reduction are minimally taken to be an estimate of processing effort: comprehenders make use of probabilistic knowledge to predict both the structure of the input they have just heard or seen, and what they may encounter next. Processing difficulty varies according to the deviations from these predictions.

Surprisal theory may not reject the notion of abstract syntactic structure (Hale, 2001, 2006; Levy, 2008a): instead, it is, in essence, agnostic about the representations and mechanisms that lead to structure-dependent interpretation, or language comprehension. Proponents of Surprisal Theory are explicit about this. E.g., Futrell, Gibson, and Levy (2020) posit:

> *"In addition to providing an intuitive information-theoretic and Bayesian view of language processing, surprisal theory has the theoretical advantage of being representation-agnostic: The surprisal of a word in its context gives the amount of work required to update a probability distribution over any latent structures that must be inferred from an utterance. [...] This representation-agnosticism is possible because the ultimate form of the processing cost function [...] depends only on the word and its context, and the latent representation literally does not enter into the equation."* (Futrell et al., 2020, p. 4)

By consequence, authors often refrain from drawing conclusions about the computational, algorithmic or implementational levels of language comprehension, which are not the focus of Surprisal Theory. There are some extensions of Surprisal Theory that do include specifications of the cognitive architecture underlying comprehension: Brouwer, Delogu, Venhuizen, and Crocker (2021) provide a probabilistic instantiation of a 'Retrieval-Integration Account', a model that contains explicit levels of representation and two mechanisms: 'retrieval', the use of a word form and the context to access word meaning; and 'integration', the mapping of the word meaning and the prior context onto a representation of the utterance. Another proposal that goes beyond the notion of modeling processing difficulty rather than the mechanisms underlying process of language comprehension, is the work by Frank and colleagues. This work uses surprisal to focus on (the cognitive reality of) representational levels during language comprehension. Such studies question the necessity of abstract representations and (hierarchical) syntactic structure during language comprehension (Frank, 2013; Frank & Bod, 2011; Frank, Bod, & Christiansen, 2012; Frank & Christiansen, 2018), but again do not offer an account of how surprisal values come to reflect comprehension nor of how particular meanings are perceived, but not others.

We highlight the representation-and-mechanism-agnosticism of distributional estimates, but our aim is to not criticize the literature around Surprisal Theory. Instead, we have two main objectives. Firstly, we wish to point the field of the

neurobiology of language to the issues that surround the use of distributional estimates that are representation-agnostic, both from a psychological and from a cognitive neuroscientific perspective. More specifically, these issues arise because we are trying to explain the process of language comprehension, rather than describe it, with a model. Secondly, by focusing on syntactic representations, we will show that results obtained using surprisal or entropy, because they are representation-agnostic, do not warrant conclusions about the latent factors driving the surprisal estimates (as in Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018).

The goal is not to discourage the use of distributional estimates in our field – in our view they are a core, crucial ingredient. In fact, in our models, distributional information derived from linguistic experience likely plays a role in shaping the process of language comprehension (e.g., Martin, 2016, 2020; Meyer, Sun, & Martin, 2020b; Slaats, Weissbart, Schoffelen, Meyer, & Martin, 2023). However, as we will see below, *despite* this role, the exclusive focus on representation-agnostic distributional information (no matter the level of representation or model used) may obscure the mechanism we need to explain how we understand.

2.2.3 Characteristics of structure: explaining compositional meaning

One of the crucial mechanisms that our field seeks to explain is our capacity for *syntax*. Syntactic information is a description of the abstract collocation, constituency or dependency relations, and domains over which functions apply in human language. This information, these patterns, are the result of a system that can be described by rules: the grammar of a language. Decades of research in theoretical linguistics has provided insights into aspects of the grammar that are necessary to explain human competence in language production and comprehension. Firstly, central to the grammar is that the rules are structure-dependent rather than item-dependent: the rules apply to grammatical categories ('parts-of-speech') and other rules, and not to the words themselves. Secondly, the structures generated by the grammar are *hierarchical*. This is a consequence of the observation that syntactic operations apply to *constituents* – (groups of) words that share a particular grammatical function – rather than to individual words (e.g. substitution: 'yesterday I saw [my best friend]' --> 'Yesterday I saw [her]'). In some cases, the same sequence of words can have more than one in-

terpretation depending on the hierarchical relation between the elements; e.g. '[old men] and women' vs. 'old [men and women]'; and 'Robin saw [the woman with the binoculars]' vs. 'Robin saw [the woman] with the binoculars'. These sequences are structurally ambiguous. Of course, only a subset of all possible word sequences is structurally ambiguous, but phrases are parsed hierarchically in all cases (Cinque, 2004; Coopmans, de Hoop, Kaushik, Hagoort, & Martin, 2021; Everaert et al., 2015; Jackendoff, 1972).

Assuming the presence of syntax – an abstract, structure-dependent rule system that applies hierarchically – in language provides an explanation for the striking linguistic capacity to generalize, produce, and understand. Our knowledge of syntax is what allows us to produce and understand combinations of words that we have never perceived together, or to create sentences with novel words (e.g., Gertner, Fisher, & Eisengart, 2006). Most importantly, however, syntax is one of the elements that determine the meaning of a sentence, another being the meanings of the individual words (*principle of compositionality*). This capacity distinguishes language from other perception-action systems and makes language behavior difficult to account for (see Martin (2020) and Everaert et al. (2015) for discussion). The study of (a computational theory of) syntax has a long history in formal linguistics, but *how* this capacity is realized in mind and brain, at the algorithmic and implementational level, is an answer the field still aims to find – or *should* aim to find.

2.3 Surprisal is a perfect descriptor, but not a mechanism

Using simulation, we show that lexical surprisal values can reflect variance that finds its origin in the capacity we seek to explain (syntax), but despite this, these values themselves do not encode syntactic structure. Instead, surprisal likely contains information that derives from this structure. See the supplementary materials at `https://osf.io/xp3r7/` and the code at `https://github.com/sslaats/surprisal` for the details of the used corpora and the model architectures. The statistical tests reported are two-tailed t-tests and Pearson's correlation. The simulations described below serve an exemplifying purpose. Using a different model architecture will likely lead to numerically different results.

2.3.1 Syntactic structure affects surprisal values

In the style of Elman (1991), we designed a phrase-structure grammar with 27 words and 4 parts-of-speech. We trained a long short-term memory model (LSTM) on the sentences generated with this grammar (see Figure 2.2 for an example). To study the effect of structure, we trained a second LSTM on the same sentences, but with the words scrambled within each sentence. This method of scrambling maintains word frequency estimates, word frequency per sentence, and sentence length, but removes all sentential structure. Both models were tested on sentences generated by the grammar that were not part of the training set.

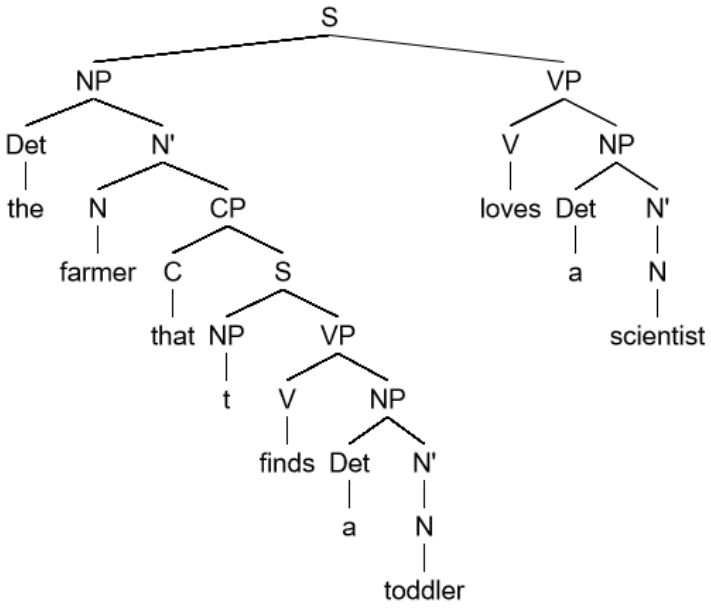


Figure 2.2: Example sentence generated using a small context-free grammar. The sentence reads: "the farmer that finds a toddler loves a scientist".

Compared to training on scrambled input, providing an LSTM model with structured input decreases surprisal values by 1.03 bit on average ($M_{struct} = 1.86$, $sd_{struct} = 0.99$; $M_{scram} = 2.88$, $sd_{scram} = 1.08$; $t(62854) = 124.08$, $p < 0.001$; CI = [1.01, 1.04], Cohen's $d = 0.98$, power = 1; Figure 2.3). This illustrates that syntactic structure in the input impacts surprisal values. Nevertheless, the distributions of surprisal values for the structured and scrambled models are highly correlated ($\rho(31426) = .92$, $p < 0.001$), suggesting that the frequency of each individual word – constant between training corpora – does the heavy lifting when it comes to surprisal estimation.

Next, we tested whether these findings scaled to a larger corpus with a wider lexicon and a variety of sentence types. To this end, we repeated the same procedure with a larger model, trained on 700.000 words from an English corpus (OpenSubtitles 2018; Lison & Tiedemann, 2016). Though smaller, here, too we observe a difference between the distributions ($M_{struct} = 6.42$, $sd_{struct} = 4.27$; $M_{scram} = 7.00$, $sd_{scram} = 3.68$; $t(160916) = 29.18$, $p < 0.001$; CI $= [0.54, 0.62]$, Cohen's $d = 0.15$, power $= 1$), with structured input decreasing the surprisal values by 0.58 bit on average (sd 1.82; Figure 2.4). Also here, however, we observed a high correlation ($\rho(80457) = 0.91$, $p < 0.001$) between the surprisal values estimated using the scrambled and structured model. In conclusion, syntactic structure affects the surprisal values of a large, naturalistic corpus in qualitatively similar ways to the effects observed in a small, constructed corpus. Thus, we conclude that (1) a decrease in surprisal is a general effect of the presence of language-like structure in sequences; and that (2) a large part of the variance in surprisal values stems from unigram probability information.

2.3.2 Surprisal does not lead to syntax

Can the underlying structure of a sequence (i.e., the latent syntactic structure that gives rise to the surface word order) be identified based on surprisal alone? To answer this question, we trained and tested two additional LSTMs on the Spanish translation of the OpenSubtitles corpus (structured and scrambled), and

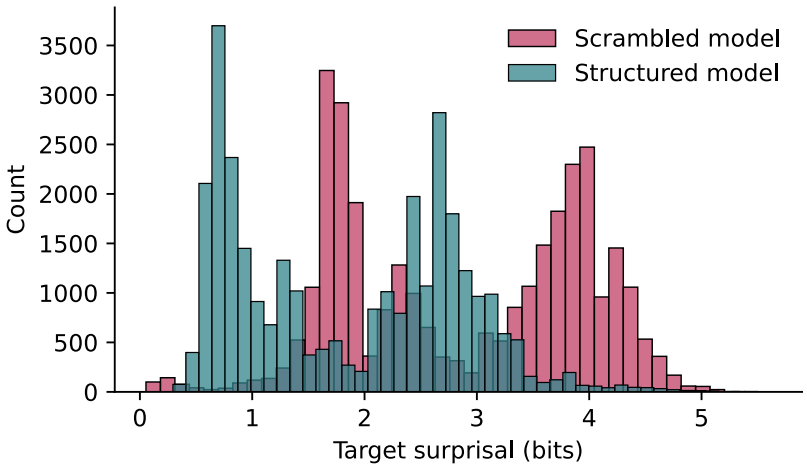


Figure 2.3: The presence of syntactic structure lowers the surprisal values of the words in sentences generated by a phrase structure grammar. Surprisal values for each word in the test set for both structured (teal) and scrambled (pink) models.

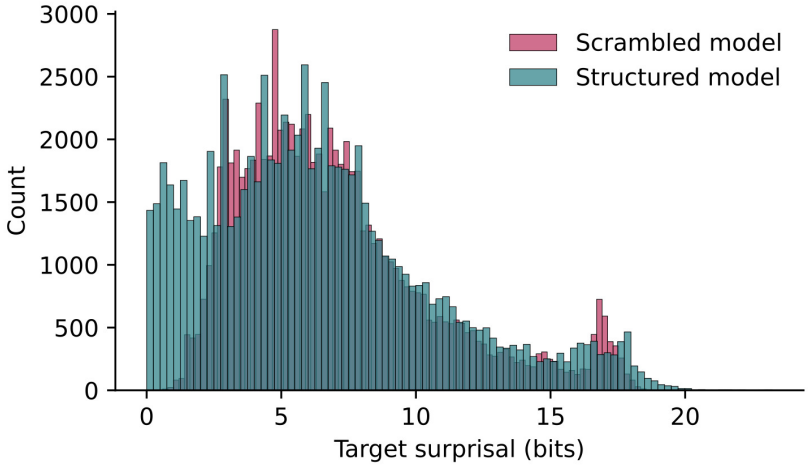


Figure 2.4: The presence of syntactic structure lowers the surprisal values of the words in sentences from the English OpenSubtitles 2018 corpus. Surprisal values for each word in the test set for both structured (teal) and scrambled (pink) models.

we used a Random Forest Classifier to classify whether (groups of) surprisal values were coming from English or Spanish. This procedure was repeated on the scrambled variants of the models.

The classifier performed above chance (structured: 63.8% (unigram) to 74.1% (10-gram)), indicating that surprisal values contain enough information for a classifier to distinguish between the two languages. However, this was also the case for the scrambled models (scrambled: 66.2% (unigram) to 84.2% (10-gram)), suggesting that structure is not the driving factor behind this above-chance performance. Instead, the classification appeared to be driven by uniqueness of surprisal values: each surprisal value was uniquely attributable to one or the other language. To remove this confound, we repeated the analysis with surprisal values rounded to the nearest 1 decimal. Doing so severely deteriorated the classification (structured 52.92% (unigram) to 67.98% (10-gram); scrambled: 58.46% (unigram) to 82.31% (10-gram)). In sum, (groups of) surprisal values contain enough information for a classifier to decide whether these values are coming from Spanish or from English, but this classification is crucially *not* dependent on structural properties of the language.[3]

Taken together, these two simulations paint the following picture. Lexical surprisal obtained from an LSTM encodes "structure" in the input, as is shown

[3]As stated by (Luce, 2003, p. 185), "the elements of choice in information theory are absolutely neutral and lack any internal structure; the probabilities are on a pure, unstructured set whose elements are functionally interchangeable."

by a general decrease in surprisal. Nevertheless, patterns of surprisal values stemming from two structurally different languages are not classifiable on the basis of structural properties. This suggests that (patterns in) surprisal values can capture regularity in language in general, but do not encode language-specific aspects of structure.

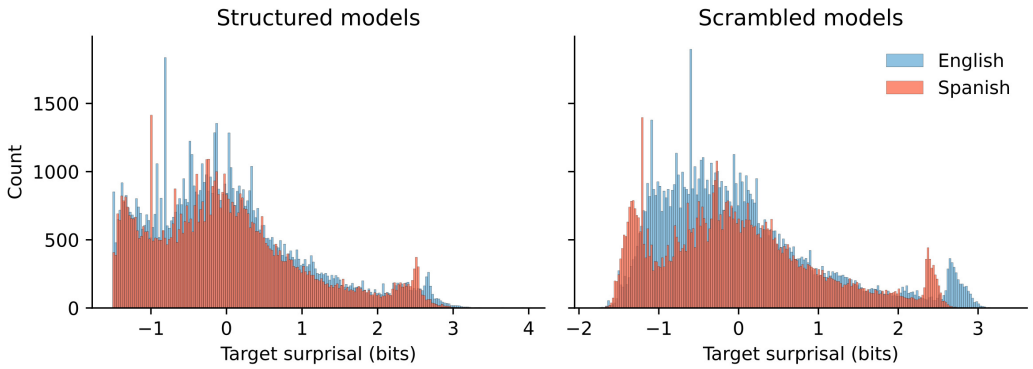


Figure 2.5: Z-scored surprisal values from the structured and scrambled models (English and Spanish). Observe the high peak in distribution in both languages; these are surprisal values for words that most often appear at the start of a sentence (Spanish: 'no'; English: 'I')

2.3.3 Understanding does not depend on surprisal

Consider the following Dutch sentences in (1). In each case, the verb *lag* ('lay') is singular, hence asking for a singular subject to the sentence. The subject is different for every sentence: it is either semantically different (man or goldfish), or it is morpho-syntactically different (singular or plural). The subjects are printed in boldface.

(1) a. Er lag een **goudvis** op straat (I = 15.4 bits)
There lay a goldfish$_{[sin]}$ in street

'A goldfish$_{[sin]}$ was lying in the street'

b. Er lag een **man** op straat (I = 7.1 bits)
There lay a man in street

'A man was lying in the street'

c. *Er lag een **goudvissen** op straat (I = 21.9 bits)
There lay a goldfish$_{[pl]}$ in street

'A goldfish$_{[pl]}$ was lying in the street'

d. *Er lag een **mannen** op straat (I = 14.9 bits)
There lay a men in street

'A men was lying in the street'

For every sentence, the surprisal value of the subject was calculated using a trigram model created with SRILM (Stolcke, 2002) trained on the Dutch OpenSubtitles 2018 corpus (Lison & Tiedemann, 2016). In other words, the surprisal value captures the surprisal of the subject in the context of the words 'lag', 'een' – both requiring the next word to be a singular noun. The first observation is that the surprisal of the subject 'goldfish$_{[sin]}$' is higher than the surprisal of the subject 'man': 15.4 and 7.1 bits, respectively. This is expected: 'man' is more frequent. If we change 'man' to its plural 'men', or 'goldfish' to its plural, the sentences become erroneous. In this case, surprisal value increases with approximately 7 bits, as well.

What is striking about this example is that the surprisal of the plural subject 'men' is similar to that of 'goldfish$_{[sin]}$' – despite the plural subject rendering the sentences grammatically incorrect and, as such, uninterpretable without repair. Although we may not often encounter a goldfish on the street, this sentence is perfectly intelligible. This illustrates a crucial feature of language: understanding does not depend on surprisal (See also: van Schijndel & Linzen, 2021).[4]

This highlights a crucial gap between distributional estimates such as surprisal and entropy and what *understanding* logically entails: reaching a single, interpretable representation of the input. Meaning can only arise when a stable representation has been formed. After all, probability distributions are not intelligible. Instead, the brain needs to converge on a discrete representation of the elements and the structure to reach understanding. Even if a large part of signal processing is probabilistic, at some point the brain has to 'decide' or converge on a stable interpretation of what we are hearing, reading, or seeing. In fact, this is

[4]Values obtained from neural network models show the same pattern. Here, we use English sentences, and score them using the model from the English simulations described above (context of 10 words). Surprisal of **men* is lower than perfectly intelligible *kite*.

a. On the street lies a **kite** (I = 14.46 bits)

b. On the street lies a **man** (I = 2.94 bits)

c. *On the street lies a **kites** (I = 16.45 bits)

d. *On the street lies a **men** (I = 7.88 bits)

More specifically, it shows that probability and grammaticality are theoretically distinct: the grammaticality of a sentence can change without affecting the surprisal values.

of one of the brain's main features: it can take in probabilistic information and map it onto deterministic representations.[5]

To summarize, surprisal values are sensitive to structure in the input, but they do not uniquely capture the structure that generated the sequence, or grammatical well-formedness. These observations lead us to conclude that distributional metrics are not suitable as *explanans* for the core capacities of language.

2.4 How surprisal can obscure the view

Several studies that have used both predictors of syntactic structure and surprisal in their models have drawn conclusions similar to those outlined above: syntactic structure is necessary to create the best model of the data (Brennan & Hale, 2019; Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Kapteijns & Hintz, 2021; Nelson, El Karoui, et al., 2017). Nevertheless, these studies report that distributional estimates, such as surprisal, explain more of the variance in the signal than do the syntactic predictors. The reason for this finding is that surprisal (even lexical) reflects variance from many latent factors. This is a consequence of the fact that surprisal estimates depend fully on the identity of the unit estimated. If we want to estimate the surprisal of a word, we need to know the identity of that word. This is the representation-agnostic character mentioned in the introduction: surprisal can – and will – parametrically reflect variation stemming from *any* domain or representational level of language, including syntax.

[5]This statement is closely tied in with the question what mental representations of abstract concepts are like – and whether (large) language models "have" abstract representations. They do, in fact, not, as was already pointed out by Fodor and Pylyshyn (1988). For example, our original toy grammar chose a word with the correct part-of-speech 78% of the time. While this means that the model has learned something about the regularities of the language, this does *not* mean that the model has learned what a noun is. More sophisticated distributional representations, such as the internal states of neural networks, can be called 'abstract' in the sense that a given pattern does not directly correspond to any individual item (e.g., when number of dimensions is reduced to below the number of types in the input). This does not mean that the model has an abstract representation, however. For example, two synonyms may have the same vector in a word2vec model, but that does not mean that the word2vec model understands 'synonymy'; the model only represents that these words are likely to occur in the same contexts, statistically speaking (and this disregards the fact that some synonyms can be used in different contexts – such as different registers –, despite having the same referential meaning!). If the internal states of a sophisticated distributional model reveal a pattern that is consistent across all words that are synonyms of another word, then we could say that this model *statistically approaches* the concept 'synonymy'. Nevertheless, this representation does not exist if we do not provide the model with a synonym. For humans, on the other hand, the concept 'synonymy' persists in the absence of synonyms. The same holds for other subconscious knowledge of language, such as what a noun is, or what plurality means.

We demonstrate this by changing our toy grammar in two ways: by adapting the grammar itself, or by changing word frequency. To edit the grammar, we changed the order of the constituents in verb phrases. The complement (a noun phrase or a complementizer phrase) now precedes the verb. In other words, we have changed the grammar from "SVO" (subject-verb-object) to "SOV" (subject-object-verb); see Figure 2.6 for an example sentence. Like in the scrambling models, doing so preserves the word frequency values as well as the number of words per sentence, but drastically changes the structure of the language. The model trained on this SOV-language was subsequently tested on the exact same test set as the previous models (structured and scrambled).

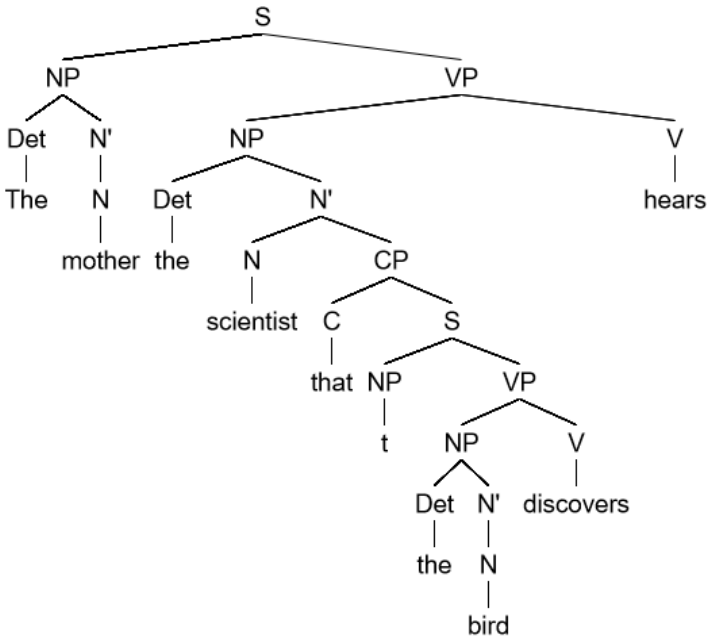


Figure 2.6: Example sentence generated using a small context-free grammar in which the constituent structure was changed to SOV. The sentence reads: "the mother the scientist that the bird discovers hears." In the original subject-verb-object structure, that would be: "the mother hears the scientist that discovers the bird."

Again, the resulting surprisal values from this SVO-trained model were significantly higher than those obtained using the original structured model (see Figure 2.7)[6] ($M_{SVO} = 1.86$, $sd_{SVO} = 0.99$; $M_{SOV} = 3.29$, $sd_{SOV} = 2.73$; $t(62854) = 87.79$, $p < 0.001$; CI = [1.41, 1.47], Cohen's $d = 0.7$, power = 1) – unsurprising, because a large number of word-to-word transitions that were present in the test set were definitely *not* present in the training set by virtue of being

[6]Notice that this SOV model is *also* structured.

ungrammatical.[7] The correlation between the results from the structured model and the SOV-model was lower, but nevertheless still there ($\rho(31426) = 0.44$, $p < 0.001$), indicating that unigram probability drives the pattern of surprisal values to an important extent.

Instead of changing the syntax, we can also change the word frequency values.[8] To do this, the word frequency parameters were changed for a few words in the original SVO grammar. Specifically, the frequency of the words 'woman', 'discovers', and 'a' was adjusted to be twice as high as the other words in their syntactic category.[9] We then tested this model on the same test set again, and indeed: there was a significant difference between these distributions ($M_{SVO} = 1.86$, $sd_{SVO} = 0.99$; $M_{WF} = 1.91$, $sd_{WF} = 0.97$; $t(62854) = 6.78$, $p < 0.001$; CI = [0.04, 0.07], Cohen's $d = 0.05$, power = 1), while the correlation between the original- and the word frequency adjusted estimates was still high ($\rho(31426) = 0.84$; $p < 0.001$); see Figure 2.8.

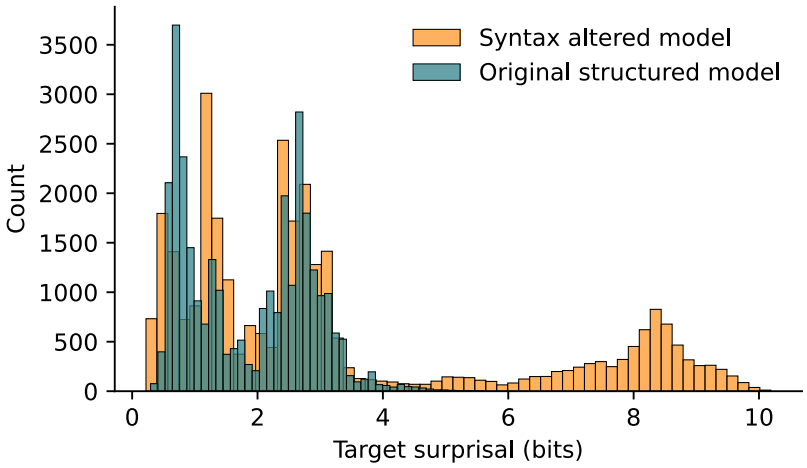


Figure 2.7: A different syntactic structure in the training input increases the surprisal values of the words in sentences generated by a phrase structure grammar. Surprisal values for each word in the test set for both SVO (teal) and SOV (yellow) models.

Indeed, both of these changes to the input/output relation – either the syntactic structure underlying the word sequences, or the frequency with which a word in a given category is selected – change the surprisal values we observe.

[7] This was crucially *not* the case in the scrambled model; any word-to-word transition was possible.

[8] Obviously, word frequency itself is a distributional variable. This toy grammar does not have a semantics, pragmatics, or any other linguistic parameter that we can tweak; here, the word frequency parameter is causal for word frequency within a category. See the supplementary materials for details.

[9] Essentially, this means that the lexical entropy in the training corpus is lower.

Logically, beyond syntax, lexical surprisal values can reflect variability from all kinds of sources: lexical category, syntactic structure, the pragmatic- and semantic context, priming, and so on (viz., any variable that affects word choice or word form). More than anything else, surprisal is a reflection of language data – which is also what we provide to our participants by presenting stimuli. Surprisal values are calculated by passing the data through a distributional filter, and thus are a prism or reflection of the data itself.[10]

Expressing the data through a distributional filter without being able to tease apart which variables in the data are contributing to a particular estimate of surprisal poses a problem when we want to use surprisal in combination with our readouts (e.g., behavioral, neural) to theorize about the cognitive architecture of the human language system. This is because while using a predictor that is derived from the data passed through a distributional filter may make it the most powerful predictor, it relegates the reasons why it is such a good predictor difficult to impossible to interpret: where do the effects come from? What linguistic factors create dynamics in surprisal? These observations should and do have important consequences for the way we develop our psycholinguistic theories, and for how surprisal comes into play in them. In short, surprisal does

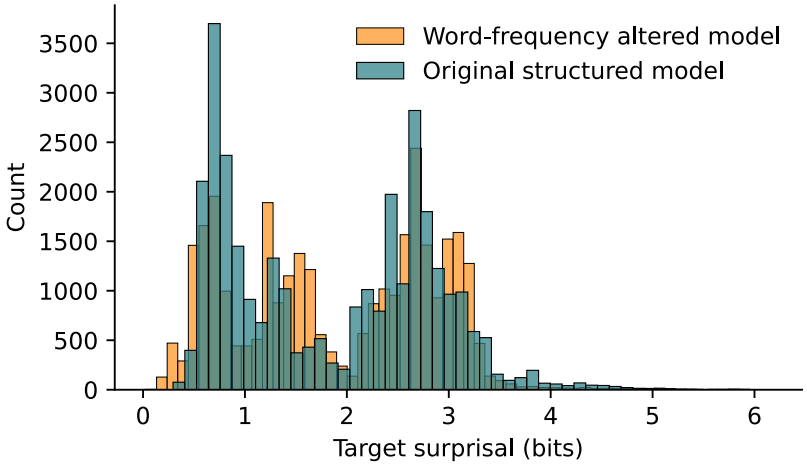


Figure 2.8: Surprisal values for each word in the test set from the corpora obtained with the original structured model and a model trained on a corpus that was adjusted for word frequency. Surprisal values for each word in the test set for both original (teal) and word-frequency altered (yellow) models.

[10] Devlin (2001, p. 21), cited in (Luce, 2003, p. 183): "Shannon's theory does not deal with 'information' as that word is generally understood. Instead, it deals with data-the raw material out of which information is obtained."

not allow us to draw conclusions about the potential mechanisms at work, e.g., in composition, even when its ability to predict the readout is robust.[11]

2.5 Theory is always wrong

Syntactic predictors in statistical models of reaction times or neural data are at a disadvantage even before running the model: the way they are constructed depends on a *theory of syntax* and of how this structure is implemented in the brain (Martin, 2020). Where surprisal values follow naturally from counting large numbers of words in a corpus,[12] to create a syntactic predictor one must choose between numerous syntactic theories (for example, a minimalist or constructivist approach), make assumptions about the parsing strategy the brain employs to reach this structure (for example, 'left-corner' or 'top-down'; Brennan et al., 2016), and, finally, assume that the brain does not make any errors when parsing the input, for example when encountering ambiguous sentences (a so-called 'perfect oracle').

Any of these theory-driven assumptions will change the predictive power of the syntactic feature. This is not a problem in and of itself, but this becomes a problem when comparing it to a data-driven feature like surprisal. A data-driven estimate will perform better than a theory-driven estimate – because the data do not err, the theorizer does (Guest & Martin, 2021). These errors, or rectifications to those errors (when a theory-driven change to a feature affects the goodness of fit, or when an explicit experimental manipulation has a certain effect on responses) provide an opportunity to adjust our theory. By contrast, using surprisal as an explanation prevents us from looking at the influence from latent factors by reflecting variance that stems *from* these factors as a second-order variable. In this way, combined with representation-agnosticism, exclusive focus on surprisal as a predictor serves to obscure the mechanisms that explain behavioral and neural responses, and the view on the casual structures of human languages that shape its instantiation in the mind and brain.

[11] For a discussion of whether large neural networks can be a "mechanistic account" of cognition, see Guest and Martin (2023).

[12] This may or may not be a simplification of what deep learning models are doing; see Carlini et al. (2021).

2.5.1 Pax grammatica

As discussed in section 2.2.3, the meaning of a linguistic expression is a function of the meaning of the individual words and the way they are combined. If the goal of the field is to understand how we understand, we must have a mechanistic account of this process, not only a predictive one. We need to understand what the relevant representations of the input are, how the brain represents them, and how the brain moves from one type of representation to another. To be more precise, we want to know (among other things), (if and) how our brain stores and accesses (quasi)lexical items, how these (quasi)lexical items are combined using knowledge of syntactic and semantic structure, (if and) how the representation of the 'words' is separable from syntactic structure, and how the brain represents the "end product" (the meaning of the utterance).

Surprisal, or the principles underlying the calculation of surprisal – alone - cannot serve as a mechanism for language comprehension, but the factors that give rise to surprisal effects likely play an outsized and valuable role in the building of a mechanistic theory of language comprehension. Consider Figure 2.9 below. When the brain moves from representing the words in the bottom row to phrases and sentences in the rows above, *something* is done to the information. The probabilistic relation between the words 'the, 'train' and 'arrived', represented by the red arrows, may carry (non-deterministic) information about how the words are structurally related in the phrase or sentence – for example, if the surprisal of 'arrived' is relatively high, this can signal the beginning of a new phrase (see Martin, 2016). But what *mechanism* is responsible for constructing the phrase (there are different proposals for this process, ranging from 'construction' to 'merge'), and (how) can this mechanism be implemented by the brain? [13]

Despite the arguments laid out in this Chapter, the notion that surprisal is not equivalent to structural information, composition, or comprehension, is not at all in conflict with the use of distributional information in sensory processing. In fact, that humans can make use of distributional information in their environment is undeniable: in the absence of any other sources of information such as meaning, prosody, or knowledge of linguistic structure, we are capable of using statistical information to segment the input stream, and, as such, 'break in' to the realm of language (Aslin & Newport, 2012; Aslin, Saffran, & Newport,

[13] Obviously, the mechanism cannot be *prediction*. Firstly, we are perfectly capable of understanding sentences where some words have high surprisal (i.e., where the language model was wrong!); and secondly, and most importantly, predicting an item does not entail understanding of the input. See e.g. Huettig and Mani (2016).

1998; Batterink & Paller, 2019; Newport et al., 2004; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Thompson & Newport, 2007; Trecca et al., 2019).

2.5.2 Toy model

In the process of comprehension, distributional information can play a similar role. Such an architecture could work as follows. For the purposes of this toy model, we assume a set of representational levels (not exhaustive by any means): phonetic, phonemic, syllabic, lexical, phrasal, sentential, and discourse-level. There is much evidence that linguistic levels of representation are separable in brain and behavior (e.g., Bai et al., 2022; Gwilliams, Linzen, Poeppel, & Marantz, 2018; Kaufeld, Bosker, et al., 2020; Krauska & Lau, 2023; Leonard & Chang, 2014; Marslen-Wilson & Tyler, 2007; Mesgarani, Cheung, Johnson, & Chang, 2014; Slaats et al., 2023; Ten Oever, Carta, Kaufeld, & Martin, 2022; Tezcan, Weissbart, & Martin, 2023). For simplicity, we will focus only on how the brain performs *combination* (Pylkkänen, 2019) of units at a given level to infer the next level of representation. Within a given level of representation, then, we assume that the brain represents a sequence of units within this level of representation, in the spirit of either working memory, resonance (Jafarian & De Persis, 2015), attractors (Pascanu & Jaeger, 2011), or by retaining activation corresponding to

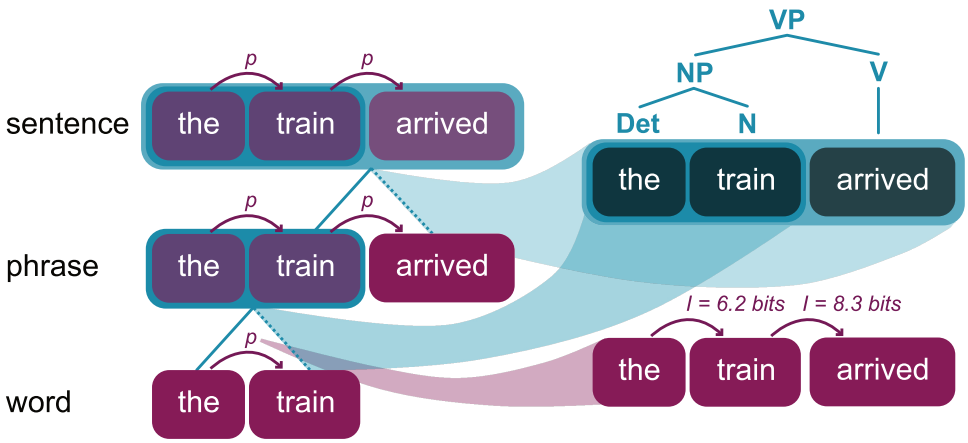


Figure 2.9: Schematic representation of the process of syntactic inference (left) and the differences between the outcome of this process (top right) and the outcome of lexical distributional information (bottom right). The surprisal values are fictional.

a unit of representation (Gwilliams, Poeppel, Marantz, & Linzen, 2018). In other words, previous input shapes the system in some way: either we retain several words prior to the processing current word, or the processing of previous words has changed the ongoing dynamics of the system such that the processing of the current word is shaped by these dynamics.

We propose that the brain represents probabilistic information *within* and *across* levels (viz., sequentially between phonemes, between words, between phrases, and bi-directionally between phonemes and words, words and phrases, etcetera) through *path dependence.* Path dependence means that the set of possible trajectories through state space is delimited by past trajectories (see Guest & Martin, 2021; Martin, 2020). By representing a sequence of units in this way, we gain access to transitional information, as well as information about how a specific unit relates to the level above; e.g., how likely it is for /χ/ to be the start of a word. Extrapolating crucial effects from acquisition research (Aslin et al., 1998; Saffran, 2001; Thompson & Newport, 2007), we assume that we remain capable of using this probabilistic information to infer the underlying structure that gave rise to the sequence – a.k.a., to transform the information into the next level of representation (see Martin (2020) for pseudocode). This could mean the use of phonemic information to infer words (e.g., Tezcan et al., 2023), or words to infer phonemic information (e.g., Martin, Monahan, & Samuel, 2017), or words to infer phrase structure or vice versa (Baese-Berk, Dilley, Henry, Vinke, & Banzina, 2019; Bai et al., 2022; Marslen-Wilson & Tyler, 1980; Thompson & Newport, 2007; van Alphen & McQueen, 2001).

What is crucial about this toy model is that probabilistic information is a factor, but it does not correspond to the representational levels themselves, nor to the computations that underlie the transformation of information that will eventually lead to a structured representation of the inferred meaning. Nevertheless, making use of probabilistic information as a *cue* to the level above (and, potentially, below; see Martin (2016) and Marslen-Wilson and Welsh (1978)) implies that these measures will affect the computations that do lead to the transformation of information, such as the inference of a phrase from words. Most of the findings in the domain of probabilistic information suggest effects of time, with most notably high surprisal being associated with longer reading times (Aurnhammer & Frank, 2019; Brothers & Kuperberg, 2021; Frank & Bod, 2011; Kapteijns & Hintz, 2021; Luke & Christianson, 2016; Monsalve et al., 2012), slower word recognition (Balling & Baayen, 2012), and so on. These findings suggest that probabilistic information may be a temporal modulator in the pro-

cess of comprehension, affecting the time-course of the computations that lead to comprehension.

Rather than being the *mechanism* that leads to comprehension of what was said, we propose that lexical surprisal is a *cue* to detecting the presence and absence of phrase boundaries – much like how transitional probabilities were viewed by Saffran, Newport, and Aslin (1996). Indeed, as mentioned in section 2.2.2, (backward) surprisal contains information about phrase structure (McCauley & Christiansen, 2019; Thompson & Newport, 2007). This means that the surprisal of a given word can provide the recipient information about whether a phrase boundary is likely there, similar to prosody, co-articulation, and others (Martin, 2016, 2020),[14] and is not an index of the process of composition itself. From a practical perspective, this suggests we should study interactions between (lexical) probabilities and syntactic operations rather than modelling surprisal and entropy as main effects – not only in behavior (e.g., Fine, Jaeger, Farmer, & Qian, 2013) but also in neuroimaging.

2.5.3 Open questions

For a human (brain) to be sensitive to a probabilistic estimate of any linguistic representation, this representation first needs to come into mental existence. What the nature of this representation is, and how it is inferred, are the difficult questions our field should aim to answer. The role of distributional cues is a part of the answer, and much about this aspect of the process is unknown. There is no consensus with respect to the nature of statistical information in the brain (how does the brain represent probability and/or uncertainty in general terms?), nor is it known what the brain 'computes' statistical information over (which representations are probabilistic, and why?), and what mechanism is responsible for this 'computation' (how does the brain keep track of probability?). In the field of language specifically, and cognitive science more generally, an important question is how the brain is capable to bootstrap structure (such as syntax) from statistics on the one hand. The reverse question is also open: does the brain re-

[14] The status of *cue* rather than *mechanism* should hold at other levels of representation, too (e.g., syllables, phonemes, phrases). Distributional information describes behavioral and neuroimaging data at all levels, ranging from the lower-level phonemic surprisal all the way to the higher-level syntactic estimates of surprisal (Brennan & Hale, 2019; Di Liberto, O'Sullivan, & Lalor, 2015; Heilbron et al., 2022). Being distributional estimates, the same problem holds: they prevent us from studying the underlying mechanism. For example, when recognizing words, the probabilistic relation between phonemes can provide information about where the input stream should be segmented into words – but it does not tell us how segmentation works at a mechanistic level, let alone how we recognize those segments as words.

fine probabilistic representations with structured knowledge, and if so, how does this work? We urge the field to consider these questions when using surprisal as a predictor for linguistic data.

2.6 Conclusion

In this Chapter, we have argued that the current focus on distributional information in the psycholinguistic and neurolinguistic literature may prevent us from uncovering the mechanisms we need to explain how we understand in two ways. Firstly, we have shown that distributional and syntactic information are functionally inseparable, because distributional information derives (in part) from syntactic information. Secondly, we have shown that being a filter on many sources of linguistic information, makes surprisal, entropy, and other distributional estimates robust and reliable predictors of human data. Nonetheless, distributional metrics are not suitable as *explanans* for the core capacities of language. Adopting surprisal and other distributional metrics as theoretical objects may instead distract us from the goal of an explanatory and mechanistic theory of language comprehension. We instead propose to view distributional information as a cue to the next level of abstraction; an aide to the mechanisms that we aim to uncover.

2.7 Appendix

Methodological information for simulations. All code is available on `https://github.com/sslaats/surprisal`.

2.7.1 Toy grammar simulations

The corpus was generated using a miniature phrase-structure grammar with four parts-of-speech: verbs (V), nouns (N), determiners (Det) and complementizers (Comp). The rules are displayed in (1).

(1) Phrase-structure rules

a. S --> NP VP

b. CP --> Comp S

c. NP --> Det N'

d. N' --> N

e. N' --> N CP

f. VP --> V NP

g. VP --> V CP

Using a small vocabulary of 27 words (see table 2.1 below), we generated a corpus of 10.000 sentences. The number of subordinate clauses was restricted to 5 irrespective of their binding position to avoid unrealistically long sentences and, more practically, an infinite loop.

Table 2.1: Vocabulary used for the simulations.

Part of speech	Words
Complementizer	*that*
Determiner	*a, the*
Noun	*woman, dog, goat, president, bird, colleague, mother, toddler, scientist, child, farmer, painter, cat*
Verb	*loves, discovers, reveals, notices, assumes, indicates, finds, senses, guarantees, teaches, hears, understands*

Scripts used for these simulations:

- **grammar.py**: specifies the toy grammar (phrase-structure rules)
- **simulate-corpus.py**: uses the grammar to generate n sentences for training of the LSTM model

- **train-model.py**: train model on toy grammar
- **train-random-model.py**: train model on scrambled output of toy grammar
- **test-model.py**: test model trained on toy grammar
- **compare-models.py**: compares the surprisal values on the test set between scrambled and structured models
- **language.csv**: the vocabulary & POS to use for simulate-corpus.py

The model weights for these simulations are in the subfolder 'model-weights'; the training corpora are in the folder 'corpora'. Surprisal values for the test sets are in the folder 'results'. All on `https://osf.io/xp3r7/`.

2.7.1.1 Simulation: 'syntax leads to surprisal'

Model training & testing We split the corpus into a training- and testing set with a ratio of 80/20, and used the training set to train a recurrent neural network model with an embedding layer of 10 nodes, a hidden LSTM layer of 64 nodes, and a linear layer mapping back to the word space. The learning rate was 0.1 and we used negative log likelihood loss as implemented in PyTorch (Paszke et al., 2019).

This yielded a 'structured model'; the input to the LSTM model was generated by a grammar. We also created a 'scrambled model'. To create the scrambled training set, we randomly shuffled the words within each sentence from the training set. This method preserves word frequency across the corpus, individual sentence length, and the number of words from a certain part-of-speech in each sentence, but removed all structure. We then trained an LSTM model with the same architecture as the structured model to obtain the scrambled model. For both the structured and the scrambled model, the input required a 10-word context, meaning that we extracted 10grams for every sentence prior to training.

We estimated surprisal values for every word in the test set using the scrambled and the structured model. The test set was identical in both cases (the output from the grammar).

Results As mentioned in the main text and shown in Figure 2.3, one can clearly see that providing the LSTM model with structured input (the blue bars) decreases surprisal values by 1.06 bit on average ($t = 127.08$, $p < 0.001$). This

clearly shows that surprisal values can reflect syntactic structure. Nevertheless, the correlation between the surprisal values from the scrambled and structured models is 0.92 ($p < 0.001$).

Model predictions The random model predicts the correct word approximately 16% of the time. The model defaults to predicting determiners with the occasional complementizer; these are the most frequent words in the corpus, and will therefore most often be correct. These two categories make up 45% of the total corpus, and there are three options: 'the', 'a', and 'that'. Out of these 45%, the network is correct at chance; $1/3^{rd}$ of the time. This yields 15% correct – so the model essentially performs at chance. The same is the case for the accuracy in part-of-speech; there are four options, and the model predicts the correct part-of-speech 27% of the time. For the structured model we see a slightly different pattern. The model predicts the correct word in approximately 28% of the cases. This model also defaults to a small set of words (noun = 'scientist', sometimes 'child'; verb = 'finds'), but these words match the correct part-of-speech 78% of the time; most of the failures are in the complement of the VP or NP, where a complementizer or a determiner are both good continuations of the sentence.

2.7.1.2 Simulation: 'surprisal obscures the view', syntax

To edit the grammar, we changed the order of the constituents in verb phrases. The complement (a noun phrase or a complementizer phrase) now precedes the verb. In other words, we have changed the grammar from "SVO" (subject-verb-object) to "SOV" (subject-object-verb); see Figure 2.6 in the main text for an example sentence. Doing so preserves the word frequency values as well as the number of words per sentence, but drastically changes the structure of the language.

Model training and testing The model parameters were the same as the previous simulation. The model trained on this SOV-language was subsequently tested on the exact same test set as the previous models (structured and scrambled).

Results The resulting surprisal values were significantly higher than those obtained using the original structured model[15] ($t=87.79$, $p < 0.001$) – unsurprising, because a large number of word-to-word transitions that were present in the

[15]Notice that this SOV model is *also* structured.

test set were definitely *not* present in the training set because they were **ruled out** by the grammar[16]. The correlation between the results from the structured model and the SOV-model was lower, but nevertheless still there ($\rho = 0.44$, $p < 0.001$). In other words, a difference between the syntax of the input to the model and the sentences or words the model is tested on, will lead to higher surprisal values.

Model predictions Lexical accuracy was 20.4%, lower than the original model; also the POS accuracy was lower than the original model (58.9%).

2.7.1.3 Simulation: 'surprisal obscures the view', word frequency

The word frequency parameters were adjusted for a few words in the original SVO grammar. Specifically, we adjusted the frequency of the words 'woman', 'discovers', and 'a' to be twice as high as the other words in their syntactic category (nouns, verbs, and determiners, respectively). Essentially, this means that the lexical entropy in the training corpus is lower.

Model training and testing The model parameters were the same as the previous simulation. The model trained on this WF-adjusted-language was subsequently tested on the exact same test set as the previous models (structured and scrambled).

Results We then tested this model on the same test set again, and indeed: there was a significant difference between these distributions ($t = 6.78$, $p < 0.001$), while the correlation between the original- and the word frequency adjusted estimates was still high ($\rho = 0.84$; $p < 0.001$). See Figure 2.8 in the main text. Lexical accuracy was slightly lower than the original (27.9%); POS accuracy was te same (76.3%).

2.7.2 Natural language simulations

The corpora used for these simulations were obtained from the OpenSubtitles project (Lison & Tiedemann, 2016).

Scripts used for these simulations:

[16] This was crucially *not* the case in the scrambled model; any word-to-word transition was possible.

- **preprocessing-opensubtitles.py**: sentence & word tokenization and interpunction removal of OpenSubtitles corpus
- **train-model-natural-1layer.py**: train model on OpenSubtitles corpus
- **test-model-natural.py**: test model trained on OpenSubtitles corpus
- **correlation-natural.py**: compares the surprisal values on the test set between scrambled and structured models
- **clustering.py**: use a RandomForestClassifier to classify surprisal values as coming from Spanish or English

The weights for these models can be made available upon request. The corpora can be downloaded from `https://opus.nlpl.eu/OpenSubtitles-v2018.php`. Surprisal values for the test sentences can be found in the folder 'results'.

2.7.2.1 Simulation: 'syntax leads to surprisal', part II

Model training & testing We trained a recurrent neural network with a 300-node embedding layer, a 600-node LSTM-layer, and a linear layer on approximately 118.000 English sentences (roughly 800.000 words) to predict the next word using a context of 10 words on a sentence-by-sentence basis. and a linear layer mapping back to the word space. The learning rate was 0.1 and we used negative log likelihood loss as implemented in PyTorch (Paszke et al., 2019).

Like before, we trained two models: a structured model, trained on intact sentences from the corpus; and a scrambled model, trained on sentences in which the word order was randomized. This method of scrambling maintains word frequency estimates, word frequency per sentence, and sentence length, but removes all sentential structure. We tested both models on the same test set of 10.000 sentences (approximately 70.000 words).

Results While the difference between the random and the structured models is much smaller, here too we observe a difference between the distributions ($t = 29.18$, $p < 0.001$); the mean difference in surprisal values is 0.58 bit (sd 1.82). Here too, however, we observed a correlation of 0.91 ($p < 0.001$) between the surprisal values estimated from the scrambled and structured model. See Figure 2.4 in the main text.

2.7.2.2 Simulation: 'surprisal does not lead to syntax'

For this simulation, we trained two additional models on the Spanish translation of the English corpus. This was a corpus of approximately 116.000 sentences (roughly 800.000 words).

Model training & testing The training and testing procedures were identical to those of the English model.

Classification Before classification, we z-scored the surprisal values to account for the possibility that one of the languages is generally more surprising than the other. Subsequently, we trained a Random Forest Classifier (100 estimators as implemented in Scikit-Learn (Buitinck et al., 2013) on 80% of these sentences to predict whether the surprisal values belonged to English or to Spanish. Since structure may be encoded in patterns of surprisal values rather than the individual values, we did this for a range of groups of surprisal values (from unigrams to 10-grams). The distribution of the surprisal values is visible in Figure 2.5 in the main text.

Results We found that the classifier was able to predict with 63.8% accuracy if a single surprisal value belonged to the English or the Spanish grammar, and this increased to 74.1% for groups of 10 surprisal values With chance at 50% and 10.000 testing items, this means that the classifier performs above chance. We could have stopped here, and concluded that we were wrong: surprisal values *do* map back onto structure. But alas, we did not. We trained the same classifier on the results from the Spanish and English *scrambled* models (the words shuffled within each sentence; see the distribution in Figure 2.5 in the main text). Despite these models not containing *any* structural information, the classifier performed at 66.2% for unigram surprisal values, and performance increased to 84.2% for 10 surprisal values. Apparently, surprisal values from the scrambled model are easier to attribute to one or the other language than those from the structured models.

Why do these classifiers work at all? Structure is not the driving factor, apparently. No, specific surprisal decimal values appeared to be one of the driving factors. The surprisal values were uniquely attributable to one or the other language due to high specificity of the values. In other words, each surprisal value was unique to either Spanish or English, and the classifier learned this (partially). We tested if the pattern in groups of surprisal values was strong enough

for the classifier to attribute the values to either language by rounding all values to 1 decimal. This preserves a potential structure-specific pattern, but removes the uniqueness. This change decreased the classifier's accuracy in both the structured- and the scrambled model (structured: accuracy ranges from 52.92% (unigram) to 67.98% (10-gram); scrambled: accuracy ranges from 58.46% (unigram) to 82.31% (10-gram)), but the accuracy values were still higher for the scrambled model than for the structured model.

3 | Delta-band neural responses to individual words are modulated by sentence processing[1]

Abstract

To understand language, we need to recognize words and combine them into phrases and sentences. During this process, responses to the words themselves are changed. In a step towards understanding how the brain builds sentence structure, the present study concerns the neural readout of this adaptation. We ask whether low-frequency neural readouts associated with words change as a function of being in a sentence. To this end, we analyzed an MEG dataset by Schoffelen et al. (2019) of 102 human participants (51 women) listening to sentences and word lists, the latter lacking any syntactic structure and combinatorial meaning. Using temporal response functions and a cumulative model-fitting approach, we disentangled delta- and theta-band responses to lexical information (word frequency), from responses to sensory- and distributional variables. The results suggest that delta-band responses to words are affected by sentence context in time and space, over and above entropy and surprisal. In both conditions, the word frequency response spanned left temporal and posterior frontal areas; however, the response appeared later in word lists than in sentences. In addition, sentence context determined whether inferior frontal areas were responsive to lexical information. In the theta band, the amplitude was larger in the word list condition around 100 milliseconds in right frontal areas. We conclude that low-frequency responses to words are changed by sentential context. The results of this study illustrate how the neural representation of words is affected by structural context, and as such provide insight into how the brain instantiates compositionality in language.

[1]Adapted from Slaats, S., Weissbart, H., Schoffelen, J.-M., Meyer, A. S., & Martin, A. E. (2023). Delta-band neural responses to individual words are modulated by sentence processing. *The Journal of Neuroscience*, 43(26), 4867-4883. doi:10.1523/JNEUROSCI.0964-22.2023.

3.1 Introduction

During language comprehension, listeners recognize words, retrieve stored information about them, and use this knowledge to combine the words into phrases and sentences. Psycholinguistic experiments have long shown that the behavioral responses to words change under the influence of the syntactic and sentential context that the words appear in (e.g., Katz, Boyce, Goldstein, & Lukatela, 1987; Marslen-Wilson & Welsh, 1978; Tyler & Wessels, 1983). In a step towards understanding how the brain builds sentence structure, the present study concerns the neural readout of this process. We ask (1) whether low-frequency neural readouts associated with words systematically change as a function of being or not being in a sentence context; and (2) whether neural readouts are modulated by purely lexical properties over and above sensory and distributional variables. We do this by contrasting MEG responses to words in sentences with word lists, the latter lacking any syntactic structure or coherent lexical and combinatorial meaning.

In psycholinguistic models, language comprehension is instantiated as a cascaded process in which information can flow bidirectionally (Marslen-Wilson & Welsh, 1978; Martin, 2016, 2020). Put simply, this means that speech sounds cue stored representations of words, and while the next words are being recognized, the retrieved information about words cues representations of phrase and sentence structure. At the same time, the already formed representations of sentences, phrases, and words cue lower-level representations: the information flows in two directions.

As words are being combined into phrases and sentences, then, responses to words change as a consequence of the top-down information flow. Indeed, a long tradition of research in psycholinguistics has shown that words in sentences are recognized faster than those same words appearing in isolation (Marslen-Wilson & Welsh, 1978; Tyler & Wessels, 1983). This effect is so powerful that it reduces effects of properties of the words *themselves*, such as word frequency. In isolation, highly frequent words are recognized faster than low-frequency words. In sentence context, this effect tends to be reduced: low-frequency words are recognized faster in sentence context than in isolation, while there is little change in recognition times for the high-frequency words (Schuberth & Eimas, 1977; Simpson, Peterson, Casteel, & Burgess, 1989).

To gain a full understanding of human sentence comprehension, the field currently faces the challenge of integrating these findings with knowledge of neural processing. Although recent studies provide insight into the neural correlates of

sentence structure (e.g., Bai et al., 2022; Brennan & Martin, 2020; Coopmans et al., 2022; Ding et al., 2016, 2018; Kaufeld, Bosker, et al., 2020; Meyer, Henry, Gaston, Schmuck, & Friederici, 2017; Nelson, El Karoui, et al., 2017; Tavano et al., 2022; Ten Oever, Carta, et al., 2022), much about the process of building these structures remains unknown (see Ten Oever, Kaushik, and Martin (2022) for discussion). Furthermore, while we know that the neural signal is sensitive to lexical information (Armeni et al., 2019; Brodbeck, Hong, & Simon, 2018; Brodbeck, Presacco, & Simon, 2018; Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Heilbron et al., 2022; Weissbart et al., 2019), we do *not* know how neural responses to words are transformed in the process of comprehension.

In this study, therefore, we aim to add to our understanding of how the brain leverages linguistic information when building sentence structure by finding a neural readout of the context effect on responses to words – above and beyond statistical predictability effects as quantified through entropy and surprisal. To this end, we analyzed a published MEG dataset by Schoffelen et al. (2019) of 102 participants listening to sentences and word lists. Despite these conditions being the main experimental manipulation in this open data set, they have not previously been directly compared. Using temporal response functions (TRFs), we disentangled delta- and theta band responses to individual words from responses to the speech envelope and word onsets, as well as entropy, and surprisal. This method allowed us to model any differences between the conditions that go beyond our difference of interest (structured/unstructured), and, as such, control for them. We compared the responses to individual words between word lists and sentences. Any differences between the lexical responses in these conditions reflect the effect of structure building on the processing of words.

The lexical response was modeled using *word frequency*. We chose this feature because word frequency is a proxy for the likely familiarity of the listener with the word and relatedly of ease of processing. Any modulation as a consequence of word frequency, therefore, captures the presence of word identity information in the signal. Furthermore, word frequency is *unigram* – in other words, it does not depend on the context. Therefore, the value corresponding to a given word is the same in a sentence and a word list. Differences between the neural readout of both conditions will therefore be due to the sentence context supplying structure and meaning, and not to the predictor itself.

We hypothesized that the delta-band responses to word frequency would be different in word lists and sentences as a consequence of the (in)availability of sentence context (Huizeling, Arana, Hagoort, & Schoffelen, 2022; Meyer, 2018;

Meyer, Sun, & Martin, 2020a; Meyer et al., 2020b). Studies that investigated the presence of lower-level features in the neural signal as a function of the availability of linguistic information suggest that lower-level features are represented by the delta-band neural signal more reliably when higher-level information is available. For example, mutual information between the speech signal and the neural signal is higher in the presence of structure and meaning (Coopmans et al., 2022; Kaufeld, Bosker, et al., 2020; Ten Oever, Carta, et al., 2022); and the strength of speech tracking is dependent on the listener's knowledge of the language (Blanco-Elorrieta, Ding, Pylkkänen, & Poeppel, 2020; Molinaro & Lizarazu, 2018), and general comprehension (Keitel, Gross, & Kayser, 2018). Following these results, we expected a stronger presence of the word frequency response (the lower-level feature) in the sentence condition than in the word list condition (the higher-level information) in the delta band specifically. Theta-band effects tend to be found as a function of acoustic rather than abstract linguistic manipulations (Blanco-Elorrieta et al., 2020; Etard & Reichenbach, 2019; Molinaro & Lizarazu, 2018; Sohoglu, Peelle, Carlyon, & Davis, 2012). In this study, we expected to observe this distinction between delta and theta-band activity through an absence of effects in the theta band.

3.2 Materials and methods

To answer our research question, we analyzed a part of the open-access large multimodal MEG dataset (N=204) MOUS (Mother of all Unification Studies) published by Schoffelen et al. (2019). In addition, we performed two types of control analyses; an analysis of a dataset published by Ten Oever, Carta, et al. (2022) and a set of simulations. Methods for all analyses are described below.

3.2.1 Participants

A total of 102 native speakers of Dutch (51 men, 51 women) with a mean age of 22 (range: 18 to 33) were included in this analysis. In this half of the dataset, participants were presented with the stimuli auditorily (as opposed to the other half, where stimuli were presented visually). All participants were right-handed, reported normal hearing, had normal or corrected-to-normal vision, and had no history of neurological, developmental or linguistic deficits. All participants provided informed consent and the study was approved by the local ethics committee (CMO – the local "Committee on Research Involving Human Subjects" in the

Arnhem-Nijmegen region) and followed guidelines of the Helsinki declaration. Participants took part in an fMRI and an MEG session, during which they listened to sentences and word lists. Only the MEG data are included in the present study.

3.2.2 Materials

The complete set of stimuli consisted of 360 natural Dutch sentences of 9 to 15 words (mean: 11.6) with varying syntactic structures, and 360 word lists. To create the word lists, the words from the sentences were scrambled such that more than two consecutive words did not form a coherent fragment. The stimuli were recorded by a female native speaker of Dutch. The sentences were pronounced naturally. The word lists were pronounced with neutral prosody and a clear pause between each word. The files were recorded in stereo at 44100 Hz. The sentences had an average duration of 4.27 seconds (sd. 0.61), and the word lists of 7.67 seconds (sd. 1.04). During the post-processing, the audio files were low-pass filtered at 8500 Hz and normalized such that all the audio files had the same peak amplitude and peak intensity. In the word list condition, the individual words were spliced together with variable silence between them. This created conditions with different acoustic properties. We address this issue in sections 3.2.4, 3.2.6, and 3.2.7 below. In both conditions, the transition from silence to speech was ramped at the onset and offset with a rise/fall time of 10ms. Word onsets and offsets were determined manually for each audio file using the Praat software (Boersma & Weenink, 2018).

The stimuli were divided over two sets, set A and set B. During the MEG session, participants were presented with 120 sentences from set A and 120 word lists from set B (or the reverse). Across participants, all stimuli were presented the same number of times in the sentence and word list condition.

3.2.3 Procedure

Prior to the task, participants read a written instruction and were allowed to ask clarification questions. The experimenter emphasized that the sentences and word lists should be attended carefully, and discouraged attempts to integrate the words in the word list condition. To familiarize the participants with the task, all participants performed a practice block with stimuli not included in the study. During the MEG measurement, the stimuli were presented in 24 blocks, alternating between sentence blocks (each containing 5 sentences) and word list blocks (each containing 5 word lists). The starting block type (either sen-

tences or word lists) was randomized across participants. At the start of each block there was a 1500ms presentation of the block type: 'zinnen' (sentences) or 'woorden' (words). The inter-trial interval was jittered between 3200-4200ms. During this period, an empty screen was presented, followed by a fixation cross.

In order to assure participants paid attention to the stimuli, 20 percent of the trials were followed by a 'Yes'/'No' question about the content of the preceding sentence/word list. Half of the questions on the sentences addressed the content of the sentence (e.g.: 'Did grandma give a cookie to the girl?') whereas the other half, and all of the questions about the word lists, addressed one of the main content words (e.g.: 'Was a grandma mentioned?'). Participants answered the question by pressing a button for 'Yes'/'No' with their left index and middle finger, respectively. While the tasks were not identical between the conditions, the randomized order of appearance of question types ensured that participants could not approach the sentences any differently from the word lists: any sentence or list trial could be followed by the word monitoring task.

The stimuli were presented via plastic tubes and ear pieces to both ears. The hearing threshold was determined individually for each participant prior to the experiment, and the stimuli were presented at an intensity of 50 dB above the hearing threshold.

The experiment was run using the Presentation® software (Version 16.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). MEG was continuously recorded with 275-channel axial gradiometer system (CTF) at a sampling frequency of 1200 Hz (cut-off frequency of the analog anti-aliasing low-pass filter was 300 Hz). Three head localizer coils were attached to the participant's head (nasion, left- and right ear canals) to determine the position of the head relative to the MEG sensors. The head position was monitored throughout the measurement. If needed, the participant was asked to reposition in order to correct for head position changes during breaks. The audio signal of the stimuli presented in the scanner were recorded along with the MEG data using an ADC-channel.

Structural MRI images for source reconstruction were acquired using a T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence with the following acquisition parameters: volume TR = 2300 ms, TE = 3.03 ms, flip-angle = 8 degrees, 1 slab, slice-matrix size = 256 x 256, slice thickness = 1 mm, field of view = 256 mm, isotropic voxel size = 1.0 x 1.0 x 1.0 mm. A vitamin-E capsule was placed as fiducial behind the right ear to allow visual confirmation of left-right consistency.

3.2.4 MEG preprocessing

The MEG data were preprocessed with custom-written MATLAB scripts using the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011, Donders Institute for Brain, Cognition and Behaviour, Radboud University, the Netherlands. See `http://fieldtriptoolbox.org`). Before filtering, the data were epoched from audio onset to audio offset. The epochs were baseline-corrected and band-pass filtered into the designated frequency band using a windowed-sinc Finite Impulse Response (FIR) filter (15 second data-padded), after which they were resampled to 120 Hz for TRF estimation.

The frequency band of interest was defined on the basis of the rate of occurrence of words in the stimuli, the differences in speech-brain coherence between conditions, and the literature (e.g., Blanco-Elorrieta et al., 2020; Donhauser & Baillet, 2020; Molinaro & Lizarazu, 2018; Weissbart et al., 2019). The word rate in the word listswas 1.5 Hz (sd. 0.1), and in the sentences 2.7 Hz (sd. 0.3). To compute speech-brain coherence, we first computed the broadband speech envelope by taking the absolute value of the Hilbert transform of the speech signal, low-passing it at 20 Hz and scaling the output between 0 and 1. We computed the magnitude squared coherence estimate of the broadband speech envelope and the MEG signal using Welch's method. The differences between word lists and sentences were estimated using a cluster-based permutation test. This revealed three peaks in the low-frequency signal; one between 1 and 3 Hz, one between 4.5 and 7 Hz, and one between 9.5 and 12 Hz. See Figure 3.1 below (see also: Lam, Schoffelen, Uddén, Hultén, & Hagoort, 2016). On the basis of these clusters and frequency bands analyzed in the literature (e.g., Donhauser & Baillet, 2020), we analyzed two frequency windows: delta (0.5-4 Hz) and theta (4 – 10 Hz). To account for differences in speech-brain coherence that were exclusively due to acoustic differences between the conditions, we included the speech envelope as a predictor in all the models of the data (see the modulation spectra in Figure 3.1 below). Details of the models are presented in sections 3.2.6 and 3.2.7.

3.2.5 Source reconstruction

MRI images were co-registered to the MEG headspace coordinate system by aligning the positions of the pre-auricular points and the nasion MEG coil to the MRI images using the MNE-Python coregistration GUI. For each participant, we reconstructed the cortical surface using the *watershed* algorithm from Freesurfer.

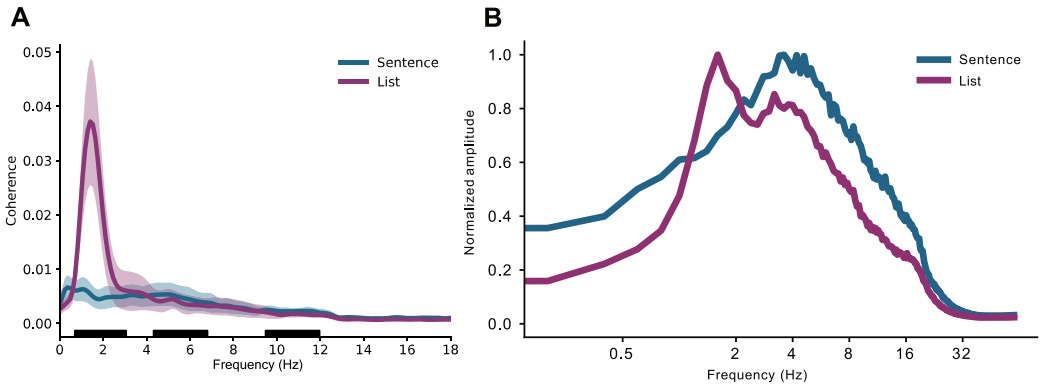


Figure 3.1: Speech-brain coherence and modulation spectra. A. Speech-brain coherence. Shaded area indicates standard deviation. Black bars indicate frequencies that were part of clusters that contributed to the significant difference between sentence- and word list coherence. B. Modulation spectra of the broadband speech envelopes (part of the TRF base model). The modulation spectra were obtained by concatenating the stimuli per stimulus type and performing a fast Fourier transform on snippets of 5 seconds. The resulting spectra were averaged.

We created a surface-based source space with 'oct6' spacing, meaning approximately 5 mm was between the source points. This generates 4098 sources per hemisphere. We created a single-layer BEM-model with surface ico downsampling of 5120, from which the lead field was computed. The sources were reconstructed using a scalar LCMV beamformer approach with a unit-noise gain beamformer to deal with depth bias. The data covariance used for computing LCMV filters was whitened using the covariance matrix of resting state data. The resting state data was band-pass filtered into the appropriate frequency band (i.e., 0.5-4 Hz for the delta band, and 4-10 Hz for the theta band). After application of the LCMV beamformer filters to the epoched MEG data, the source-localized epochs were morphed to fsaverage for group statistics. These source-localized, morphed epochs were then entered into the pipeline for temporal response function estimation. Source localization failed for 11 participants due to convergence issues for the noise covariance matrix or missing resting state data ($N_{source} = 91$).

3.2.6 Temporal response functions

In order to characterize the effect of linguistic structure and meaning on the neural response, we estimated temporal response functions (TRFs) to different acoustic and linguistic features. This approach has been used to determine re-

sponses to different linguistic features, ranging from the speech envelope and phonemic information (Di Liberto et al., 2015; Donhauser & Baillet, 2020), to lexical information (Broderick et al., 2018; Weissbart et al., 2019) and even syntactic embedding (Nelson, El Karoui, et al., 2017). The response function of interest here is the response to word frequency, as this is a unigram feature and therefore has the same per-word values in both conditions.

The TRFs were estimated using linear regression. We modelled the neural response by convolving the TRF kernel with the stimulus representation signal. In summary, this method reduces to a multivariate multiple linear regression, where we used lagged time series of stimulus features as predictors. The model equation reads as:

$$y_c(t) = \sum\sum x_f(t)\beta_f(t-\tau_k) + \eta(t) \tag{3.1}$$

Where $\{y_c\}_t, \{x_f\}_t, \{\beta_f\}_t$ represent the recorded MEG signal of channel c, the input feature f and its temporal response function respectively. $\{\eta\}_t$is a gaussian noise process accounting for measurement noise. We are using a time discrete representation of each signal, where their values are sampled at discrete time intervals $t_k = \frac{k}{F_s}$, with sampling frequency F_s. This linear model can be easily rewritten in its vectorized form and further concatenated such that we model at once all channel equations independently. We estimate the coefficients of the TRFs $\hat{\beta}_f$ by minimizing the squared error between the measured MEG signals and the reconstructed signal obtained from equation (1) while keeping the norm of TRFs coefficients, $||\beta||_2$ low to avoid overfitting. This minimization problem is solved in a closed form by:

$$\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T Y \tag{3.2}$$

Where $Y \in \mathbb{R}^{N\times C}$ is the matrix representation of the measured MEG signal (for C channels arranged column-wise, each with N data samples); $\hat{\beta} \in \mathbb{R}^{(K.F)\times C}$ contains the estimated TRFs with K lags, F features for all C channels; $X \in \mathbb{R}^{N\times(K.F)}$ is a matrix containing all lagged feature time series of length N; λ is a regularization coefficient and I_d the identity matrix. The regularisation coefficient is needed to avoid overfitting which in this case translates to the square matrix $X^T X$ not being full rank. Numerically, small eigenvalues or simply ill-conditioned matrices suffice to make the inversion unstable and thus will require regularization. In our case this happens when features present some amount of autocorrelation (as columns of X are time-lagged version of other columns. Continuous regres-

sors such as the acoustic envelope (see section 3.2.7 “Stimulus Representation”) will present strong autocorrelation and thus call for regularization.

In equation 3.1, the vector of weights $\beta_f(t)$ represents the coefficients parameterizing the temporal response functions. They form a time course reminiscent of an event related potential that tells us at which point in time (and, potentially, where) a feature modulates the neural signal. Thus, an increase at a certain lag for a given feature reflects an increase in the associated brain response to this feature at that given sensor and at the given time lag after stimulus onset. The concept of stimulus onset, especially for a continuous regressor such as the envelope, here reduces to a situation where the brain would be stimulated by an impulse of sound. Eventually, we estimate, from a system identification perspective, the transfer function mapping input to output when the brain is considered as a linear time-invariant system.

To evaluate how our models perform at reconstructing the neural data, we computed the Pearson’s correlation coefficient between the true data and data reconstructed using the estimated TRFs. The correlation between the reconstruction and the original MEG indicates how much of the variance in the neural signal is explained by the features. The TRFs were not estimated on the same portion of data used to score the model. As further explained in section 3.2.8 “model fitting”, we used a nested cross-validation procedure to tune the regularization parameter, estimate the TRF coefficients and finally score the resulting model. Unless specified otherwise, all analyses described below were done with custom made Python scripts using MNE-Python (Gramfort et al., 2013). The whole analysis was conducted both in sensor- and in source space.

3.2.7 Stimulus representation

Its multivariate character makes the TRF especially suitable for the current analysis: it allows for controlling for differences between conditions that are not currently under discussion by modelling them. To characterize the speech signal and part of its linguistic content, we constructed five different features: word frequency (the feature of interest), and four control features; the speech envelope, word onsets, entropy, and surprisal.

The *speech envelope feature* was computed for each stimulus by taking the absolute value of the Hilbert transform and down sampling it to 120 Hz to match the down-sampled MEG sampling rate. The envelope feature was added to represent the acoustic response and as such captures the difference between condi-

tions observed in the cerebro-acoustic coherence that was caused by differences in the acoustic input (see Figure 3.1 A and B).

The *word onset feature* was added to capture broadly any time-locked response to word onset for which the variance is not already explained by other features. As such, this feature can also capture any effects of segmentation that were different between the conditions. The word onsets and offsets were transcribed manually for each stimulus. We used a train of unit impulses, where the feature signal is one at the word onset sample and zero otherwise:

$$x(t) = \sum_{words} \delta\left(t - t_{onset}\right) \tag{3.3}$$

These impulse trains were convolved with a Gaussian kernel with a standard deviation of 15ms. Such temporal smoothing has the effect of inflating the autocorrelation of the signal. We designed the width of this smoothing such that the smoothed impulses end up with energy spanning a comparable frequency band as to our continuous regressor (envelope). The Fourier Transform of a gaussian is also a gaussian, and the 15ms standard deviation of the temporal smoothing kernel equates to a spectral standard deviation of 21.22Hz. This ensured that all features required a similar degree of regularization in the regression analysis, and made it possible to include impulse-like features such as word onsets and the envelope in the same regularized regression. Notably, this also translates into some uncertainty about or knowledge of the exact word onset timings.

Like the word onset feature, the *word frequency feature* was constructed as an impulse train of zeros everywhere but at word onset. Here we used the respective word frequency value to modulate the height of the impulses. We used the log-transformed value of occurrence per million words, obtained from the SUBTLEX-NL corpus (Keuleers, Brysbaert, & New, 2010):

$$x_{wf}(t) = \sum_{words} -\log\left(p\left(w\right)\right) \times \delta\left(t - t_{onset}\right) \tag{3.4}$$

where $P(w)$ represents the unigram probability estimated from occurrence per million words.

If a word did not exist in the corpus, the fallback value of 0.301 (log/million) was used, corresponding to the lowest word frequency in the corpus. The values were z-scored across all stimuli. The resulting signal was convolved with the same Gaussian kernel as the word onset feature.

The *entropy feature* consists of lexical entropy, a weighted probability measure that quantifies the uncertainty about the upcoming word on the basis of the

previous words. It provides a numeric answer to the following question: given the n previous words, with what degree of certainty can we predict the upcoming word?

$$H\left(w_i|w_{i-n}...w_{i-1}\right) = -\sum p\left(w_i|w_{i-n}...w_{i-1}\right)\log\left(p\left(w_i|w_{i-n}...w_{i-1}\right)\right) \quad (3.5)$$

The value was derived from a trigram model trained on the NLCOW2012 corpus using WOPR (van den Bosch & Berck, 2009). If a value was missing, the average of all entropy values was used. Like the word frequency feature, the entropy values were z-scored relative to all stimuli and inserted in a stick function, after which the stick function was convolved with the same Gaussian window. This feature was added to ensure that any effects on the word frequency feature were of a compositional semantic and structural nature, rather than a probabilistic one.

The *surprisal feature* reflects how surprising a given word is in its immediate context. From an information-theoretic perspective, this reflects the information content, or self-information, of a word. It was calculated as the log10-transformation of the conditional probability of a word, which was taken from the same trigram model as the entropy values. This means that surprisal is always based on the two preceding words: given the two preceding words, how high was the chance that the observed word would, indeed, appear? If the chance was low, surprisal is high. The feature was constructed in the same way as the word frequency and entropy features; the values were z-scored across all stimuli, inserted in a stick function at word onsets, and convolved with the Gaussian window.

$$I\left(w_i|w_{i-n}...w_{i-1}\right) = -\log_{10}\left(p\left(w_i|w_{i-n}...w_{i-1}\right)\right) \quad (3.6)$$

Since the three numerical lexical features (frequency, entropy, surprisal) might be correlated to some extent, we need to assert that the degree of multicollinearity present in our stimulus representation will not hinder the TRF coefficient interpretation. We checked whether the Variance Inflation Factor (VIF) was below 5 (considered a relatively conservative measure of multicollinearity; Sheather, 2009; Tomaschek, Hendrix, & Baayen, 2018). The VIF was computed by correlating the z-scored entropy, surprisal, and word frequency values, and taking the diagonal of the inverted correlation matrix. This was done for all the stimuli,

and for both conditions separately. The VIF was never higher than 5; the highest VIF was for Surprisal at 4.8 in the word list condition.

3.2.8 Model fitting

The features were fitted in a cumulative manner to assess the contribution of each feature. This led to a total of seven models per frequency band: an Envelope model, consisting of only the speech envelope; an Onset model, consisting of the speech envelope and the word onset features; and a Frequency model, consisting of the speech envelope, word onset, and word frequency features; an Entropy model, containing the speech envelope, word onset, and entropy features; a Surprisal model, consisting of the speech envelope, word onset, and surprisal features; and cross-combinations of those with- and without the word frequency feature. An overview of all models and the corresponding features is displayed in table 3.1 below.

Before model fitting, the data was split pseudo-randomly into a training- and testing set at an 80/20 ratio. Care was taken that the sentences and word lists were evenly divided across the training and test sets. The sentence- and word list models were each trained on 96 out of 120 trials. The regularization parameter was optimized individually per participant, frequency band and model (but not per condition) using an eight-fold cross-validation procedure with 20 log-spaced values around the eigenvalues of the covariance matrix of the lagged speech envelope ($\lambda = 60470.9$) ranging from $\lambda \times 10^{-3}$ to $\lambda \times 10^{3}$. The best regularization parameter was determined as the value for which the average (across sensors) reconstruction accuracies were highest. Occasionally, reconstruction accuracies would not increase with a higher degree of regularization; instead, increasing the regularization would leave the reconstruction accuracy at the same value, until overregularization occurred and reconstruction accuracy went down. In this case, the highest lambda value before a drop in accuracy occurred was chosen to ensure some degree of regularization. Each model was fitted on the complete training set using the regularization parameter from the cross-validation procedure, yielding the TRFs.

In the analysis of the source-localized MEG data, the manipulations were simplified due to computational limitations. The two maximal models were fitted, with word frequency as the only difference: the Entropy/Surprisal model, consisting of the speech envelope, word onsets, entropy, and surprisal features; and the full model, consisting of all features. The cross-validation procedure was

brought down to five-fold with ten log-spaced values around the eigenvalue of the stimuli (60470.9) ranging from $\lambda \times 10^{-2}$ to $\lambda \times 10^{2}$.

3.2.9 Model evaluation

Each model was evaluated by convolving the estimated TRFs with the unseen stimuli from the test data set. This yields, in essence, a prediction of the neural signal according to the model. The predicted neural signal was then correlated with the original neural signal from the test set using the Pearson product-moment correlation on a sensor-by-sensor or source-by-source basis. For every individual participant, this yielded a set of sensor- or source-based *reconstruction accuracies* for each model.

Table 3.1: The fitted encoding models.

Model name	Feature				
	Envelope	Word onset	Entropy	Surprisal	Word frequency
Envelope	×				
Onset	×	×			
Entropy	×	×	×		
Surprisal	×	×		×	
Frequency	×	×			×
Entropy / Surprisal	×	×	×	×	
Entropy / Frequency	×	×	×		×
Surprisal / Frequency	×	×		×	×
Full	×	×	×	×	×

Note. **An × indicates that a feature was included in the model.**

The TRF analysis has two deliverables: first, the TRF (the development of the estimated coefficients across time), which is an ERP-like waveform that captures how the neural signal changes as a function of (e.g.) word frequency; and, second, the reconstruction accuracy, which is a metric of model fit. Here, we wanted to know (1) whether the responses to word frequency differ between sentences and word lists in time and space, so we compared the TRFs between conditions; and (2) whether the presence of the word frequency response differed between sentences and word lists, so we tested whether the word frequency predictor contributed differently to the reconstruction accuracy of a model in the two conditions.

Throughout, evaluation for statistical significance of the difference between TRFs was done using cluster-based permutation tests. Cluster-based permutation tests address the null hypothesis of exchangeability across conditions by a Monte Carlo estimate of the randomization distribution of a cluster-based test statistic, optimizing statistical sensitivity while controlling the false alarm rate.

Here, we used the T-statistic as the test statistic. In these tests, we create matrices of all sensors and samples. Then, we compute the difference between two conditions and express it as a T-statistic for each of these data points. The T-values are thresholded at an a priori threshold, and the thresholded T-values are summed across clusters on the basis of spatial and temporal adjacency. The significance of the resulting largest cluster's test statistic is compared to 1024 of similarly obtained test statistics, after random permutation of the condition labels. We used the function *spatio_temporal_cluster_test* from the MNE-Python library (Gramfort et al., 2013) with the t-statistic as the test statistic and 1024 permutations.

To assess whether the responses to word frequency differed qualitatively between conditions in sensor space, the difference between the word frequency TRFs for the sentence and word list conditions was evaluated using a cluster-based permutation test. In addition, to characterize the response in each condition separately, we performed two cluster-based permutation tests with the same methods in which we contrasted the response against zero in each condition separately. In total, we performed three cluster-based permutation tests on the sensor TRFs: one on the difference between conditions, and one on the TRF for each of the two condition separately (against zero). In all cases, we calculated the threshold on the basis of the t-distribution with a significance level of $5x10^{-8}$ with 101 (number of participants – 1) degrees of freedom. This equals three times the recommended threshold for the number of participants. The threshold was increased to yield the most informative results (i.e., to ensure not every sensor and time-lag would be significant). Subsequent comparisons were done with a threshold calculated using a Bonferroni adjusted significance level (i.e., divided by two) to correct for multiple comparisons; all else was the same.

In addition, we wanted to evaluate whether there was a latency difference between the responses in the two conditions. To this end, we compared the responses from the sentences and word list conditions in a cross-correlation. The cross-correlation was done on the grand-average TRF waveforms of overlapping sensors between conditions from the clusters resulting from the one-sample tests. We sequentially cross-correlated each sensor, and normalized the values by dividing them by the maximal value from the cross-correlation for that sensor. We then obtained the peaks for every sensor. This number corresponds to the "lag" at which the two signals had the highest correlation, and show how different the responses are in time. Subsequently, we shifted the sentence response in time by the number of samples of the peak. We then correlated the shifted sentence

response and the original word list response. To check for significance, we performed the same procedure for randomly selected channels and repeated this process 10000 times.

In source space, we compared the TRFs for word lists and sentences using a cluster-based permutation test in two time-windows on the basis of the results from the analysis in sensor space: 200-400ms and 500-700ms post stimulus onset (further: PSO), respectively. We did this to get a more reliable estimate of the spatial distribution of the effects, although cluster-based permutation tests account only for a difference between the distribution overall, therefore any spatial or temporal differences are approximations and inconclusive (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019). The threshold was set to the t-distribution with an alpha of 0.025 (98.75 and 1.25^{th} percentile) to correct for multiple comparisons, with 90 (number of participants-1) degrees of freedom. Sources along the medial wall were excluded.

In the sensor space analysis, the reconstruction accuracies were averaged over sensors and submitted to a linear mixed model using lme4 in R (Bates, Mächler, Bolker, & Walker, 2015). The model had the factor *condition* (two levels: sentence and word list), and a random intercept for *participant*. In addition, the model contained three binomial factors *frequency*, *entropy*, and *surprisal*, describing whether a feature was (1) or was not (0) in the model in order to calculate a slope for each feature separately.

$$accuracies \sim condition \cdot (frequency + entropy + surprisal) + (1|participant) \tag{3.7}$$

We used a stepwise variable selection to evaluate the contribution of each of these factors. To evaluate the contribution of a given factor (or interaction), a model with the factor was compared to a model without it, and the goodness-of-fit statistics were compared using a chi-square test. If the removal of a factor did not decrease goodness-of-fit, the next factor was removed. When the removal of a given feature or interaction significantly decreased model fit, the removal of features was stopped. The prefinal model should then describe the data best. As a final check, the AIC of the models was compared using the R-package *AICcmodavg* (Mazerolle, 2020). Post-hoc t-tests were done between the Entropy/Surprisal and Full model to evaluate whether the effects held between the largest models.

In source space, a cluster-based permutation test was done to localize the interaction effect using the function *permutation_cluster_test* from the MNE-Python library. The test statistic was an F statistic from a two-way ANOVA with factors Condition (levels: word list, sentence) and Model (levels: Entropy / Surprisal, Full) with a threshold determined using the function *f_threshold_mway_rm*, equally from the MNE-Python library. The data was permuted 1024 times.

3.2.10 Control analysis I: data from Ten Oever, Carta, Kaufeld & Martin (2022)

The word lists were presented with variable silences between words. The sentences, on the other hand, were natural, with pauses occurring sparingly. This caused differences of word rate and signal length between the conditions that may affect our results. To examine potential effects of the pauses in the word list condition, we analyzed a second dataset of 16 participants listening to word lists and sentences using the same methods. Importantly, the word lists in this condition were naturally spoken, as were the sentences. This means that there were no pauses between the words in the word list condition, and there was coarticulation between words (Kaufeld, Bosker, et al., 2020). The data were supplied by Ten Oever, Carta, et al. (2022).

Participants A total of 20 native speakers of Dutch (4 men, 16 women with a mean age of 39.5) participated in the experiment. Four participants were excluded from this analysis due to a variety of reasons (e.g., session was not completed). All participants were right-handed, reported normal hearing, had normal or corrected-to-normal vision, and had no history of neurological, developmental or linguistic deficits. All participants provided informed consent. The study was approved by the ethical Commission for human research Arnhem/Nijmegen (project number CMO2014/288). Participants were remunerated for their participation.

Materials The stimuli were identical to the stimuli used in Kaufeld, Bosker, et al. (2020). The experiment consisted of three conditions in total: sentences, Jabberwocky, and word lists. Only the sentences and the word lists are analyzed here. The stimuli consisted of 10 words, which were all disyllabic except for “de” (the) and “en” (and). Sentences had a fixed syntactic structure of two coordinate clauses: [Adj N V N conj Det Adj N V N], e.g. ‘timid heroes pluck flowers and the

brown birds gather branches'. The word lists were scrambled versions of these sentences, and care was taken that there were no plausible internal combinations of words. The stimuli were recorded by a female native speaker of Dutch at a sampling rate of 44.1 kHz (mono). After recording, any pauses were normalized to ~150 ms in all stimuli and the intensity was scaled to 70 dB using the Praat voice analysis software (Boersma & Weenink, 2018).

Participants were asked to perform four different tasks on these stimuli: a passive listening task, a syllable recognition task, a word recognition task, and a word combination recognition task. In this analysis, we did not distinguish between tasks. For a description of the tasks performed, see Ten Oever, Carta, et al. (2022).

Procedure At the beginning of each trial, participants were instructed to look at a fixation cross presented at the middle of the screen on a grey background. The audio was presented binaurally through tubes after an interval randomly jittered between 1.5 and 3 seconds. One second after audio offset, the task prompt (e.g., the syllables or words for recognition) was presented, which required participants to press a button on a button box. There were eight blocks of approx. 8 minutes. After each block, participants could take a break, during which the head position was corrected. MEG was recorded using a 275-channel axial gradiometer CTF MEG system at a sampling rate of 1200Hz. After the session, head shape was collected using the Polhemus digitizer (using as fiducials the nasion and the entrance of the ear canals as positioned with earmolds).

MEG preprocessing The MEG data were processed with custom-written Python scripts using MNE-Python (Gramfort et al., 2013). As in the main analysis, the raw MEG data was filtered using a windowed-sinc Finite Impulse Response (FIR) filter between 0.5 and 4 Hz for the delta band, and 4 and 10 Hz for the theta band, after which the data was epoched from audio onset to audio offset and resampled to 120 Hz for TRF estimation.

Stimulus representation In this analysis, we used the *envelope, word onset,* and *word frequency* representations from the main analysis. For a full description, see section 3.2.7.

Model fitting We used the model-fitting approach described in section 3.2.8. We fit three models: Envelope (with only the *envelope* feature), Onset (*envelope*

and *word onset* features), and Frequency (*envelope, word onset,* and *word frequency* features). The data was split pseudo-randomly into a training and a testing set at an 80/20 ratio, ensuring that the sets contained 50% items from each condition. The regularization parameter was optimized individually per participant and model, using an eight-fold nested cross-validation procedure with 20 log-spaced values around 60000 ($\lambda = 60000$) ranging from $\lambda \times 10^{-2}$ to $\lambda \times 10^{2}$.

Model evaluation For model evaluation, we used the procedure described in section 3.2.9 of the main text.

Statistical analysis Like in the main analysis, we assessed whether the responses to word frequency qualitatively differed between conditions by evaluating the difference between the word frequency TRFs for the sentence and word list conditions using a cluster-based permutation test. In addition, to characterize the response in each of the conditions separately, we performed two additional cluster-based permutation tests with the same methods in which we contrasted the response against zero in each condition separately. In total, we performed three cluster-based permutation tests on the TRFs: one on the difference between conditions, and one on the TRF for each condition separately (against zero). In all tests, we calculated the threshold on the basis of the t-distribution with a significance level of 0.05 with 16 (number of participants – 1) degrees of freedom. Only clusters with a p-value smaller than 0.01 were considered. Subsequent comparisons were done with a threshold calculated using a Bonferroni adjusted significance level to correct for multiple comparisons; all else was the same. For comparison to the main analysis, we also compared the word onset response between conditions with the methods described above.

To evaluate the effect of *word frequency* in each condition, we compared the reconstruction accuracies from the Onset and Frequency models in interaction with condition. The reconstruction accuracies were averaged over all sensors (conservative measure). After checking for normality and sphericity through (1) visual inspection of QQ-plots and histograms, (2) statistical testing using the Shapiro-Wilk test, Anderson-Darling test, and D'Agostino's K^2 test for kurtosis and skewness as implemented in SciPy, and (3) the Mauchly test for sphericity as implemented in Pingouin (Vallat, 2018) the averaged reconstruction accuracy values were submitted to a repeated measures ANOVA using Statsmodels.

3.2.11 Control analysis II: Simulations

Using simulations, we evaluated whether the inter-word interval impacts TRF model evaluation. We did this by simulating raw MEG data consisting of a signal (different impulse responses) and a variable amount of noise.

The simulated response was equivalent to the forward model, namely a noisy output of a convolution between a predefined kernel (the ground truth for the TRF estimate) and an impulse train (for the input signal). We generated those data with variable amount of noise (i.e., explicitly manipulating the broadband signal-to-noise ratio) and with varying inter-stimulus interval (ISI) while keeping the signal length the same and the number of impulses, or events, constant (in which case shorter inter-stimulus interval results in the end portion of the output signal containing only noise).

We then scored the forward model by computing both the R^2 score and the Pearson's correlation coefficient between the reconstruction $\widehat{y}$ and the true signal using a test portion of the data, not used to estimate the coefficients β. Importantly, we then computed the scores in two ways: (1) from the fixed signal length data described above; since we also used a fixed number of impulses, or events, this resulted in a portion of the stimulated output signal to contain only noise; (2) or from a shortened signal, where we truncated all signals to the last stimulus event. This resulted in shorter signals for shorter ISI.

3.2.12 Data and code accessibility

The code is available at `https://osf.io/ky9bj/`, with the exception of the pre-processing scripts. The pre-processed data is available upon request. The raw data can be downloaded from the Donders Repository at `https://data.donders.ru.nl/collections/di/dccn/DSC_3011020.09_236?0`.

3.3 Results

3.3.1 Behavioral results

We compared participants' responses to the task that was present in both conditions, which targeted one of the main content words (e.g.: 'Was a grandma mentioned?'). To balance the number of trials included in the accuracy scores, we took a random subset of questions from the word lists (12 or 13 trials). The average proportion of correct responses was higher in the sentence condition

($mean_{sent} = 0.88$; $sd_{sent} = 0.08$) than in the word lists ($mean_{list} = 0.72$; $sd_{list} = 0.14$; $t = 10.08$, $p < 0.001$), meaning that participants remembered the words from the sentences better than the words from the word lists (see figure 3.2 below).

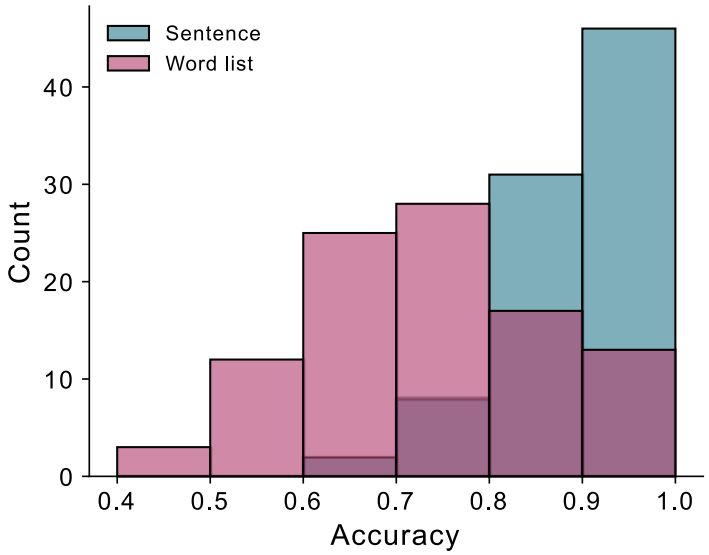


Figure 3.2: Accuracy scores for the behavioral task performed during the MEG recording. The accuracy scores include responses to word monitoring only. The word list accuracy scores are a random subset of the full set of responses to balance the number of trials (n= 12) in the word list and sentence conditions.

3.3.2 Delta band

Sensor-level analysis The cluster-based permutation test revealed differences between word lists and sentences in three clusters between 0 and 700ms. Figure 3.3A suggests that the peak of the response to word frequency was delayed by approximately 300ms in the word list condition. To evaluate if this was the case, we conducted one-sample cluster-based permutation tests and computed the cross-correlation between the two conditions for overlapping sensors from the clusters in both conditions. The one-sample cluster-based permutation test revealed a response in temporal areas in both conditions, that peaks around 250 milliseconds in the sentence condition, and around 600 milliseconds in the word list condition (see Figure 3.3 B and C).

The cross-correlation on overlapping sensors between the two conditions (time-courses and sensors visible in Figure 3.4A below) revealed a high correlation

between the word list and the sentence responses at a delay of 330 milliseconds (mean $r = 0.9$). Random sampling of sensors and lags revealed the distribution shown in Figure 3.4D; the observed values are in the upper 0.05% percentile, indicating that the observed correlation is likely not caused by chance.

Because we wondered whether the delay could be due to the differences in presentation rate, we examined differences between the TRFs for the other word-level feature that was numerically identical between conditions: word onsets (unit-spike-train in both conditions). We compared the word onset response from a model with only the envelope and word onset features. This model is equivalent to an ERP analysis which corrects for overlapping event windows (as is the case in the sentence condition) and controls for acoustic differences. A small delay, of approximately 100ms, appears in this model. This delay is in accordance with findings of an ERP-analysis on high- versus low-constraining contexts (León-Cabrera, Rodríguez-Fornells, & Morís, 2017; Liu, Shu, & Wei, 2006). Importantly, this model collapses over variance caused by the lexical features included in the full model (word frequency, entropy, and surprisal). In other words, this underspecified model attributes variance that is in fact due to word frequency, entropy, or surprisal, to the word onset predictor. When we include the other lexical predictors in the model and compared the conditions again, no such difference between the word onset responses is observed (Figure 3.3D). In this response, there were some differences around time-point zero, before as well as slightly after; these differences may indicate differences in temporal expectancy of word onset between conditions.

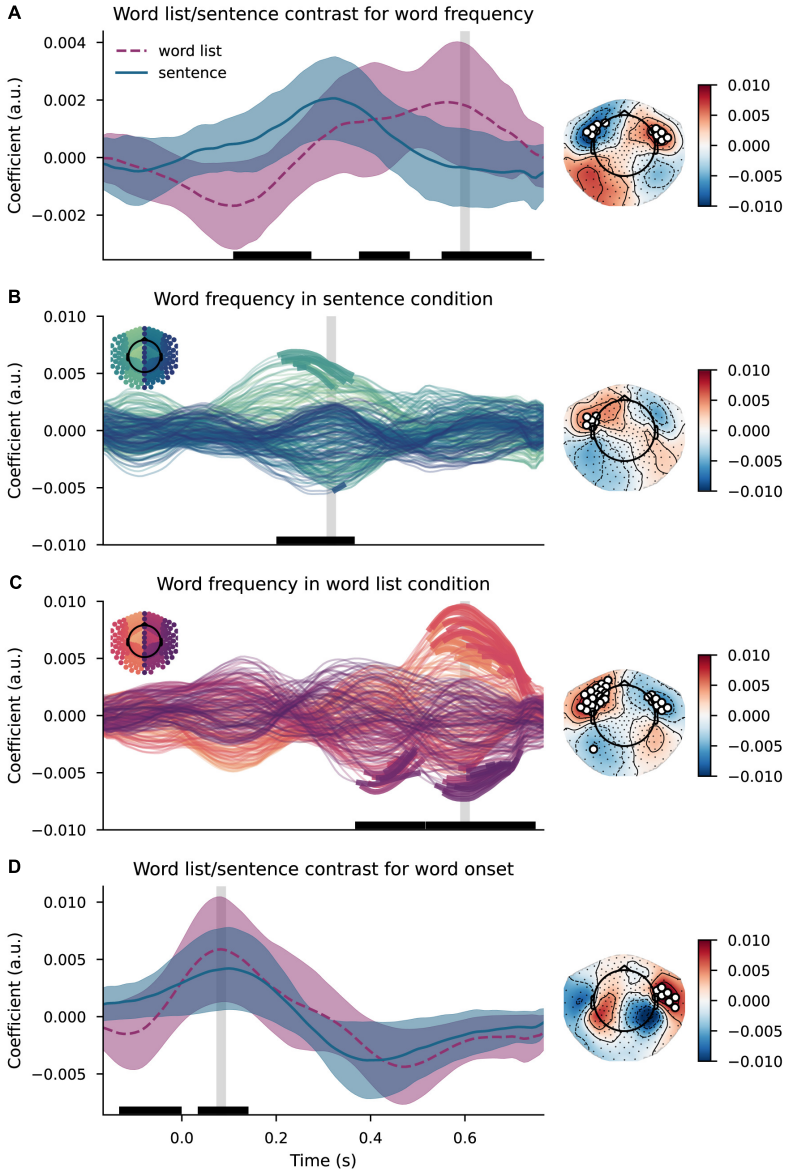


Figure 3.3: Delta-band effects (sensor level). (A) The word frequency TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Shaded area indicates standard deviation. (B) word frequency TRF in the sentence condition. Individual lines represent sensors. Sensors in bold contributed to the clusters that allowed us to reject the null-hypothesis. (C) word frequency TRF in the list condition. Individual lines represent sensors. Sensors in bold contributed to the clusters that allowed us to reject the null-hypothesis. (D) The word onset TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Shaded area indicates standard deviation. Vertical gray lines indicate the time points of the scalp maps.

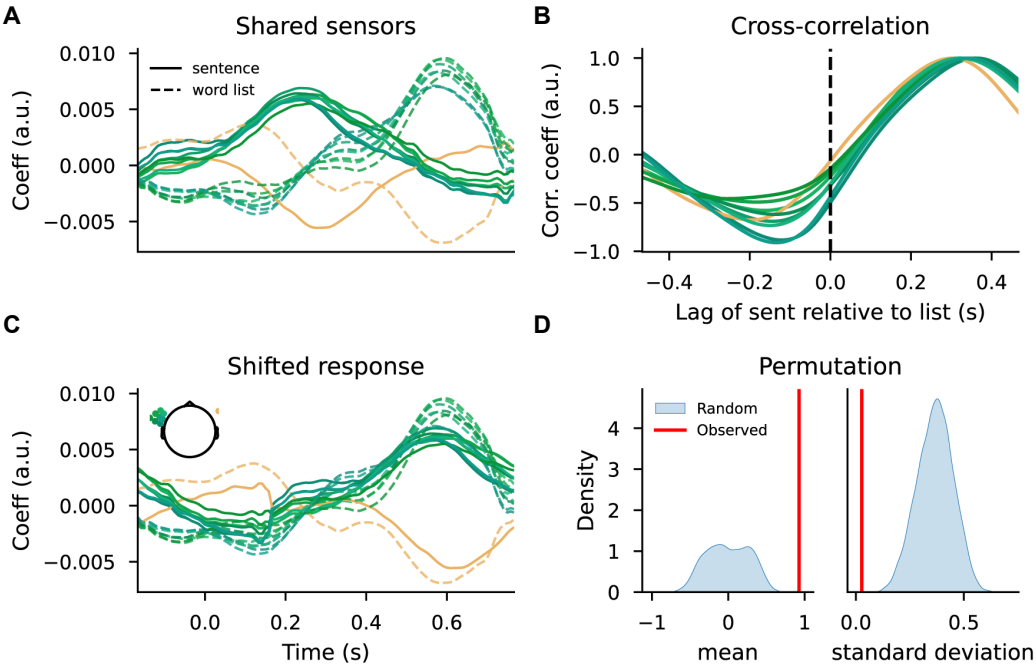


Figure 3.4: Cross-correlation analysis. (A) TRF time-courses for shared sensors between the sentence (solid lines) and word list (dashed lines). Colors indicate sensor position. (B) Cross-correlation between the sentence- and word list responses for overlapping sensors between conditions from the clusters (scaled between -1 and 1). Colors indicate sensor position. (C) The shifted response from the sentence condition (solid lines) to overlap with the word list condition (dashed lines). Colors indicate sensor position. (D) Kernel density plots of means and standard deviations from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation. Coeff.: coefficient.

The reconstruction accuracies were evaluated with the model $accuracies \sim condition \cdot (frequency + entropy + surprisal) + (1|participant)$. The explanatory value of the interaction between condition and each of the lexical factors was evaluated; each interaction significantly improved model fit (frequency: $\chi^2(1) = 6.88$, $p < 0.01$; entropy: $\chi^2(1) = 4.48$, $p < 0.05$; surprisal: $\chi^2(1) = 7.24$, $p < 0.01$), so the full model was interpreted. The results of this model are summarized in table 3.2 below.

Reconstruction accuracies were higher in the word list condition than in the sentence condition ($\beta = 1.67 \cdot 10^{-2}, SE = 9.43 \cdot 10^{-4}, t(1530) = 17.69, p < 0.01$). As can be seen in Figure 3.5A, each feature contributed positively to the reconstruction of the neural signal in the sentence condition; less so in the word

list condition, hinting at an interaction effect. Indeed, the factor *frequency* interacted with *condition* ($\beta = 2.47{\cdot}10^{-3}, SE = 9.43{\cdot}10^{-4}, t(1530) = 2.63, p < 0.01$), showing that reconstruction accuracies improved more from the addition of the word frequency predictor in the sentence condition, than in the list condition (Figure 3.5B). Further, although we will not discuss these effects, *entropy* and *surprisal* interacted with *condition*, as well (entropy: $\beta = 2.00 \cdot 10^{-3}, SE = 9.43 \cdot 10^{-4}, t(1530) = 2.12, p < 0.05$; surprisal: $\beta = 2.54 \cdot 10^{-3}, SE = 9.43 \cdot 10^{-4}, t(1530) = 2.69, p < 0.01$).

Table 3.2: Results of the LME on the reconstruction accuracies in the delta band.

Factor	β-coefficient	SE	df	t-value	p-value
(Intercept)	$8.61{\cdot}10^{-2}$	$1.82{\cdot}10^{-3}$	1306	47.22	***
Word frequency	$3.61{\cdot}10^{-4}$	$6.66{\cdot}10^{-4}$	1530	0.54	n.s.
Surprisal	$6.24{\cdot}10^{-4}$	$6.66{\cdot}10^{-4}$	1530	0.94	n.s.
Entropy	$-3.88{\cdot}10^{-4}$	$6.66{\cdot}10^{-4}$	1530	-0.58	n.s.
Condition	$-1.67{\cdot}10^{-2}$	$9.43{*}10^{-4}$	1530	-17.69	***
Word frequency * condition	$2.47{\cdot}10^{-3}$	$9.43{\cdot}10^{-4}$	1530	2.63	**
Surprisal * condition	$2.54{\cdot}10^{-3}$	$9.43{\cdot}10^{-4}$	1530	2.69	**
Entropy * condition	$2.00{*}10^{-3}$	$9.43{\cdot}10^{-4}$	1530	2.12	*

Note. **SE: standard error; df: degrees of freedom; n.s. not significant; * p < 0.05; ** p < 0.01; *** p < 0.001**

To gain more insight into the effect of *frequency*, we performed post-hoc t-test comparing the two largest models (Entropy/Surprisal and Full). These tests confirmed that the word frequency predictor enhanced reconstruction accuracy in the sentence condition (t(101)=5.35; p < 0.01), but not in the word list condition (t(101)=-0.15, p = 1) (Bonferroni-corrected).

Finally, we hypothesized that the higher reconstruction accuracy in the word list condition was due to the salience of isolated words, possibly evoking a larger auditory response. If this is true, a model with only the envelope predictor, and no word-level feature, should also fit the list condition better. To evaluate this hypothesis, we compared the reconstruction accuracies (averaged over all sensors) for the Envelope model between conditions. This model was not included in the analyses of the word frequency effect. And indeed, this was the case: reconstruction accuracies were higher for word lists than sentences using only the envelope as predictor (t(101)=13.40, p < 0.01).

In sum, the response to word frequency differed between word lists and sentences. The TRFs in sensor space revealed a left-lateralized frontotemporal response to the feature that peaked around ~250ms post word onset in the sentence condition, and around ~600ms in the word list condition. The sentence

effect is in line with other studies that used word frequency as a feature in TRF models of natural language comprehension (Brennan & Hale, 2019; Weissbart et al., 2019). A cross-correlation analysis between a set of left (and one right) temporal and frontal sensors that were involved in the response in both conditions suggested that the word list response peaks ~300ms later. The reconstruction accuracies in sensor space suggests that the word frequency predictor explains more variance over and above acoustics, entropy, and surprisal in the sentence condition, but not in the word list condition.

Source reconstruction In source space, we compared the TRFs for word lists and sentences using a cluster-based permutation test in two time-windows on the basis of the results from the analysis in sensor space: 200-400ms and 500-700ms post stimulus onset, respectively. The cluster-based permutation test on the TRFs from the source reconstructed MEG revealed two clusters in the early time-bin, and four clusters in the late time-bin. In line with the analysis in sensor space, coefficients were higher in the sentence condition than in the word list

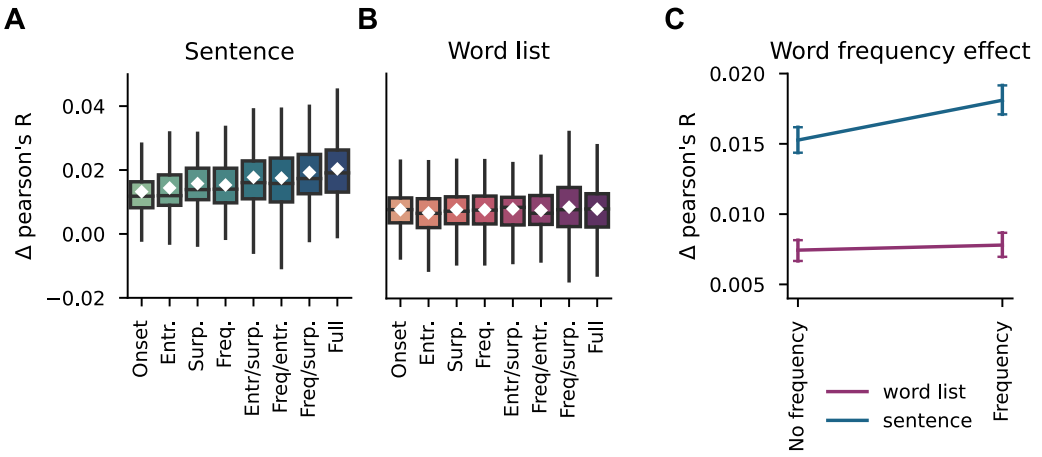


Figure 3.5: Reconstruction accuracies in the delta band. (A) Reconstruction accuracy difference with the envelope model for each model in the sentence condition. Middle line indicates the median, the withe diamond indicates the mean. (B) Reconstruction accuracy difference with the envelope model for each model in the word list condition. Middle line indicates the median, the withe diamond indicates the mean. (C) The interaction between condition and frequency on the reconstruction accuracies. Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval. Entr.: entropy, surp: surprisal, freq: frequency.

condition in the early time-bin (200-400ms PSO). These differences appeared bilaterally in the posterior superior- and middle frontal gyrus (dorsolateral- and dorsomedial prefrontal cortex) and cingulate gyrus (Figure 3.6A). In the right hemisphere, the cluster extended to the inferior frontal gyrus (Figure 3.6A).

In the late time-bin (500-700ms PSO, Figure 3.6B), coefficients were higher in the word list condition than in the sentence condition in three out of four clusters. Those clusters appeared in the left hemisphere in the posterior temporal lobe across the superior, middle, and inferior gyri/sulci, the temporal pole, and the parahippocampal gyrus. In the right hemisphere, the effects appeared in superior temporal, inferior parietal, and caudal frontal areas, as well as cingulate gyrus. In a final cluster in the late time bin, the coefficients were higher in the sentence- than in the word list condition. This cluster spanned left inferior frontal areas, orbital cortex, as well as a small portion of the middle frontal gyrus.

In addition, we observe a difference between the responses in left orbitofrontal and ventrolateral prefrontal cortex – including the inferior frontal gyrus. In this area, the response peaks in the late time-bin in the sentence condition only. That this area is where we find a difference in late time lags is not surprising given the large literature implicating the left inferior frontal cortex, or Broca's area, in syntactic processes processes (Friederici, 2011, 2012, 2015; Hagoort, 2013, 2015; Matchin & Hickok, 2020).

Given our finding that the word list response appeared delayed in comparison to the response in the sentence condition, we also considered responses in the sentence- and word list conditions separately through one-sample cluster-based permutation tests. Here, we observed a widespread response in both conditions; and indeed, this response appears in the *early* time window in the sentence condition (Figure 3.6C), and in the *late* time window in the word list condition (Figure 3.6F).

As we already observed in the contrast, in the late time window, the response to word-internal information encompasses the left posterior superior- middle- and inferior temporal gyrus (including parahippocampal gyrus) and the temporal poles, as well as bilateral somatosensory areas in both conditions. These areas are traditionally associated with lexical (and) semantic memory (Binder & Desai, 2011; Hagoort, 2013, 2015). Furthermore, as we observed in the early time-window, this response includes the bilateral dorsolateral prefrontal cortex. These areas are part of the dorsal attention network and have been implied to control activation and selection of information stored in temporoparietal cortices

(Binder & Desai, 2011). In addition, like we observed in the contrast between conditions, in the sentence condition a late response appears in the left inferior frontal gyrus (Figure 3.6E). This response was absent in the word list condition.

We compared the reconstruction accuracies using a cluster-based 2x2 ANOVA. There were no significant differences (all $p > 0.1$).

Taken together, these findings indicate that (1) much, but not all, of the response to word internal information is shared between conditions in space; (2) the response develops differently in time, with a delay in the word list condition; and (3) word internal information modulates activity in the left inferior frontal gyrus only in the presence of a coherent context.

3.3.3 Theta band

Sensor-level analysis In the theta band, the cluster-based permutation test revealed no differences between the word list and sentence TRFs for the word frequency feature (see Figure 3.7). The one-sample tests indicated, however, a response between 100 and 200 milliseconds in the word list condition that was absent in the sentence condition.

Like in the delta band, the full model was *accuracies* $\sim$ *condition* $\cdot$ *(frequency + entropy + surprisal) + (1|participant)*. Removing the interaction between *frequency* and *condition*, or the interaction between *surprisal* and *condition*, decreased model fit (marginally; frequency: $\chi^2(1) = 3.80$, $p = 0.051$; surprisal: $\chi^2(1) = 3.95$, $p < 0.05$), but removing the interaction between *entropy* and *condition* did not ($\chi^2(1) = 0.47$, $p = 0.49$). We continued with the model *accuracies* $\sim$ *condition* $\cdot$ *(frequency + surprisal) + entropy + (1|participant)*. The AIC comparison confirmed that this model was the best descriptor of the data. The results of this model are summarized in Table 3.3.

In theta, too, there was a main effect of condition ($\beta = 2.09 \cdot 10^{-3}, SE = 6.90 \cdot 10^{-4}, t(1530) = 3.02, p < 0.01$), with reconstruction accuracies being higher in the word list condition than in the sentence condition (see Figure 3.8). In addition, there was a main effect of *frequency* ($\beta = 1.17 \cdot 10^{-3}, SE = 5.64 \cdot 10^{-4}, t(1530) = 2.07, p < 0.05$) indicating that generally, the addition of word frequency improved reconstruction accuracy. The interaction between *frequency* and *condition* approached, but did not reach significance ($\beta = 1.56 \cdot 10^{-3}, SE = 7.97 \cdot 10^{-4}, t(1530) = 1.95, p = 0.051$), indicating a potential trend for the frequency effect to be larger in the sentence condition than in the word list condition (Figure 3.7).

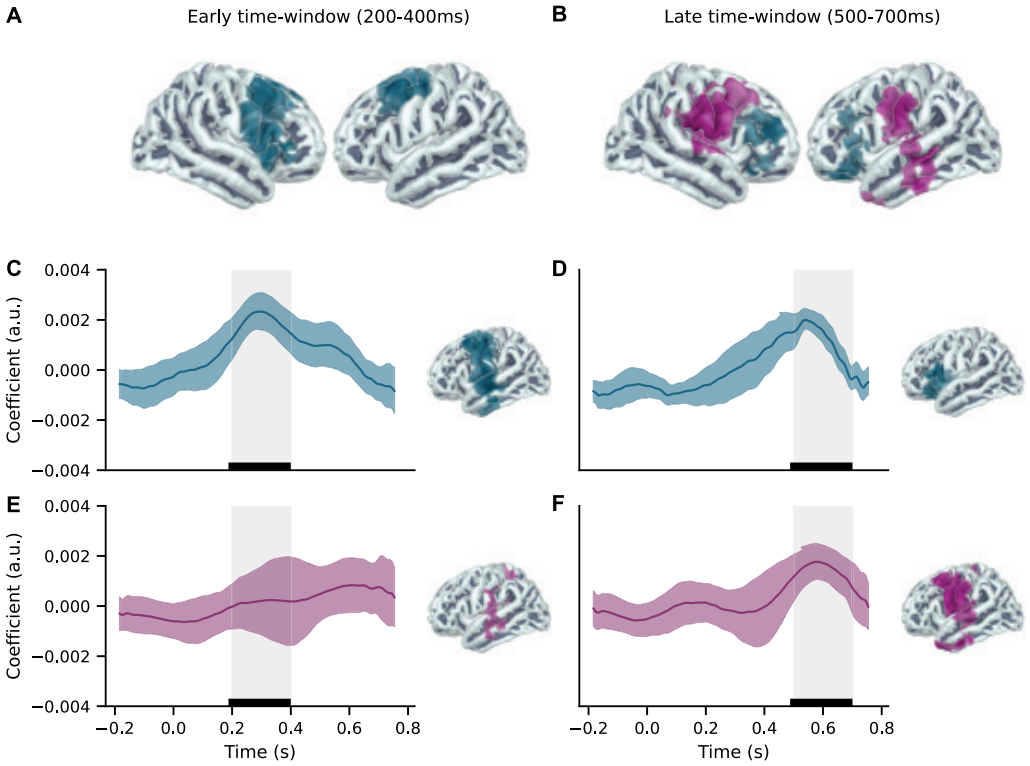


Figure 3.6: Clusters from the delta-band TRFs in source space. Left column: early time-window (200-400ms). Right column: late time-window (500-700ms). (A) Top left: Differences between the word list and sentence responses to word frequency in the early time window. Blue indicates that coefficients sentence > word list; pink indicates word list > sentence. (B) Top right: Differences between the word list and sentence responses to word frequency in the late time window. Blue indicates that coefficients sentence > word list; pink indicates word list > sentence. (C) Middle left: sentence condition. TRF and spatial distribution of one-sample cluster in early time-window. Time-window is indicated in grey. (D) Middle right: sentence condition. TRF and spatial distribution of one-sample cluster in late time-window. Time-window is indicated in grey. (E) Bottom left: word list condition. TRF and spatial distribution of one-sample cluster in early time-window. Time-window is indicated in grey. (F) Bottom right: word list condition. TRF and spatial distribution of one-sample cluster in late time-window. Time-window is indicated in grey. Shaded areas in blue and pink indicate SD.

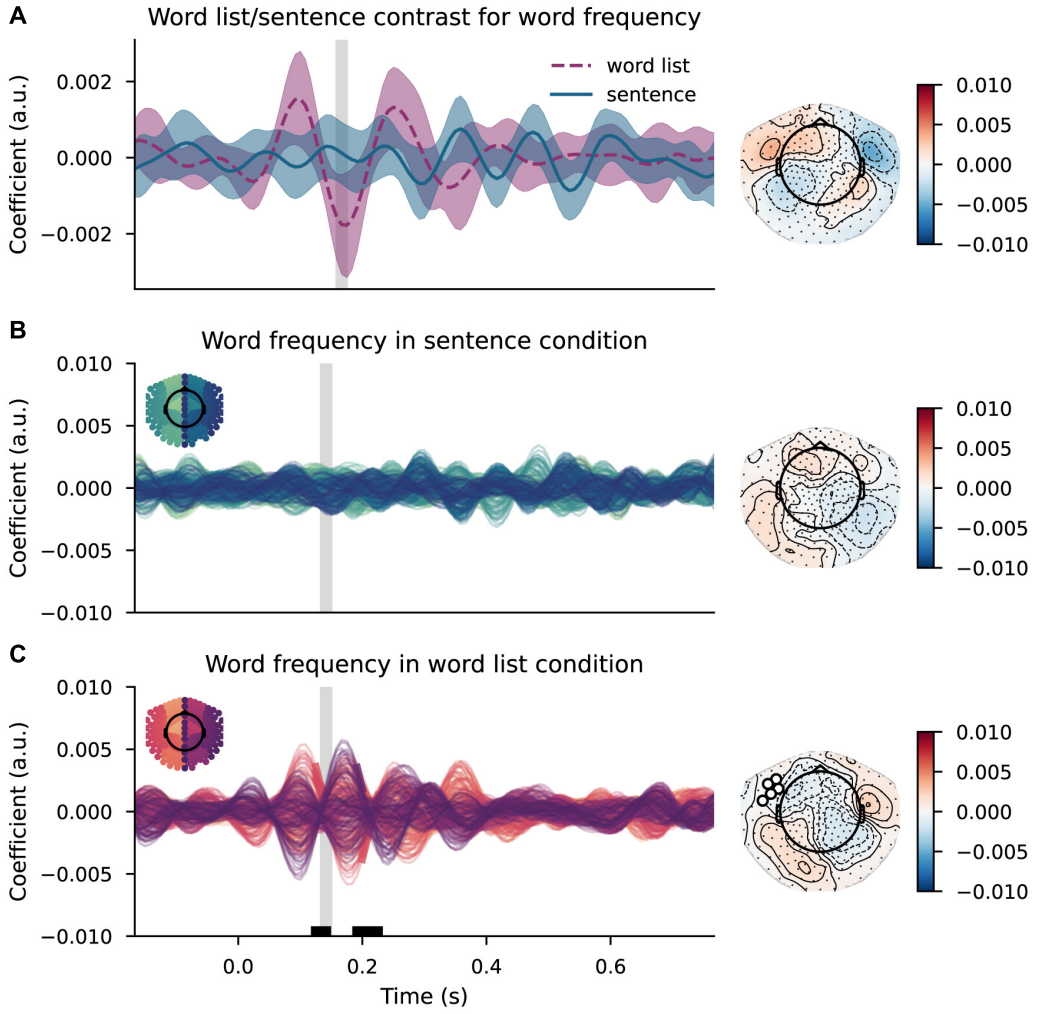


Figure 3.7: Theta-band effects (sensor level). (A) The word frequency TRF in both conditions in the theta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters. Shaded area indicates standard deviation. (B) The word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. (C) The word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Vertical gray lines indicate the time points of the scalp maps.

With respect to the other predictors, there was a positive effect of *entropy* ($\beta = 2.43 \cdot 10^{-3}, SE = 3.99 \cdot 10^{-4}, t(1530) = 1.95, p < 0.01$), and an interaction between *condition* and *surprisal* ($\beta = 1.55 \cdot 10^{-3}, SE = 7.92 \cdot 10^{-4}, t(1530) = 1.99, p < 0.05$), indicating that *surprisal* enhanced reconstruction accuracies more in the sentence condition than in the word list condition.

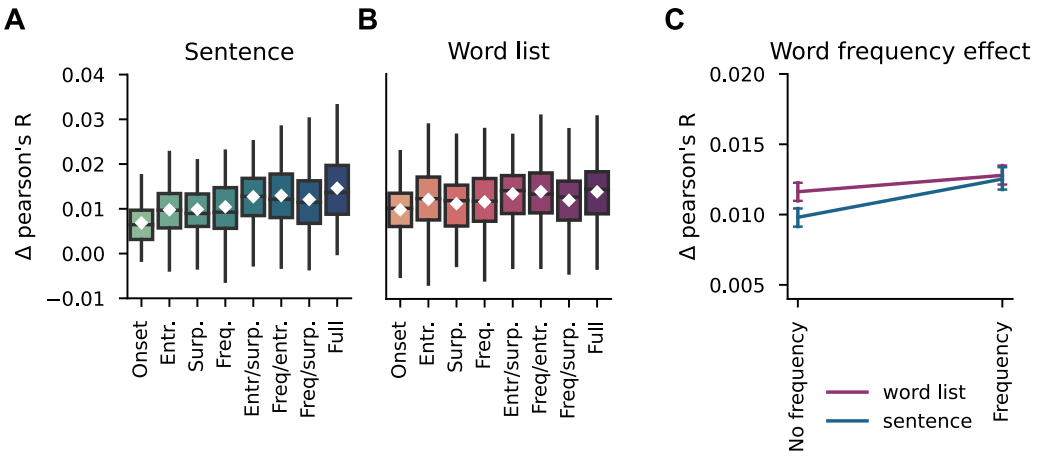


Figure 3.8: Reconstruction accuracies in the theta band. A) Reconstruction accuracy difference with the envelope model for each model in the sentence condition. Middle line indicates the median, the withe diamond indicates the mean. (B) Reconstruction accuracy difference with the envelope model for each model in the word list condition. Middle line indicates the median, the withe diamond indicates the mean. (C) The interaction between condition and frequency on the reconstruction accuracies (p = 0.051, see section 3.3.3). Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval. Entr.: entropy, surp: surprisal, freq: frequency.

Again, we performed post-hoc t-tests comparing the two largest models (Entropy/Surprisal and Full) to gain more insight in the effect of word frequency on the reconstruction accuracies. These showed that the word frequency predictor enhanced reconstruction accuracies in the sentence condition (t(101) = 5.67; p < 0.01), but not in the word list condition (t(101) = 1.48; p = 0.57). There were no effects of condition for these two models (all p = 1).

Source-reconstruction Given that the permutation test in the sensor-based analysis did not reveal any effects in the theta band and we could not select time-bins a priori, we performed a cluster-based permutation test on the full TRF. This revealed two clusters in the right hemisphere between 100 and 250ms.

Table 3.3: Results of the LME on the reconstruction accuracies in the theta band.

Factor	β-coefficient	SE	df	t-value	p-value
(Intercept)	$4.02 \cdot 10^{-2}$	$1.29 \cdot 10^{-3}$	1382	31.04	***
Word frequency	$1.17 \cdot 10^{-3}$	$5.64 \cdot 10^{-4}$	1530	2.07	*
Surprisal	$7.26 \cdot 10^{-4}$	$5.64 \cdot 10^{-4}$	1530	1.29	n.s.
Entropy	$2.43 \cdot 10^{-3}$	$3.99 \cdot 10^{-4}$	1530	6.10	***
Condition	$-2.09 \cdot 10^{-3}$	$6.90 \cdot 10^{-4}$	1530	-3.02	**
Word frequency · condition	$1.55 \cdot 10^{-3}$	$7.97 \cdot 10^{-4}$	1530	1.95	n.s.
Surprisal · condition	$1.59 \cdot 10^{-3}$	$7.97 \cdot 10^{-4}$	1530	1.99	*

Note. **SE: standard error; df: degrees of freedom; n.s. not significant; * p < 0.05; ** p < 0.01; *** p < 0.001**

Both of these clusters reflect a larger amplitude across right frontal and temporal areas for the TRF in the word list condition than the sentence condition, as can be seen in the plots of the time courses of the clusters in figure 3.9 below. These effects, although visible in figure 3.7A-C, did not reach significance in the sensor-analysis, potentially due to the stringent threshold (recommended value multiplied by three) chosen there.

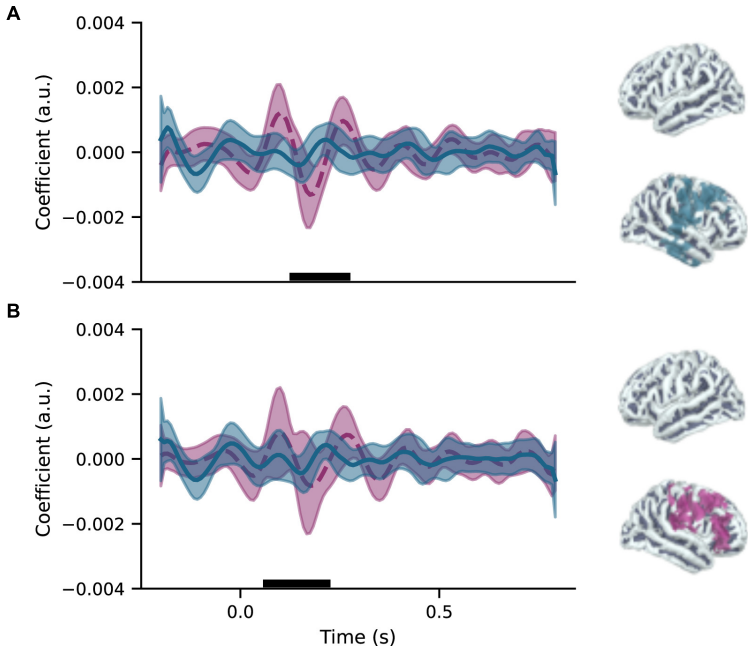


Figure 3.9: Clusters from the theta-band TRFs in source space. Blue indicates that coefficients sentence > word list; pink indicates word list > sentence. (A) top: right-lateralized cluster where TRF sentence > word list. (B) bottom: right-lateralized cluster where TRF word list > sentence. Shaded areas in blue and pink indicate SD.

3.3.4 Control analysis I: Data from Ten Oever, Carta, et al. (2022)

In the delta band, the cluster-based permutation test revealed no significant differences between the word frequency response in the word lists and sentences. To evaluate if this was due to there being no detectable responses or no difference between conditions, we performed one-sample cluster-based permutation tests. Here we observed a response in the sentence condition over a large array of left-posterior sensors that was significant from word onset to about 400 milliseconds. The peak appears around 200 milliseconds (Figure 3.10A). Although figure 3.10B suggests a potential response around 400 milliseconds in the word list condition, there were no significant clusters. As in the main analysis, there were no significant differences between conditions in the responses to word onset.

The absence of a difference between the conditions and the lack of a detectable response in the word list condition alone make the results from this analysis difficult to interpret in relation to the main analysis. The large difference between the sample sizes (N=102 vs N=16, respectively) may play a role in this difference. We performed a power analysis on the difference between the conditions in the control analysis using the average t-values from the time-points and sensors taken from the significant clusters from the same contrast in the main analysis. This showed that power would increase on average by 30.7% when taking a sample of 102 participants, with three clusters reaching a power of above 96%. This suggests that the control analysis did not have enough power to reject or confirm the hypothesis that the delay in the response in the word list condition is caused by the different temporal dynamics in the original analysis. We therefore refrain from drawing conclusions on the basis of this finding.

Nevertheless, the ANOVA on the reconstruction accuracies revealed a main effect of model ($F(1,15)=38.01$; $p < 0.01$), indicating that the word frequency predictor enhanced reconstruction accuracy, and an interaction between condition and model ($F(1,15)=6.79$; $p < 0.05$), suggesting that this effect was larger for the sentence condition than for the word list condition (Figure 3.10C). There was no main effect of condition ($p = 0.16$).

In the theta band, there were no significant effects on the TRF waveforms nor on the accuracy values (figure 3.11).

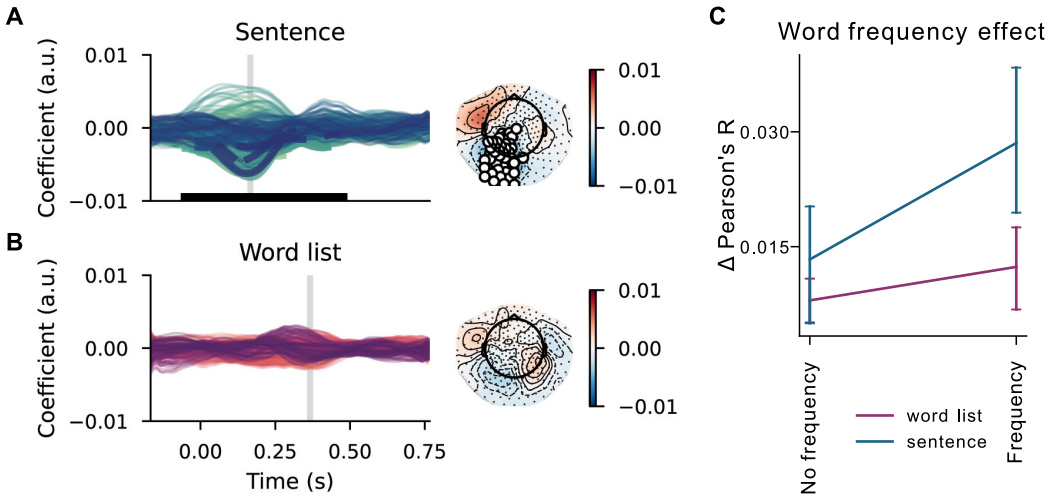


Figure 3.10: Delta-band effects in the extra data. (A) Word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters. (B) Word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters (none). (C) The interaction between condition and frequency on the reconstruction accuracies. Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval. Vertical gray lines indicate the time points of the scalp maps.

3.3.5 Control analysis II: Simulations

In order to evaluate the effect of differences in inter-stimulus interval (i.e., pauses), we simulated raw MEG data consisting of a signal (different impulse responses) and optional noise. Strikingly, the inter-stimulus interval has no direct influence on the reconstruction score, although the length of the segment on which we estimate the score does (Figure 3.12). In this case, the difference in interstimulus interval – which eventually leads to a difference in data length – shows how the bias in score observed between conditions is solely due to the difference in duration. The bias, however, is constant, and should be controlled for when directly comparing models *within* conditions. Moreover, we actually observe the opposite effect in our MEG analysis: the absolute scores for the longer segment of data (the word lists) are higher than the shorter segment of data (the sentences). This means that our score differences exist above and beyond any bias generated from the stimulus difference.

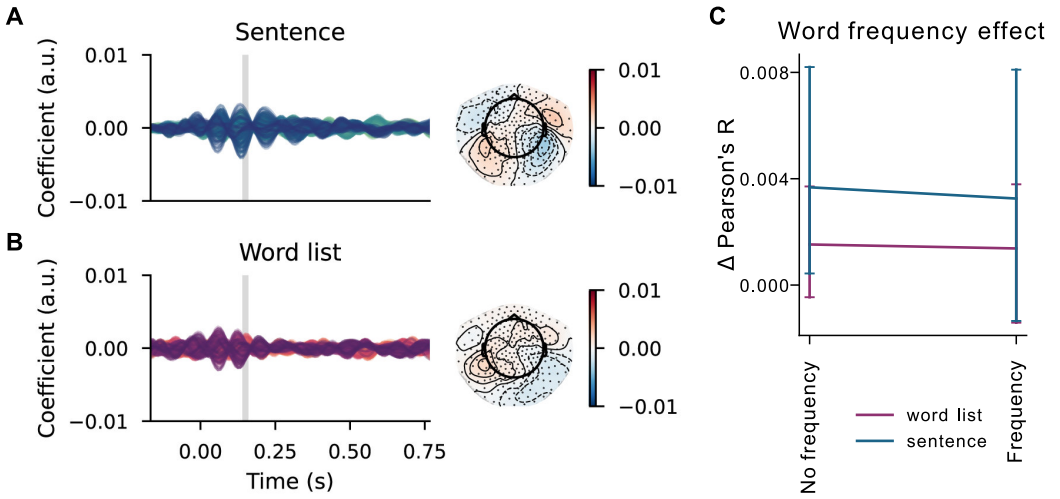


Figure 3.11: Theta-band effects in the extra data. (A) Word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters. (B) Word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters (none). (C) The (lack of an) interaction between condition and frequency on the reconstruction accuracies. Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval. Vertical gray lines indicate the time points of the scalp maps.

3.4 Discussion and conclusions

In this study, we asked whether low-frequency neural readouts associated with words systematically changed as a function of being in a sentence context, and whether neural readouts were modulated by purely lexical properties over and above sensory- and contextual distributional variables. We contrasted responses to word frequency for words in sentences with word lists, the latter lacking any syntactic structure and combinatorial lexical meaning. We hypothesized that the delta-band, but not theta-band, responses to word frequency would be different in word lists and sentences as a consequence of the (in)availability of sentence context. Specifically, following findings from speech tracking, we expected a stronger presence of the word frequency response in the sentence condition.

Our findings showed that the delta band response to word frequency differs between word lists and sentences in time and, albeit minimally, in space. In both conditions, word internal information modulates a response across the left tem-

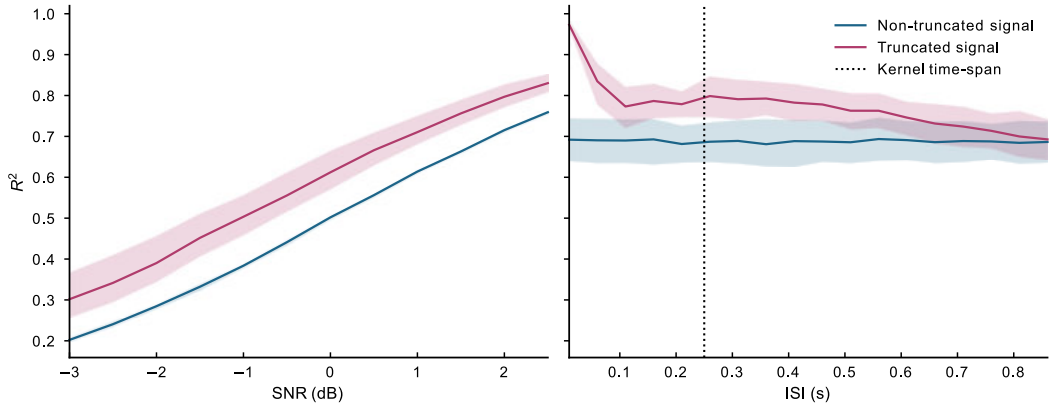


Figure 3.12: Influence of interstimulus interval (ISI), data length, and noise on score (reconstruction accuracy; R^2). The left panel shows the (proportional) influence of broadband signal-to-noise ratio (SNR) on score. The right panel shows that for every interstimulus interval value, the same score is measured if the data length is kept constant; and the score deflated for longer signals as more noise is being evaluated in the scoring.

poral lobe and the frontal cortex. However, this response occurred about 300ms earlier in the presence of coherent sentence context. In addition, in sentence context, word internal information could be seen to modulate activity in the left inferior frontal gyrus at around 600ms post word onset; a response that is absent when a word is not embedded in a sentence. Furthermore, the word frequency feature explains more variance over and above the other features in the sentence condition than in the word list condition. In the theta band, there were only minimal differences between the conditions. We will discuss our results in more detail below.

In psycholinguistic theories of word recognition, word frequency is often modeled as the baseline of activation or the prior probability of a word (e.g., the Logogen model (Morton, 1969); Cohort model (Marslen-Wilson, 1987); Shortlist-A and B, (Norris, 1994; Norris & McQueen, 2008)). We assume therefore that the neural readout associated with word frequency represents neural activity during the process of word recognition. Our results provide direct evidence that this process happens differently depending on whether the structure-building of sentence comprehension is also occurring. We know that words are recognized faster when they are embedded in a coherent sentence context (Marslen-Wilson & Welsh, 1978; Tyler & Wessels, 1983); this is reflected in the delayed word list response to word frequency (see also: Lam et al., 2016).

Furthermore, the reconstruction accuracies in sensor space suggest that the response to word frequency explains more variance in the sentence condition than in the word list condition. This may seem contradictory to findings from psycholinguistics. Indeed, the behavioral effect of word frequency, when assessed with reaction time measures, diminishes in sentence context (Schuberth & Eimas, 1977; Simpson et al., 1989; Tyler & Wessels, 1983). Put differently, words with a low frequency are recognized more slowly than words with a high frequency. This does not necessarily mean that lexical information explains less variance in the neural signal. In fact, studies that consider metrics like mutual information between the brain and the speech signal find that the brain represents aspects of the speech signal more reliably when more linguistic information is present (e.g., Kaufeld, Bosker, et al., 2020; Ten Oever, Carta, et al., 2022), while the acoustic information in speech matters less for word recognition when the word is embedded in a sentence (Boothroyd & Nittrouer, 1988; Mattys, Davis, Bradlow, & Scott, 2012). In general terms, these findings suggest that the brain represents lower-level features more reliably when higher-level information can be inferred, while the lower-level information itself becomes less important for the outcome of the task. Indeed, that words are represented more robustly when sentence context is provided is reflected in the accuracy scores on the word monitoring task performed in this study: participants were more likely to correctly remember whether a word was mentioned or not when they had been presented with a sentence, than when they heard a word list.

There are two causes for this finding. Firstly, the perceptual salience of the words in the word list condition leads to a large response to the speech envelope; the response to lexical features then are of *relatively* lower power, and explain less of the variance in the signal relative to the lower-level features. Secondly, as a consequence of words being embedded in larger structures – phrases and sentences – word frequency is likely present in a larger neural network in the sentence condition than in the word list condition (Martin, 2020). The signal is therefore reconstructed better in a wider array of sensors, leading to an overall larger increase in reconstruction accuracies. As discussed below, the presence of the effect in the control analysis favors the latter interpretation. The propagation of lexical information to a wider network is additionally reflected in the differences between conditions in the inferior frontal gyrus at approximately 600ms. This interpretation is consistent with findings that show that sentence structure influences the dynamics and distribution of neural signals (Bai et al., 2022; Blank, Balewski, Mahowald, & Fedorenko, 2016; Coopmans et al., 2022;

Grodzinsky, Pieperhoff, & Thompson, 2021; Matchin, Brodbeck, Hammerly, & Lau, 2019; Matchin, Liao, Gaston, & Lau, 2019; Schell, Zaccarella, & Friederici, 2017; Ten Oever, Carta, et al., 2022).

Importantly, both the TRF and the reconstruction accuracy effects of sentence context on the representation of word-internal information are independent of (1) the contextual probability predictors surprisal and entropy, and (2) sensory information in the speech envelope. Each of these predictors are undoubtedly important for how the neural signal represents lexical information (e.g.; sensory: Doelling, Arnal, Ghitza, and Poeppel (2014); and probability: Weissbart et al. (2019)). Given that these influences were accounted for by the encoding model, the differences that remain imply a role for abstract structure and meaning on the transformation of low-frequency neural readouts associated with words (or more minimally, associated with purely lexical features). These conclusions are in line with findings on the visual part of the dataset, not analyzed here (Huizeling et al., 2022).

Striking also is the difference between the effects in the delta and theta band. In the theta band, the responses to word frequency differed between conditions only slightly: the amplitude of the response was larger in the word list condition than in the sentences in the right frontal and temporal hemisphere around 100 milliseconds – possibly indicating that word frequency in interaction with contextual information tunes sensory sampling. The addition of the word frequency predictor had a small effect on the reconstruction accuracies, which was present only in the post-hoc analysis. In general, theta band activity appears to be more sensitive to perceptual aspects of the stimulus than to linguistic aspects. For example, tracking of sound by theta band activity persists even in the absence of linguistic information (Molinaro & Lizarazu, 2018), while it is affected when acoustic edges in the stimulus are experimentally manipulated (Doelling et al., 2014). However, in line with the differences that we do see, Donhauser and Baillet (2020) showed that the gain of early theta responses varies according to the contextual uncertainty of speech. The results from the present analysis are consistent with an account in which the theta band is important for speech processing, but not as central for the representation of higher-level features such as lexical-internal information. At the same time, the process reflected by theta modulations during language comprehension is likely to be influenced linguistic context.

In addition to the linguistic differences, there was a variable pause between the words in the word list condition only. To examine the potential effect of

this additional difference between the conditions on our results, we ran several simulations. The simulations showed that the interstimulus interval (ISI) between events modelling word-like responses has no effect on model evaluation and TRF estimation. However, there will be a constant bias in the model score that is proportional to the broadband signal-to-noise ratio (where the noise is the additive contribution beyond variance explained by the linear model). This bias is not directly due to the differences in ISI, but rather to the fact that we are integrating a larger portion of data in the list condition – thus more noise to contribute to the score. As such, any model comparison contrasting scores within condition will eliminate the constant bias. Furthermore, this bias leans toward deflating the score of the model evaluated on the longest segment of data (the word list condition). We found that with the envelope alone, the scores in the list condition were *higher* than the scores of the sentence condition; this is in direct contrast with the expectations from the simulations. From these simulations we conclude therefore that the delay in the TRF waveform and the interaction effect in the reconstruction of the neural signal are not just due to difference in signal length between the word list and sentence condition. The next question is, then: what are the potential cognitive effects of silence between the words? There are three potential effects: (1) higher perceptual saliency of each word, already mentioned above; (2) decreased word rate; and (3) absence of phonological cues between words, such as prosody and co-articulation.[2] We consider phonological cues to be consequences of as well as cues to the sentence context; they would be different between word lists and sentences in naturalistic conditions as well. The first two, however, need some consideration.

As was mentioned above, the perceptual difference between two consecutive words is much smaller than the difference between silence and a word. This effect was visible in the speech-brain coherence for both conditions (Figure 3.1; coherence was much higher in the word list condition in the delta band), and caused overall higher reconstruction accuracy in the word list condition. Importantly, in the analysis on a second dataset in which this difference between conditions did not exist, the interaction effect between word lists and sentences was replicated: the word frequency feature explained more variance over and above the envelope- and word onset predictors in the sentence condition than in the word list condition. Furthermore, we stipulated that a general delaying effect on word processing generated by the decreased word rate in the word

[2]A reviewer suggested we add a prosody predictor. We constructed a prosody predictor by extracting the prosody contour using Parselmouth, a Praat wrapper for Python. Running the analysis with this extra predictor did not qualitatively change the results.

list condition would be visible at *other* features as well. Nevertheless, the word onset feature – the only feature besides word frequency that was numerically identical between conditions – did not show such difference. These findings indicated that it was only the response to *word-internal* information that was delayed, and suggests that the brain processes lexical information later in the absence of a coherent sentence context.[3] Taken together, this indicates that the effects described in this work are unlikely to be driven by silence.

In summary, this study suggests that delta band, and to a lesser extent, theta band, responses to word-internal information are affected by sentence context in *time* and in *space*. Given that a difference in encoding of a strictly lexical feature persists when context-driven lexical features like entropy and surprisal are added, we conclude that low-frequency responses to word internal information are changed by sentential structure and meaning, and not by probabilistic differences alone. In the delta band, a lexical response across the posterior and anterior left temporal lobe and the bilateral parietal lobe, is delayed in the absence of sentence context. In addition, a word's embedding in sentence context determines whether inferior frontal areas are responsive to lexical information. In the theta band, a larger amplitude in the word lists at about 100 milliseconds across the right frontal and parietal areas suggests that linguistic information can tune sensory sampling. In addition, this study shows that the TRF can be used to model acoustic differences between stimuli when measuring higher-level linguistic effects (see also Bai et al., 2022). The results of this study speak to how the neural representation of words is affected by the linguistic structure of sentence context, and as such provide beginning insight into how the brain instantiates compositionality in language processing.

[3] Of course, the possibility that the brain performs **linguistic** computations conservatively in time - at the highest speed necessary and lowest speed possible – is not excluded. Such a mechanism would lead to any word-level processing beyond segmentation being 'slowed down' in the case of decreased word rate or even word presentation in isolation. Further research is required to examine this possibility.

4 | Surprisal is not enough: Additive effects of grammaticality and lexical surprisal in self-paced reading

Abstract

Language comprehension requires the integration of information from a wide variety of sources, both from the input and retrieved from memory. The present study contributes to a growing literature examining how probability and uncertainty shape language comprehension in close collaboration with grammatical knowledge. There is much evidence for the influence of abstract structure building and probabilistic processing alike. This suggests that the process of language comprehension is not only shaped by morphosyntactic processing, or that it is wholly determined by the statistical probability. Instead, it is likely that both factors play a role. Here we asked about their impact during the process of binding subject and verb in Dutch (subject-verb agreement). In an online self-paced reading paradigm, we tested whether lexical surprisal affects the use of grammatical information. The results indicated that both lexical probabilistic information as well as grammatical information are needed to describe reading time data from a subject-verb agreement paradigm. This is in direct contrast with proposals that model this phenomenon and language comprehension more generally using exclusively lexical probabilistic information. At the same time, the results suggested that the morphosyntactic cue provided by the subject or the verb in subject-verb agreement in Dutch is stronger than the cue of contextual lexical probability as used here: the ungrammaticality effect was not altered by lexical probability. In addition, the data provided some evidence that lexical probability is leveraged more reliably when the constraints placed by the grammar are obeyed. Taken together with previous findings, the results suggest a process of language comprehension in which grammatical cues, as well as contextual probabilistic cues, are weighted on the basis of their reliability.

4.1 Introduction

Language comprehension requires the integration of information from a wide variety of sources, both from the input and retrieved from memory. Two sources that are important for the generation of linguistic meaning that both depend on linguistic knowledge stored in memory are lexical and morphosyntactic knowledge. To form a clause, lexical information must be combined with morphosyntactic information. How this happens at a mechanistic level is to date unknown.

An important step in the formation of a clause is the establishment of an agreement relation between the subject and the verb. This process, subject-verb agreement, has received considerable attention in the psycholinguistic literature, and has given rise to a mechanistic account of sentence processing called *cue-based retrieval* (Bock & Miller, 1991; Brehm, Hussey, & Christianson, 2020; Lewis, Vasishth, & Dyke, 2006; Nicol, Forster, & Veres, 1997; Pearlmutter, Garnsey, & Bock, 1999; Tanner, Nicol, & Brehm, 2014; Van Dyke & McElree, 2006; Vasishth, 2001). On this view, subject-verb agreement is established by the retrieval of the preceding noun from memory on the basis of retrieval cues at the verb. At the same time, current work on self-paced reading and other behavioral metrics suggests that contextual lexical probability shapes the processing of language input in a non-trivial way: words that are statistically unpredictable from the context are consistently associated with longer reading times than words that are statistically predictable from the context (Aurnhammer & Frank, 2019; Brothers & Kuperberg, 2021; Monsalve et al., 2012, among others). These findings are interpreted in a model called *surprisal theory* (Hale, 2006, 2016; Levy, 2008a; Levy & Gibson, 2013).

There is much evidence for abstract structure building and probabilistic processing alike, both of which are reviewed below. This suggests that the process of language comprehension is not only shaped by morphosyntactic processing, or that it is wholly determined by the statistical probability. Instead, it is likely that both factors play a role (Martin, 2016, 2020). Here we asked about their impact during the process of binding subject and verb in Dutch. In an attempt to understand the cognitive architecture for language that can give rise to both effects of a morphosyntactic nature, and effects with a lexical probabilistic origin, we evaluate whether probabilistic contextually-driven lexical pre-activation interacts with the establishment of subject-verb number agreement in Dutch. To this end, we perform a self-paced reading study in which we pit probabilistic lexical information against subject-verb agreement. The results are interpreted in

light of models of language processing (production/comprehension): cue-based retrieval, surprisal theory, and language processing as cue-integration.

4.1.1 Subject-verb agreement in comprehension

Subject-verb agreement, the systematic co-occurrence of agreement features between the subject and the verb, encodes a formal link between the subject and the predicate. In Dutch, like other Germanic languages, there is subject-verb agreement in number, as shown in example (1). In (1a), the subject *de zwemster* 'the swimmer' is singular, as is the verb *won* 'won'; a correct agreement relationship is established. Compare this to (1b), where *de zwemster* remains singular, while the verb *wonnen* 'won' is now plural; this is an agreement error, and the sentence is anomalous.

(1) a. De zwemster won de competitie.
The $\text{swimmer}_{\text{F,SIN}}$ won_{SIN} the competition

'The swimmer won the competition.'

b. *De zwemster wonnen de competitie.
the $\text{swimmer}_{\text{F,SIN}}$ won_{PL} the competition

?'The swimmer won_{PL} the competition.'

c. De zwemsters wonnen de competitie.
The $\text{swimmers}_{\text{F,PL}}$ won_{PL} the competition

'The swimmers won the competition.'

d. *De zwemsters won de competitie.
The $\text{swimmers}_{\text{F,PL}}$ won_{SIN} the competition

?'The swimmers won_{SIN} the competition.'

For sentence comprehension, this entails that upon encountering a plural noun as a subject, such as *zwemsters* 'swimmers' in (1c), the reader or interlocutor can expect a plural verb form (*wonnen* 'won_{PL}', in (1c)).

Although speaking, signing and writing requires establishing agreement between the subject and verb in every sentence, speakers, signers and writers occasionally make agreement errors. This often happens when the subject noun and the verb are separated by other materials containingnouns, which can function as candidates (*attractors*) for the agreement relation (Bock & Miller, 1991).

This is a phenomenon known as *agreement attraction*. Consider the example in (2) below.

(2) a. The key to the cabinet was locked in the desk.

b. The key to the cabinet *were locked in the desk.

c. The key to the cabinets was locked in the desk.

d. The key to the cabinets *were locked in the desk.

Studies in several languages have shown that speakers were more likely to choose the wrong number on the verb when the first noun (*the key*) was singular, and the second noun was plural (*the cabinet(s)*) example (2c/d) than in the reversed configuration (Bock & Miller, 1991) (i.e., *cabinet;* see (3)). This asymmetry has been explained by the marked plural form inadvertently overriding the unmarked singular form upon establishing the agreement relation (Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001).

(3) a. The keys to the cabinet were locked in the desk.

b. The keys to the cabinet *was locked in the desk.

c. The keys to the cabinets were locked in the desk.

d. The keys to the cabinets *was locked in the desk.

In addition, there is a large literature on the processing of subject-verb agreement in comprehension (Acuña-Fariña, Meseguer, & Carreiras, 2014; Hagoort, Brown, & Groothusen, 1993; Lago, Acuña Fariña, & Meseguer, 2021; Mancini, Postiglione, Laudanna, & Rizzi, 2014; Nicol et al., 1997; Pearlmutter et al., 1999; Tanner & Bulkes, 2015; Tanner et al., 2014; Wagers, Lau, & Phillips, 2009). This literature has shown that comprehenders are sensitive to subject-verb agreement violations in the input. We will call this the *ungrammaticality effect*.

Pearlmutter and colleagues (1999) investigated subject-verb agreement in comprehension using a self-paced reading task. In this task, participants are asked to read sentences. The sentence is not visible at once; instead, the participant presses a button to reveal the next word (or phrase). Upon this button press, the previous word (or phrase) disappears. The speed with which the participant moves from word to word has proven to be indicative of several linguistic processes. For example, the speed of word reading depends on word length (Barton, Hanif, Eklinder Björnström, & Hills, 2014), and participants read more slowly when the sentence is ambiguous (Traxler, 2005). Similarly, in the study by

Pearlmutter and colleagues, the self-paced reading task (and a subsequent eye-tracking study) revealed the ungrammaticality effect: participants read verbs with an agreement error more slowly than correct verbs.

Experiments using electroencephalography (EEG) have provided an additional signature of the ungrammaticality effect (Hagoort et al., 1993; Osterhout & Mobley, 1995; Tanner, Grey, & van Hell, 2017; Tanner et al., 2014). These studies presented participants with grammatical and ungrammatical sentences while their EEG was recorded. Ungrammatical sentences elicit a larger P600-component (i.e., more positive values in a time-window between 400 and 800 milliseconds after word onset with a peak around 600 milliseconds) than grammatical sentences, which was in some cases preceded by a left-anterior negativity (LAN) (Osterhout & Mobley, 1995).

The ungrammaticality effect, as shown with the P600 (+LAN) or reading times, is relatively robust. It has been replicated in different languages, among which Italian (Mancini et al., 2014), Spanish (Acuña-Fariña et al., 2014), and Dutch (Hagoort et al., 1993). Wagers and colleagues (2009) showed that the increased reading times for agreement errors persist until the third word after the grammatical error. In addition, when the subject and verb are immediately adjacent, ungrammaticality effects are found for both plural and singular subjects, although the effect is larger for singular subjects. The difference between singular and plural subjects increases when the distance between the subject and the verb is increased: practically, this means that the ungrammaticality effect gets smaller for plural (but not singular) subjects when the subject and verb are separated by intervening material (Wagers et al., 2009).

This number imbalance in comprehension – where an error appears more difficult to process when the subject is singular and the verb is plural rather than in the reverse configuration – is reminiscent of the attraction effects found in production: attractors ('cabinets') cause more speech errors when the head noun ('key') is singular and the attractor is plural, than the other way around. And indeed, attraction effects as well as the disbalance between singular and plural have been found in comprehension as well. Nicol et al. (1997) presented participants with sentences of the types in examples (2) and (3) and asked them to perform several reading tasks (a maze task and a series of sentence classification tasks). The study revealed that participants read the sentences with a singular head noun and plural attractors (mismatch condition) more slowly than sentences with a singular head noun and a singular attractor (match condition), showing the presence of agreement attraction in comprehension. Like in pro-

duction, they found no effect of attractor match when the subject was plural. In other words, during comprehension, participants show the same asymmetrical pattern as during production.

Importantly, the presence of an attractor has the capability to reduce the ungrammaticality effect (Wagers et al., 2009): if there is a number mismatch between the (singular) subject ('key') and the attractor ('cabinets'), the difference between a correct (singular) and incorrect (plural) number on the verb is smaller than when there is no number mismatch between the subject and the verb. In other words, the presence of attractors has the capacity to render a sentence "less ungrammatical" for the reader. This effect has been shown to be visible on the P600 as well, with reduced P600 amplitudes in agreement attraction contexts (2c-d) relative to sentences were attraction effects are absent (2a-b) (Tanner et al., 2014). While Wagers et al. (2009) did not report attraction effects in grammatical sentences, some studies suggest that attraction effects exist in grammatical sentences too, although they are smaller than those in ungrammatical sentences (Dillon, Mishler, Sloggett, & Phillips, 2013; Villata, Tabor, & Franck, 2018).

These findings are usually interpreted in a cue-based retrieval model (Lewis et al., 2006; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006; Vasishth, 2001). The main thesis of cue-based retrieval is that the comprehender uses retrieval cues to retrieve preceding words (often, the subject) from memory in order to integrate the word into the sentence. Cues for retrieval are abstract syntactic properties, such as [+subject] and [+singular] for a singular noun like 'key' in (2a), and [-subject], [+singular] for 'cabinet' in that same sentence. These cues are signaled by the verb. In other words, a verb with [subject] and [singular] features requires a noun that has [+subject] and [+singular] to establish agreement.

The retrieval process is hypothesized to be 'direct access' rather than a serial search through memory because retrieval time has been found to be constant irrespective of the distance from the target (in: Martin, 2018). Instead, the memory representations that are to be retrieved are thought to be *content addressable*, which means that memory representations are organized and retrieved on the basis of their content directly, rather than on the basis of their linear distance from the current input (Martin & McElree, 2008, 2009, 2011). Speakers use different features as cues to retrieve the appropriate representation from memory, including syntactic, semantic, and morphological features. In the case of subject-verb number agreement, these features on the verb are used as cues to retrieve

the subject from memory and establish the dependency. Within this framework, agreement attraction effects arise because some features of the verb correspond to several preceding items in the sentence (the subject, but also the attractor), leading to multiple representations to be considered in parallel and as such interfering with retrieval. Similarly, the ungrammaticality effect is hypothesized to arise because the syntactic position of the subject noun will match the syntactic cues of the verb, but the number cues mismatch, creating interference during the retrieval process (Lago et al., 2021).

Another well-known framework is the *feature percolation model* (Bock & Eberhard, 1993; Eberhard, 1997), briefly alluded to above, which was extended from production to comprehension by Nicol and colleagues (1997). In this framework, ungrammaticality effects must arise through the "backward checking" of number features: given the features on the verb, the subject NP is checked for a match. The agreement attraction effects are modeled to arise because the number feature on the attractor percolates upward and changes the number representation on the subject NP. In other words, instead of the head noun, the attractor noun – both part of the same subject NP – now determines the number of the subject. This leads to an error when checking the features. The asymmetry between singular and plural is thought to arise from a *markedness effect*: only the marked plural feature percolates upwards and affect the representation of the head noun; the unmarked singular feature does not. Notice that the feature percolation model differs radically from the cue-based retrieval model: in the feature percolation model, the memory representation of the head noun is effectively altered, while it is not in the cue-based retrieval account.

Despite this large difference between the two accounts, neither of these predicts the difference between ungrammatical and grammatical attraction effects in comprehension: there is stronger evidence for attraction effects in ungrammatical sentences than in grammatical sentences (e.g., Wagers et al., 2009). The feature percolation account predicts attraction effects should appear similarly in ungrammatical and grammatical sentences. The cue-based retrieval account, on the other hand, suggests that attraction affects may have the opposite directionality in grammatical sentences: if the head noun and the attractor both share their number feature with the verb, the retrieval of the head noun may be impeded due to stronger competition from the attractor.

A good explanatory model appears to contain aspects of both of these models. Yadav and colleagues (2023) performed a computational modeling study in which they compared several implementations of the two accounts, includ-

ing hybrid models containing aspects of both accounts. Comparisons between model fit to the data of 17 studies looking into attraction effects in grammatical sentences during comprehension suggested the data was best described by a hybrid model, which contained feature percolation and grammar-driven cue-based retrieval. In this model, percolation of the number feature on the attractor sometimes affects the representation of the head noun before retrieval occurs. If this happens, retrieval is impeded because the number feature on the head noun no longer matches the number cue on the verb (Yadav et al., 2023).

In the present study, we built upon this literature and investigated whether the subject-verb number agreement processing in grammatical- and ungrammatical conditions can be affected – not only by intervening attractors, but also – by a probabilistic contextual variable: lexical surprisal.

4.1.2 Lexical probability in comprehension

In recent years, influence of probabilistic contextual information on the comprehension of incoming linguistic material has (re)gained importance in linguistic experiments and analyses. The general idea is that the probability of the input given the immediate linguistic context plays an important role in shaping the response to the current input, both in reading and in spoken and signed language comprehension (Hale, 2001, 2006, 2016; Levy, 2008a, 2008b; Levy & Gibson, 2013). Much of the evidence has come from studying the effect of the probability of a word in context on reaction time measures such as self-paced reading (Aurnhammer & Frank, 2019; Kapteijns & Hintz, 2021; Lowder et al., 2018; Monsalve et al., 2012).

The probability of a word in context is often quantified using *surprisal* or *entropy* (as proposed by Shannon (1948)). A word's surprisal (in bits) is the base two log-probability of a word given any number of preceding words and as such quantifies the extent to which a word was expected (surprisal is low) or unexpected (surprisal is high). Entropy, on the other hand, is a quantification of the uncertainty about the current word: it is the weighted average surprisal of all possible continuations of the sentence given any number of preceding words. If entropy is high, the uncertainty about the continuation is high – either because there are many alternatives, or because all alternatives are equally likely and it is difficult to pick one. If entropy is low, the uncertainty about the word is low – either because there are few alternatives, or because one of the alternatives is much more likely than the others.

Both surprisal and (derivatives of) entropy have been found to predict reading time measures. For example, higher surprisal is associated with longer reading times in self-paced reading and eye-tracking studies (Aurnhammer & Frank, 2019; Brothers & Kuperberg, 2021; Frank & Bod, 2011; Lowder et al., 2018; Luke & Christianson, 2016; Monsalve et al., 2012; Sharpe, Reddigari, Pylkkänen, & Marantz, 2018), meaning that words that are less expected given the recent context are read more slowly. Similarly, words that decrease the uncertainty about the rest of the sentence – decrease of entropy – are also associated with longer reading times (Frank, 2013). All of these studies suggest that the probability of a word given the preceding words plays a role in the processing of that word.

More relevant for the present study, Ryu and Lewis (2021) suggested that surprisal from GPT2, a large language model based on Transformer architecture, is useful for modeling psycholinguistic effects. They extracted surprisal values for the stimuli from Wagers and colleagues (2009) and assessed whether the patterns in the surprisal values matched the expected direction of the effects; i.e., higher surprisal values for conditions in which higher reading times are reported in the literature. Indeed, the authors showed that surprisal values from GPT2 successfully simulate the presence of attraction effects. Surprisal values were higher for ungrammatical target verbs than for grammatical ones; and interference of attractors also drove up surprisal values.

In this light, it is unsurprising that surprisal and/or entropy are powerful predictors of human reading time data. Sometimes they are so powerful that it is difficult to find separable effects of syntactic computations and effects of (lexical) probability (e.g., Kapteijns & Hintz, 2021). This has reignited the old discussion surrounding the necessity of abstract, hierarchical syntactic structure for a psycholinguistic theory of language comprehension (Frank & Bod, 2011; Frank et al., 2012; Frank & Yang, 2018; Pulvermüller & Assadollahi, 2007).

That lexical probability is a good predictor of human reaction time data fits well with a framework called *surprisal theory* (Hale, 2001, 2006, 2016; Levy, 2008a, 2008b; Levy & Gibson, 2013). This theory has focused on modeling a variety of effects observed in psycholinguistic studies of sentence comprehension (e.g., syntactic ambiguity resolution) in a probabilistic framework. The main aim of the framework is to predict where in a sentence the comprehender will encounter processing difficulty. The theory is that listeners, signers and readers make use of probabilistic knowledge to predict (1) the structure of the input they have just encountered, and (2) what they may encounter next. The deviation

from this prediction, as quantified by surprisal, should be the best predictor of reading times. While surprisal theory is a successful descriptive theory, it lacks a mechanistic component: what representations are used for surprisal estimation (syntactic, lexical, or even syllabic? All of the above?), and how does the human mind calculate and represent these values?

subsectionLanguage comprehension as (probabilistic) cue integration

In light of the evidence for the importance of probabilistic information during language comprehension, several studies have examined if and how lexical probabilistic information interacts with a variety of grammatical processes, including grammatical agreement. Indeed, there is evidence that agreement is affected by lexical probability. In a self-paced reading experiment, Brehm, Hussey & Christianson (2020) examined the role of word frequency of the attractor on attraction effects. The study revealed that word frequency indeed affects attraction: high-frequency nouns did not elicit attraction effects, while low-frequency nouns did. In other words, for low-frequency nouns, the reading times on the verb differed between singular and plural, but they did not do so for high-frequency nouns. This suggests that lexical probability (at least of the attractor) plays a role in the establishment of morphosyntactic relations.

Furthermore, Loerts and colleagues (2013) studied the interaction between a different kind of agreement – gender agreement – and lexical probability in an ERP study. The authors presented participants with Dutch sentences in which cloze-probability was manipulated, such that the context was either highly constraining (i.e., the target word was predictable from the context) or not constraining (i.e., the target word could not be predicted from the context). The target word was a noun that was preceded by a determiner or an adjective, a construction which in Dutch requires gender agreement. This agreement relation was manipulated to be correct or incorrect, to elicit a P600 (i.e., an ungrammaticality effect). The study revealed that the P600 appeared earlier in conditions where the word was predictable from the context. This again suggests that lexical predictability interacts with morphosyntactic processing.

Departing from the cue-based retrieval framework, Campanelli and colleagues (2018) studied the interaction between memory retrieval and contextual predictability in a study with object relative clauses. The authors manipulated the predictability of the target verb that agreed with the target subject of the object relative clause (high/low predictability condition). At the same time, participants were presented with a list of three nouns that they did or did not need to remember during the presentation of the stimulus sentence (load/no load

condition). In some cases, the list contained a noun that would be a plausible candidate for the subject of the target verb (interference/no interference). The study revealed that the interference effect induced by the memorization of a list of nouns was cancelled out by a highly predictable verb (i.e., the interference effect was only in the low predictability condition). The authors suggest that accumulated evidence may selectively pre-activate the subject noun in memory and make it more available for retrieval compared to the distractors, consequently minimizing interference. This is in line with the hypothesis that the (distributional) relevance of an element given the context determines the level of pre-activation of this element (Campanelli et al., 2018).

Similarly, Tung and Brennan (2023) used EEG to show that lexical predictability affects the use of linguistic cues during ellipsis resolution in Mandarin Chinese. In contrast to Campanelli and colleagues, who manipulated the predictability of the verb, and Brehm and colleagues, who manipulated the frequency of the attractor, this study manipulated the predictability of the target by changing the verb. In addition, the authors manipulated the grammaticality of the target, and the semantic relatedness of the distractor to invoke interference effects. They investigated how the semantic relatedness of the distractor and the predictability of the target affect the difference between grammatical and ungrammatical ellipsis resolution (reminiscent of Loerts et al. (2013)). They showed that the P600-effect as a result of the grammaticality manipulation was affected by both the semantic relatedness of the distractor and the predictability of the target: when the target was predictable, the P600-effect was reduced, and more so when interference from the distractor was high. These findings are in line with those from Campanelli and colleagues (2018).

Given the importance of the findings from the agreement literature and the strong effects of probability, and taking into account the observed interactions, Martin (2016; 2018; 2020) proposed a model that integrates the architecture of the cue-based retrieval model with probabilistic processing. This model, entitled language processing as cue integration (LPCI), suggests that language comprehension is a process of cue-based inference. The cues can come from the input signal (be it text, sign, or speech) and from linguistic representations stored in memory (e.g. grammatical knowledge, lexical knowledge). In the model, the linguistic representations are hypothesized to form a hierarchy of levels, with levels ranging from phonetic features to phrases and larger sentential or event structures. Each level of representation is a cue to higher levels of representation, creating a cascaded architecture in which activation can spread to higher

levels before inference of a given level is complete. For example, syllabic representations are cues to lexical representations, lexical representations cue phrasal representations, and so forth. The representation that results from this cascade of cue-based inference is an abstract, hierarchical representation of the inferred meaning.

The relationship between a given cue and the "target" (i.e., the higher level of representation), and by consequence, the representations themselves, are hypothesized to be probabilistic in nature; some cues have a higher reliability than others. This has the advantage that the model can account for inference at higher levels of representation both when processing at a given stage is not complete (e.g., lexical decision before the end of a word; Grosjean, 1980), or cannot be completed (e.g., a phoneme is hidden by a cough, and yet the word is perceived; Warren, 1970). Similarly, through this mechanism the probability of a lexical item may affect the formation of relations between lexical items, such as subject-verb agreement.

4.1.3 The present study

In the present study, we used self-paced reading to examine whether the probability of the target word given the context (i.e., the lexical surprisal) affects the establishment of subject-verb agreement. We presented participants with sentences in which the target word for the establishment of the agreement relation (in this case, the subject noun) either did or did not agree with the preceding verb in number to elicit the ungrammaticality effect. In addition, the target noun was manipulated such that it had low- or high surprisal given the sentence context (a preposition phrase, the verb, and a definite determiner). An example stimulus is shown in table 4.1 below.

These manipulations allow us to ask whether the lexical probability of the target noun affects the processing of subject-verb agreement. The models outlined above (cue-based retrieval, surprisal theory, and LPCI) make different predictions when it comes to such an experimental paradigm.

Cue-based retrieval The current theory of cue-based retrieval predicts a slowdown for the ungrammatical relative to the grammatical sentence: the mismatch between the number cue of the noun and the preceding verb interfere with retrieval, causing a slowdown. The model does – to our knowledge – not readily include an architecture to make predictions about whether the surprisal of the target will interfere with this process.

Table 4.1: Example stimulus set.

Cat. surprisal	Agr.	Context	Target	Spill-over region	Surprisal	WF*
Low	Correct	In de herfst controleert de *In the fall checks the* "*In the fall, the steward checks the trees for illnesses*"	beheerder *steward*	de bomen op ziektes *the trees for illnesses*	11,2	2,00
Low	Incorrect	In de herfst controleert de *In the fall checks the* *"*In the fall, the stewards checks the trees for illnesses*"	beheerders **stewards*	de bomen op ziektes *the trees for illnesses*	13,4	1,26
High	Correct	In de herfst controleert de *In the fall checks the* "*In the fall, the blonde woman checks the trees for illnesses*"	blondine *blonde woman*	de bomen op ziektes *the trees for illnesses*	19,16	2,00
High	Incorrect	In de herfst controleert de *In the fall checks the* *"*In the fall, the blonde women checks the trees for illnesses*"	blondines *blonde women*	de bomen op ziektes *the trees for illnesses*	22,21	1,40

Note. Log10 word frequency from the SUBTLEX-NL corpus.

(A) Cue-based retrieval hypothesis

- $RT_{incorr} > RT_{corr}$

Surprisal theory While surprisal theory aims to explain when a processing difficulty will arise, the lack of a specific mechanistic 'back-end', so to speak, makes it relatively difficult to derive specific predictions. Let us go over the problem. The simplest prediction that surprisal theory could make is that reading times are fully proportional to surprisal. Since lexical surprisal captures lexical probability, and ungrammatical sequences of words have low probability, this metric of surprisal should capture the ungrammaticality that arises from an agreement manipulation (and they do; see below). However, it is possible that surprisal values are extracted from some representation of the syntactic parse (e.g., Levy, 2008a). In this case it depends on the instantiation of agreement in this parser whether or not there is a difference of surprisal for the grammatical and ungrammatical sequences. In other words, what the surprisal value of the target noun is (and, therefore, how well it predicts our reading time data), depends wholly on the representation of the stimulus the surprisal value is calculated over.

For simplicity, we assume a strict version of surprisal theory, most reminiscent of what is advocated by Frank and colleagues (Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018). This framework proposes a lexicalist view on sentence processing. Under this assumption, *lexical* surprisal should fully predict reading times. Given that surprisal captures ungrammaticality as explained above, surprisal values should capture this slowdown, as well; a prediction that

is directly in line with the findings from Ryu and Lewis (2021). In other words, the ungrammaticality effect does not need to be modelled separately.

(B) Surprisal theory hypothesis (strong)

- RT $\sim$ surprisal
- $\text{RT}_{\text{incorr\_low}} = \text{RT}_{\text{corr\_high}}$

LPCI As outlined above, the LPCI model combines aspects from cue-based retrieval with a probabilistic view on language processing. The model predicts that lexical distributional information affects processing of the target noun, with a facilitative effect if a word is expected in the context (low surprisal). In addition, according to this model, the comprehender uses linguistic cues from memory to process the input. In line with the cue-based retrieval model, therefore, a grammaticality effect is expected, with longer reading times for the ungrammatical relative to the grammatical condition. Besides these two predictions, however, the model makes another prediction: lexical probability and agreement may have contradicting forces, leading to an interaction effect. A lexical probabilistic cue may activate the target noun and thus ease its processing, but the number feature on the noun is not congruent with the number feature on the verb, as such creating a processing difficulty. The result of these two forces working simultaneously could result in a *smaller* ungrammaticality effect for low-surprisal words than for high-surprisal words. Importantly, however, the lexical and phrase-level representations generated during reading are separable in this model. This predicts that the incorrect and correct agreement conditions do not have the same reading times, even when the lexical surprisal values are identical.

(C) LPCI hypotheses

- $\text{RT}_{\text{incorr}} > \text{RT}_{\text{corr}}$
- RT $\sim$ surprisal
- $\text{RT}_{\text{incorr\_high}} - \text{RT}_{\text{corr\_high}} > \text{RT}_{\text{incorr\_low}} - \text{RT}_{\text{corr\_low}}$
- $\text{RT}_{\text{incorr\_low}} \neq \text{RT}_{\text{corr\_high}}$

4.2 Methods

4.2.1 Participants

88 Dutch native speakers were recruited form the participant pool of the Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. All participants had normal or corrected-to-normal hearing and reported no neurological disorders, dyslexia, SLI or other language disorders. Participants provided informed consent and received monetary compensation for their participation (€6,-). The study was approved by the ethics board of the Faculty of Social Sciences of Radboud University under the umbrella approval assigned to prof. dr. Antje S. Meyer (ECSW2014-1003-196).

4.2.2 Materials

We used a 2 x 2 within-participant design with the factors 'surprisal' with two levels (high surprisal, low surprisal) and 'agreement' with two levels (correct subject-verb number agreement, incorrect subject-verb number agreement). The stimuli consisted of 40 quadruplets of sentences. Across the four versions of a sentence the target word was manipulated for agreement and surprisal; i.e., a given sentence frame (the pre-target words and spill-over region) was used to create an item for every condition ('low surprisal/correct', 'low surprisal/incorrect', 'high surprisal/correct', 'high surprisal/incorrect'. This ensured that the same sentence frame occurred in once each condition. The two manipulations were shown in table 4.1 above. All sentences followed an initial P – det – N – V – det – N structure, with the final noun (the sixth word) being the subject of the sentence and the target word. Adding an adjunct to the beginning of the sentence impedes the subject to move to its preferred position in first position. This is because Dutch is a *verb-second* language: in a declarative root clause, the verb will end up in second position. As a consequence, the verb can precede the subject when the first position is filled. This leads to a V-S-O order. In such sentences, a reader or interlocutor who encounters a plural verb form can expect a plural NP as a grammatical subject. Agreement was manipulated between verb and the subject, which always directly followed each other (a determiner intervenes between the subject noun). The agreement error was plural in half of the items and singular in the other half.

The structure described above – specifically, the preverbal preposition phrase – was chosen to ensure enough context to drive surprisal values on the tar-

get. Since surprisal depends in part on the frequency of the word itself and (therefore) tends to correlate with word frequency (Slaats et al., 2023, see also Chapters 2 and 3 in this dissertation), having enough context allows for the selection of words that have different surprisal values but relatively similar word frequency values. This was done in the following way.

A potential low surprisal context + target combination (both correct and incorrect) was submitted to a script that extracted the surprisal value of the target word from a version of GPT2 that was retrained for Dutch (de Vries & Nissim, 2021). If the surprisal value for the correct target (in table 4.1: 'beheerder') was below 12, the word frequency value was used to extract from the SUBTLEX-NL corpus (Keuleers et al., 2010) the 100 words that were nearest to the target in their word frequency value. These were then combined with the potential context and their surprisal values were derived from the Dutch version of GPT2. The words that had a surprisal value of above 14 were returned by the script for human evaluation. A word that led to a context-target combination that was grammatical and plausible was chosen for the stimulus, as long as the incorrect number of this word had a similar word frequency as the low surprisal incorrect number alternative. In continuation, an end to the sentence was added to match the context and both targets as well as possible. Since agreement effects tend to be found up until the third spill-over word (Wagers et al., 2009), the end of the sentences consisted of four words: three for the spill-over, and one for the wrap-up effect (longer reading times on the last word of a sentence; Just & Carpenter, 1980).

In general, the aim was to keep the surprisal values of incorrect/low surprisal stimuli below 14 bits. This could not always be ensured, however: in 8 sets, the incorrect/low surprisal had a surprisal value that was higher than 14 bits. In all cases except one, however, the correct/high surprisal values were still higher than the incorrect/low surprisal values. In the one exception, the incorrect/low surprisal value was 14.47 bits, while the correct/high surprisal value was 14.08 bits.

The correct sentences were judged for plausibility on a 7-point Likert scale by 5 native speakers of Dutch, who also provided informed consent. Sets in which one or more sentences were judged to be fairly implausible (with a mean below 4.0) were removed; this led to a selection of 40 sets. The stimuli were divided over four lists, such that each list contained one item from every set. In this way, participants were never presented with the same sentence frame or target word more than once. There were no surprisal or word frequency differences between

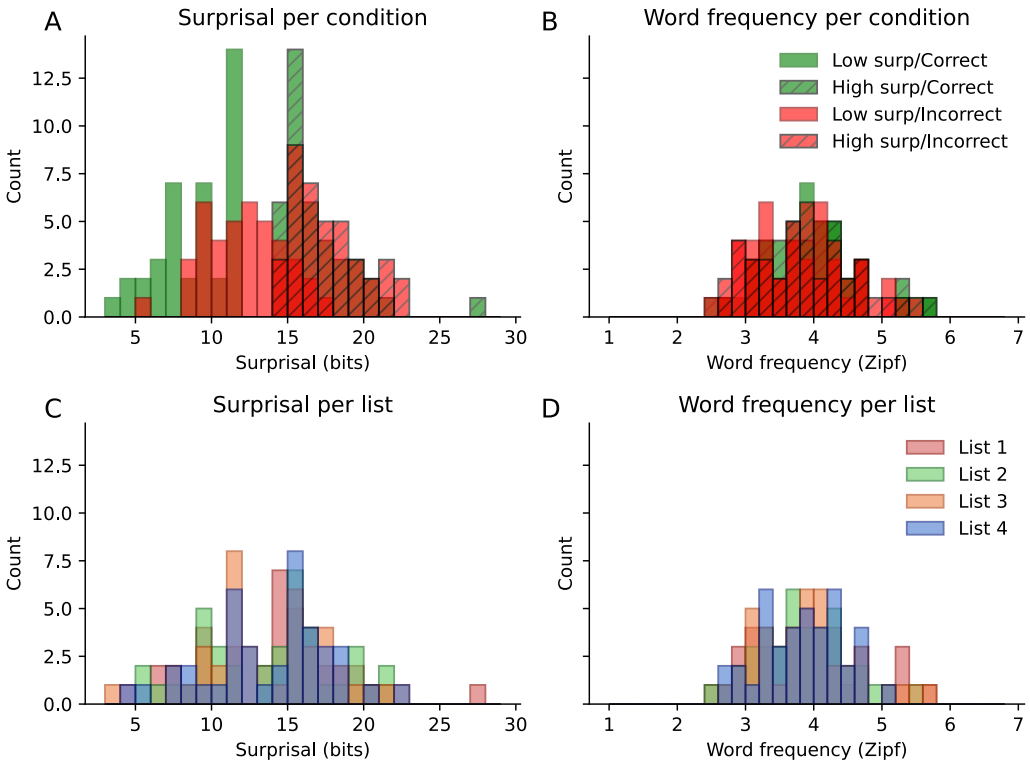


Figure 4.1: The distribution of surprisal and word frequency values on the target word. (A) Surprisal values for the target word per condition. (B) Word frequency values for the target word per condition. (C) Surprisal values for the target word per stimulus list. (D) Word frequency values for the target word per stimulus list.

lists ($F_{surprisal}(3, 36) = 0.099$, $p_{surprisal} = 0.96$; $F_{WF}(3, 36) = 0.092$, $p_{WF} = 0.96$; see also Figure 4.1C and D).

A two-way ANOVA on the surprisal values with factors *surprisal* ('low', 'high') and *agreement* ('correct', 'incorrect') performed with Statsmodels in Python 3.7 (Seabold & Perktold, 2010) revealed that there was an interaction between the *agreement* and *surprisal* factors ($F(1,1) = 6.57$, $p = 0.01$) on the surprisal value, in addition to main effects of agreement ($F = 27.01$, $p < 0.01$) and surprisal category ($F = 275.60$, $p < 0.01$). This suggests that surprisal values are higher in the high surprisal category than the low surprisal category – as intended –, that incorrect agreement leads to higher surprisal values – which is inevitable, since language models are trained on correct agreement (almost) exclusively –, and that the effect of agreement on surprisal values is stronger in the low surprisal category than the high surprisal category. This can be seen in Figure 4.1A. We will get back to this in the analysis. The two-way ANOVA on the word frequency values revealed no effects (all $p > 0.1$).

The filler sentences were 90 grammatical sentences taken from Creemers and Meyer (2022) and 30 grammatical sentences from a large-scale MEG/fMRI study (Schoffelen et al., 2019). 60 sentences from Creemers and Meyer (2022) made reference to two characters; 20 of those included a pronoun that was ambiguous with respect to its referent. The remaining 30 sentences from this set contained only one character. The 30 sentences from the MEG/fMRI study all contained a subordinate clause. Otherwise, the structures were varied. The filler sentences were the same in the four lists. With 120 grammatical fillers, 20 grammatical test items and 20 ungrammatical test items the study had a ratio of 14,3% ungrammatical sentences, which is similar to other agreement studies with self-paced reading (e.g., Wagers et al., 2009). In total, the experiment consisted of 1793 words, which is an average of 11.21 words per sentence. All experimental sentences contained 10 words. All sentences are shown in appendix I, 4.6.

4.2.3 Procedure

The experiment was performed online using Frinex (Framework for Interactive Experimentation), a tool for web experiments developed at the Max Planck Institute for Psycholinguistics. Participants were sat behind their own computer and read the sentences in a word-by-word self-paced reading paradigm.

Upon pressing the control key, the first word of the sentence appeared. The other words appeared masked; instead of the letters, a corresponding number of dashes was shown on the screen. The space bar was used to proceed to the next word. The previous word was masked again. Participants read 6 sentences in a practice phase. These sentences were all followed by a comprehension question addressing the content of the sentence, in which participants were forced to choose between two options by a button press. In the test phase, 20% of the sentences were followed by a comprehension question. The questions were evenly divided over trial types. The position of the correct answer was balanced. 12 experimental sentences were followed by a question. These questions never addressed the target word, and the references to the context and the spill-over region were balanced.

Each participant read one out of four lists. Due to the nature of the Frinex software, the order of these lists was fixed (this is accounted for in the analysis; see 4.2.4below). The experimental trials were pseudo-randomly interspersed across the whole list of 160 items; care was taken there were always two or more filler items in between two experimental trials.

4.2.4 Analysis

We recorded reading times (RTs) for every word, and responses to the comprehension questions. We then calculated the participants' accuracy on the comprehension questions. All participants had accuracy scores of above 75% (mean: 93.3%; standard deviation: 3.9%). The data of all 88 participants was included in the analysis. We analyzed the RT-data word by word, in which we separately inspected effects of surprisal and agreement at the target word and the three spill-over words. The RT-data was analyzed with linear mixed-effects models using the *lme4* package (Bates et al., 2015) in R (version 4.3.1). P-values were computed using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

Preprocessing Unreasonably short and long latencies (<100 ms and >2500 ms) were excluded prior to analysis, which resulted in a loss of 0.5% of the data. Response times were log-transformed (natural log) in order to reduce the positive skew that is inherent to reading times (Baayen & Milin, 2010). Log response times more than 2.5 standard deviations from the condition mean (per participant) were excluded. The responses to one sentence were not recorded correctly for one participant. In total, there was a loss of 2.77% of the data.

The log-transformed RTs were regressed against several factors that are often found to affect RT-data in self-paced reading experiments: word frequency (higher word frequency predicts faster RTs; Schilling, Rayner, & Chumbley, 1998), word length (in number of letters; longer words predict slower RTs; Barton et al., 2014), and the position of the sentence in the stimulus list (sentences presented later in the list predict faster RTs; Hofmeister, 2011). The word frequency estimates were the Zipf values from the SUBTLEX-NL database (Keuleers et al., 2010). The estimates of word length and list position were log-transformed with a natural logarithm and centered; the word frequency factor was centered. All words from the test sentences (i.e., not only the target word and the spill-over region) were used for these models.

Since word length and word frequency have been found to interact such that the effect of frequency is larger for longer words (Barton et al., 2014; Johnson & Rayner, 2007; McGinnies, Comer, & Lacey, 1952; Postman & Adis-Castro, 1957), we performed model comparison on this regression, as well, starting out with a model in which word length and word frequency interact, and comparing it to a model that did not include this interaction. The comparison revealed that an interaction between word frequency and list position and a main effect of

sentence order in the fixed effects, with random slopes for word length and word frequency, provided the best fit to the data ($\chi^2(1) = 228.21$; $p < 0.01$). This model showed a main effect of word frequency, with faster RTs for words of higher frequency (β = -6.33·10^{-3}, SE = 3.04·10^{-3}, t(33790) = -2.08, p = 0.037), a main effect of list position, with faster RTs for words at later positions (β = -1.06 ·10^{-1}, SE = 5.77·10^{-3}, t(88.04) = -18.32, $p < 0.01$), and, in line with previous findings, an interaction between word length and word frequency (β = 2.88·10^{-2}, SE = 1.90·10^{-3}, t(33980) = 15.13, $p < 0.01$). The per-word residuals of this model for the target word and the spill-over region were submitted to further analysis.

Statistical analysis Because of the interaction between the surprisal category ('low', 'high') and agreement ('correct', 'incorrect') on the surprisal values of the target word mentioned in section 4.2.2 above, the residual RTs were analyzed for effects of agreement and surprisal in two ways: once with a categorical factor of surprisal ('low', 'high'), and once with a continuous factor of surprisal (numerical). By including surprisal as a numerical factor we model the differences in surprisal value between the agreement-groups. For this reason, we describe the analysis with surprisal as a continuous factor here, and include the results for the categorical surprisal factor in appendix II, 4.7.

For each word and surprisal factor, we used a maximal-to-minimal model comparison approach. The maximal model included surprisal, agreement and the interaction. In addition, the model included a number factor ('singular', 'plural') which coded the correct inflection of the noun given the verb, in interaction with agreement to account for any potential differences between plural and singular correct nouns on the ungrammaticality effect. Departing from this model we step-wise removed factors. We first reduced the random effects structure following the 'keep it maximal' principle. We fitted every model with the maximal random effects structure possible given the paradigm and model (by-participant intercept, by-participant slopes for agreement, surprisal, number, agreement * number, and agreement * surprisal), and all possible combinations of these. We adopted the largest random effects structure (i.e., the highest number of parameters) that converged and was non-singular (Barr, Levy, Scheepers, & Tily, 2013). If two models with the same number of random factors both converged, we used the Akaike Information Criterion (AIC) to determine which model best fitted the data. Having chosen the random effects structure for the maximal fixed effects structure, we reduced the fixed effects structure by comparing each model to a

version with fewer fixed factors using a chi-square test. In this case, if there was no difference between two models, the factor was removed.

In case of an interaction effect, estimated marginal means were extracted using the emmeans package. Simple-effects analyses were performed using the *joint_tests* function. P-values were Bonferroni corrected for multiple comparisons.

4.3 Results

Table 4.2 and Figure 4.2 below provide an overview of the means and standard deviations of the original reading times (after outlier rejection). As described above, all analyses were performed on the residual log-transformed reading times. These were the result of a regression of the log-transformed reading times against factors *list position*, *word length,* and *word frequency*, as well as an interaction between the latter two. Any influence of these factors is removed. The residuals are displayed in Figure 4.3.

We observed the following patterns. (1) There was an effect of agreement in all analyzed regions (target, spill-over 1, spill-over 2, and spill-over 3), with longer reading times in sentences with incorrect number agreement than sentences with correct number agreement. At the target, the effect of agreement was somewhat obscured by an effect of number: the agreement effect reversed direction when the verb is plural (i.e., incorrect singular nouns were read faster than the correct plural nouns). (2) Surprisal had a weaker effect than agreement. We observed that higher surprisal values were associated with longer reading times at the target, spill-over 1 and spill-over 2. At spill-over 1, surprisal appeared to affect reading times only in grammatical conditions; there were no other interactions between surprisal and agreement. (3) There was also an effect of verb number. This effect was visible at the target, spill-over 1 and spill-over 2, with longer reading times for words following plural verbs than words following singular verbs. As mentioned in (1), the effect of number partially overrode the effect of agreement at the target. At spill-over 1, the effect of number affects reading times in correct- but not incorrect sentences – reminiscent of the effect of surprisal. The full results and model comparison statistics are described below.

Table 4.2: Means (and standard deviations) for reading times in the four analyzed word positions.

Cat. surprisal	Agr.	Target		Spill-over 1		Spill-over 2		Spill-over 3	
Low	Correct	332.24	(89.10)	342.84	(82.48)	338.72	(79.82)	349.32	(81.43)
	Incorrect	347.21	(87.02)	398.83	(126.41)	386.10	(109.57)	383.66	(95.22)
High	Correct	334.28	(96.49)	346.30	(82.44)	348.11	(83.33)	352.79	(87.43)
	Incorrect	332.03	(98.93)	377.29	(136.29)	370.54	(102.08)	363.27	(95.06)

Note. Target: word position 5; spill-over 1-3: word positions 6-8.

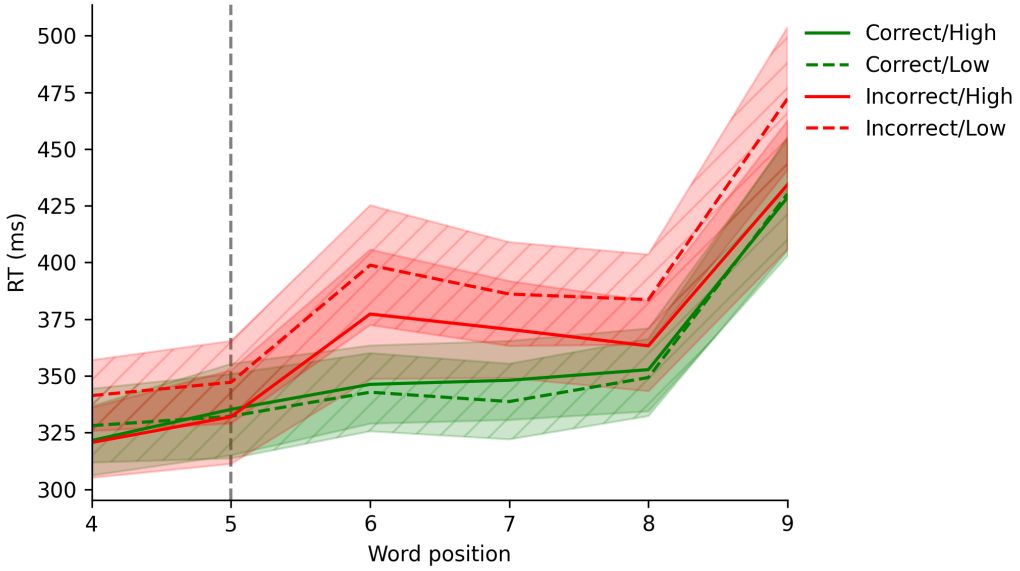


Figure 4.2: (Unanalyzed) raw reading times in the analyzed window (word positions 5, 6, 7, and 8) in two agreement conditions (correct, incorrect) and two surprisal conditions (high, low). Shaded areas represent 95% confidence intervals. Grey dashed line indicates the target word. Words in positions 4 and 9 were not analyzed (4 is pre-target, 9 shows the increased RTs associated with the wrap-up effect).

4.3.1 Target word

Reducing the random effects structure revealed that the largest random effects structure that lead to non-singular fit included random slopes for agreement and a random intercept. Model comparison showed that the interaction between surprisal and agreement did not contribute to model fit ($\chi^2(1) = 0.05$; $p = 0.82$). The main effect of surprisal did, however ($\chi^2(1) = 11.55$; $p < 0.01$), as did the interaction between agreement and number ($\chi^2(1) = 21.27$; $p < 0.01$). The interpreted model included a fixed effect of surprisal, the interaction between agreement and number, and it had by-participant random slopes for agreement and a random intercept. This model, summarized in table 4.3 below,

revealed a main effect of surprisal, indicating that higher surprisal values lead to longer reading times at the target word. In addition, there was a main effect of agreement, showing the ungrammaticality effect: incorrect nouns had longer reading times than correct nouns. Unsurprisingly, these models also showed an interaction between number and agreement.

Table 4.3: The output of the interpreted linear mixed effects model of the residual log reading times at the target word (the subject). Model: residual log RT ~ surprisal + noun_number · agreement + (1 + agreement | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-1.28e^{-01}	1.63e^{-02}	1,01e ^{03}	-7.83	1.19e^{-14}	***
Surprisal	3.22e^{-03}	9.47e^{-04}	3,32e ^{03}	3.40	6.77e^{-04}	***
Agreement	3.72e^{-02}	1.26e^{-02}	3,17e ^{02}	2.95	3.40e^{-03}	**
Correct number	4.77e^{-02}	1.15e^{-02}	3,30e ^{03}	4.16	3.22e^{-05}	***
Agreement * correct number	-7.49e^{-02}	1.62e^{-02}	3,29e ^{03}	-4.62	3.99e^{-06}	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Simple effects comparisons revealed that the effect of agreement was in fact reversed for sentences with a plural preceding verb: the incorrect singular nouns were read faster than the correct plural nouns ($F(1,274.96)=9.71$, $p < 0.01$). In sentences with singular preceding verbs, the ungrammaticality effect had the expected directionality ($F(1)=8.65$, $p < 0.05$). Comparison in the opposite direction (i.e., the effect of number per agreement condition) revealed that the effect of number was significant in grammatical sentences, with plural nouns following a plural verb being read more slowly than singular nouns following singular verbs ($F(1, 3300.94) = 17.31$, $p < 0.01$); in the incorrect condition, there was a trend for singular nouns following plural verbs to be read faster than plural nouns following singular verbs ($F(1, 3300.03) = 5.60$, $p = 0.072$). This can be seen in Figure 4.4D.

In sum, at the target word we observed an effect of surprisal in the expected direction: higher surprisal values were associated with longer reading times. In addition, there was an effect of agreement, with longer reading times for ungrammatical than grammatical targets. This effect was partially overridden by the effect of number, which shows up as plurals receiving longer reading times than singulars. In the case of a plural verb, the effect of number on the target was so strong that the ungrammatical singular noun was read faster than the grammatical plural noun. There was no interaction between agreement and surprisal.

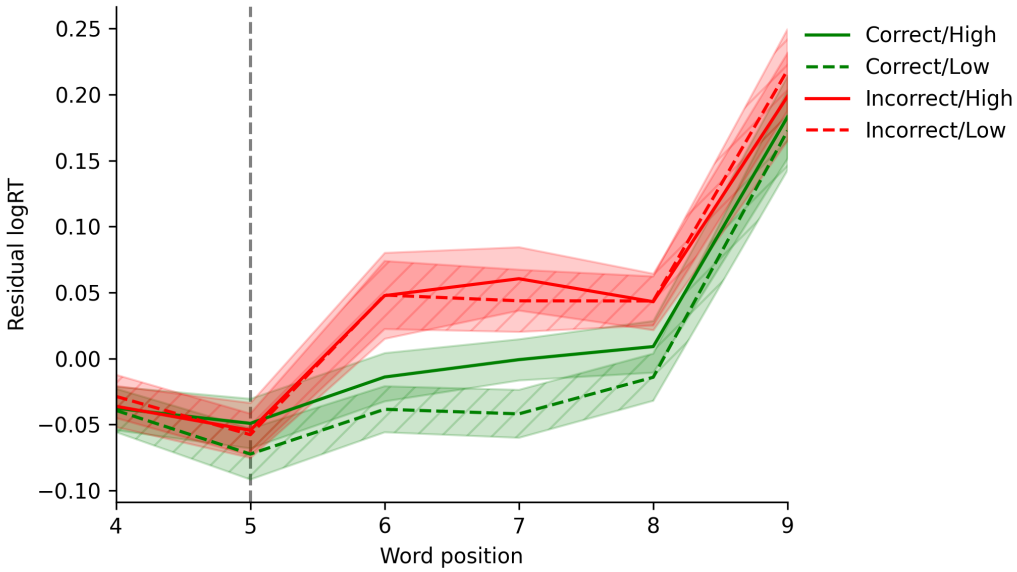


Figure 4.3: Residual log reading times in the analyzed window (word positions 5, 6, 7, and 8) in two agreement conditions (correct, incorrect) and two surprisal conditions (high, low). The residuals are obtained from a model that regressed the original reading times against list position, word length, and word frequency. Shaded areas represent 95% confidence intervals. Grey dashed line indicates the target word. Words in positions 4 and 9 were not analyzed (4 is pre-target, 9 shows the increased RTs associated with the wrap-up effect).

4.3.2 Spill-over 1

Two models converged and did not have singular fit: the models with either random slopes for surprisal, or random slopes for agreement (but not both at the same time). Comparison of the AIC values suggested that random slopes for agreement yielded better fit to the data ($AIC_{agreement} = 538.20$; $AIC_{surprisal} = 593.04$). Model comparison for fixed effects showed a trend for the interaction between agreement and surprisal to contribute to model fit ($\chi^2(1) = 3.41$; $p = 0.065$). Since this value approached significance and is the interaction of interest, we will maintain this interaction and further reduce from there.

Reduction from the model that contains an interaction between surprisal and agreement revealed that the interaction between agreement and number contributed significantly to model fit ($\chi^2(1) = 9.04$; $p < 0.01$). This means that here, we interpret the full model, with interactions between agreement and surprisal, and agreement and number, a random intercept, and random slopes for agreement. This model revealed main effects of agreement, surprisal, and number. The output is shown in table 4.4 below. The agreement effect showed that

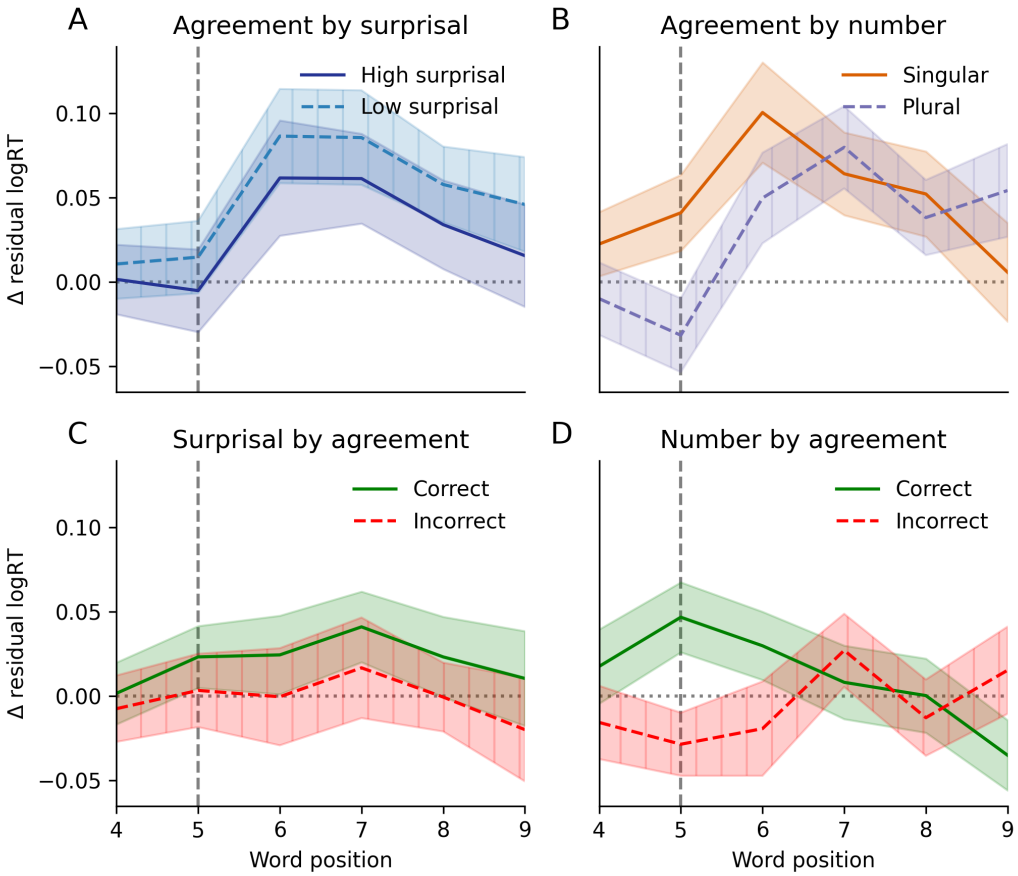


Figure 4.4: Pairwise comparisons for the interactions between agreement and surprisal and agreement and number. Shaded areas represent 95% confidence intervals. The vertical grey dashed line indicates the target word (noun). (A) The ungrammaticality effect for the high- and low surprisal categories. Y-axis corresponds to the difference between incorrect and correct agreement. The zero-point on this axis, traced by a horizontal dotted line, indicates no ungrammaticality effect. (B) The ungrammaticality effect for singular and plural preceding verbs. Y-axis corresponds to the difference between incorrect and correct agreement. The zero-point on this axis, traced by a horizontal dotted line, indicates no ungrammaticality effect. (C) The difference between high surprisal and low surprisal for the correct and incorrect agreement conditions; i.e., the reverse of (A). Y-axis corresponds to the difference between high surprisal and low surprisal. The zero-point on this axis, traced by a horizontal dotted line, indicates no effect of surprisal. (D) The difference between sentences with a plural and singular verb per agreement condition, i.e. the reverse of (B). Y-axis corresponds to the difference between sentences with plural and singular verbs. The zero-point on this axis, traced by a horizontal dotted line, indicates no effect of verb number.

words in ungrammatical sentences received longer reading times. The effect of surprisal showed that higher surprisal values were associated with longer reading times. The effect of number showed that words following plural verbs were associated with longer reading times than words following singular verbs, as well as a significant interaction between agreement and number plural. The interaction between agreement and surprisal approached but did not reach statistical significance.

Simple effects revealed that the effect of agreement existed for both plural and singular verbs (singular: $F(1,210.58) = 38.36$, $p < 0.01$; plural: $F(1,187.92) = 8.84$, $p < 0.05$); see Figure 4.4B. The reverse configuration revealed that the interaction between agreement and number was caused by a significant effect for number in grammatical sentences, with plural nouns following plural verbs being read more slowly than singular nouns following singular verbs ($F(1,3312.86) = 7.37$, $p < 0.05$), but not in ungrammatical sentences ($F(1,3302.67) = 2.38$, $p = 0.49$), where singular and plural nouns incorrectly following plural and singular nouns, respectively, were read at similar speed; see Figure 4.4C. In addition, and more importantly, simple effects of the interaction between agreement and surprisal revealed that surprisal led to longer reading times in grammatical sentences ($F(1,3311.24) = 7.79$, $p < 0.05$) but not in ungrammatical sentences ($F(1,3322.63) = 0.003$, $p = 1$), causing the ungrammaticality effect to be larger for low surprisal than high surprisal words; this can be seen in the second panel of Figure 4.5 below, and, to some extent, in Figure 4.4A (per surprisal category).

In summary, at the first spill-over word we observed three main effects and two interactions. There was a main effect of agreement: words following ungrammatical targets received longer reading times than words in grammatical sentences. In addition, there was a main effect of surprisal: higher surprisal values were associated with longer reading times. Finally, the main effect of number suggested that generally speaking, words following plural verbs received longer reading times than words following singular verbs. The effect of agreement interacted with both surprisal and number. We observed that the effect of number and the effect of surprisal were significant only in grammatical sentences.

4.3.3 Spill-over 2

In the second word following the target word, both random slope for number and a model with only a random intercept converged and were non-singular. However, upon running other fixed effects structures for reduction, the model with a random slope for number led to singular fit. For this reason, we opted for

Table 4.4: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the first spill-over word. Model: residual log RT ~ surprisal + noun_number · agreement + (1 + agreement | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-9.2310^{-2}	2.00e^{-02}	2.52e^{03}	-4.61	4.24e^{-06}	***
Agreement	1.51e^{-01}	3.35e^{-02}	2.17e^{03}	4.51	6.80e^{-06}	***
Surprisal	3.72e^{-03}	1.33e^{-03}	3.31e^{03}	2.79	5.26e^{-03}	**
Correct number	3.33e^{-02}	1.23e^{-02}	3.31e^{03}	2.72	6.63e^{-03}	**
Agreement * surprisal	-3.80e^{-03}	2.06e^{-03}	3.31e^{03}	-1.85	0.06465	.
Agreement * correct number	-5.24e^{-02}	1.74e^{-02}	3.30e^{03}	-3.01	2.65e^{-03}	**

Note. Signif. codes: * p < 0.05; ** p < 0.01; *** p < 0.001

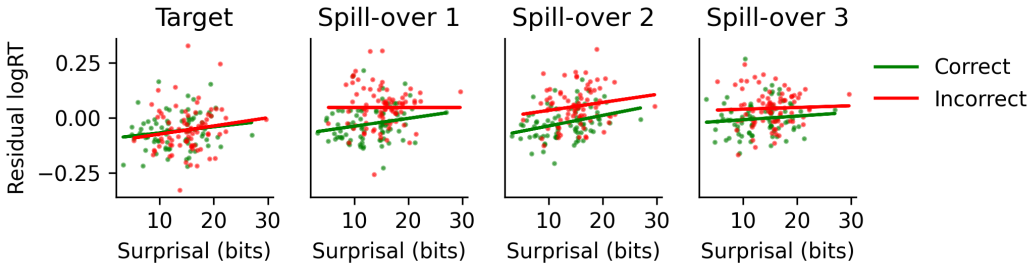


Figure 4.5: Interaction between continuous surprisal and agreement for the different word indices analyzed. The dots in the scatter plot show the residual log reading times per test item averaged over participants. N.B. the slopes shown in this figure were estimated over the averages per item using *numpy.polyfit* for visualization purposes. These estimates closely approach but do not directly correspond to the slopes estimated in the linear mixed effects model, which are reported in the text.

a random intercept only. Model comparison for fixed effects revealed no contribution of the interaction between surprisal and agreement to model fit ($\chi^2(1) = 0.14$; $p = 0.71$), whereas surprisal ($\chi^2(1) = 23.19$; $p < 0.01$) did contribute. The interaction between number and agreement did not improve model fit ($\chi^2(1) = 1.04$; $p = 0.31$). The main effect of number did contribute ($\chi^2(1) = 6.07$; $p < 0.05$), as did agreement ($\chi^2(1) = 58.54$; $p < 0.01$). The interpreted model contained fixed main effects of surprisal, agreement, and number, and a random intercept. The output is shown in table 4.5 below. There were main effects of agreement and surprisal, suggesting that both manipulations drive response times. In addition, there was a main effect of number, with longer reading times for nouns following plural verbs than nouns following singular verbs.

In sum, at the second spill-over words, no interactions remained. We observed the effects of surprisal, agreement and number all in the same direction as in the

Table 4.5: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the second spill-over word. Model: residual log RT ~ agreement + surprisal + correct number + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-9.12e^{-02}	1.52e^{-02}	1.80e ^{03}	-6.01	2.27e^{-09}	***
Surprisal	4.57e^{-03}	9.45e^{-04}	3.38e ^{03}	4.84	1.38e^{-06}	***
Agreement	6.43e^{-02}	8.36e^{-03}	3.38e ^{03}	7.68	2.01e^{-14}	***
Correct number	2.01e^{-02}	8.14e^{-03}	3.38e ^{03}	2.47	0.0138	*

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

previous two words: longer reading times are found for higher surprisal values, ungrammatical sentences, and sentences with a plural verb, respectively.

4.3.4 Spill-over 3

At the third word after the target word – and the last one analyzed here – the effect of surprisal appeared to subside. The only possible random effects structure was a random intercept. Comparison for fixed effects showed no contribution from the interaction between agreement and surprisal ($\chi^2(1) = 0.57$; $p = 0.45$), and no contribution from surprisal ($\chi^2(1) = 1.53$; $p = 0.22$) with respect to model fit. The interaction between number and agreement also did not contribute to model fit ($\chi^2(1) = 0.86$; $p = 0.35$), nor did the main effect of number ($\chi^2(1) = 0.48$; $p = 0.49$). Agreement did contribute positively to model fit ($\chi^2(1) = 34.48$; $p < 0.01$). The output of this model, with a random intercept and a fixed effect of agreement, is shown in Table 4.6 below. The model revealed a main effect of, with higher RTs for incorrect agreement, indicating that slower reading times after an incorrect target word persists (at least) until the third word after the agreement error. This can be seen in Figure 4.3 above.

In sum, at the last word analyzed, only the effect of agreement remained.

Table 4.6: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the third spill-over word. Model: residual log RT ~ agreement + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-2.86e^{-03}	7.77e^{-03}	1.57e ^{02}	-0.368	0.713	
Agreement	4.62e^{-02}	7.84e^{-03}	3.39e ^{03}	5.887	4.32e^{-09}	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.4 Discussion

In this study, we aimed to gain insight into the cognitive architecture of language that can give rise to both effects of morphosyntactic nature and effects with a probabilistic origin. To this end, we evaluated whether contextually driven lexical probability interacts with the establishment of subject-verb number agreement in Dutch. We performed a self-paced reading study in which we pit probabilistic lexical information against subject-verb number agreement, by constructing sentences with high- and low surprisal target nouns that did or did not agree in number with the verb. To construct a constraining context, and due to verb-second constraints in Dutch, the verb preceded the target noun. On the basis of three theoretical frameworks, hypotheses were formed: cue-based retrieval, surprisal theory, and language processing as cue integration (LPCI). In brief, cue-based retrieval models predict an ungrammaticality effect, with longer reading times for incorrect agreement than for correct agreement. A strong interpretation of surprisal theory suggests that reading times should be captured by surprisal alone; there should be no separable ungrammaticality effect. LPCI suggests that both factors should affect reading times, and that the ungrammaticality effect may be smaller for low surprisal words than for high surprisal words as a consequence of path dependence; however, the ungrammaticality effect cannot be eliminated.

The results revealed that agreement and surprisal both affected reading times. Incorrect agreement led to longer reading times on the whole time-window analyzed (the target word and three spill-over words). This effect was found for both plural and singular verbs, despite the effect being reversed for plural verbs on the target word: singular nouns following plural verbs were read faster than plural nouns following plural verbs. This was only the case at the target word; the sign of the effect had the predicted directionality from the first spill-over region onward, with incorrect singular nouns receiving longer reading times than correct plural nouns. In addition, higher surprisal values were associated with longer reading times. This effect was observed most clearly when using continuous surprisal as a predictor, with the effect of surprisal present at the target word and two out of three spill-over regions. There was only limited evidence for an interaction between surprisal and agreement: at the first spill-over word was there a marginal interaction between the two factors. Upon further analysis, this interaction appeared to be driven by surprisal affecting reading times in grammatical sentences, but not in ungrammatical sentences. In the other regions of interest there were no further interactions.

4.4.1 Agreement and lexical predictability

The finding that both surprisal and agreement drive reaction times points toward theories that incorporate both probabilistic and grammatical processing, such as LPCI, and it excludes the strong interpretation of surprisal theory. Lexical surprisal is capable of capturing agreement errors: indeed, surprisal is higher for incorrect nouns (through essentially capturing the probability of a word form). However, surprisal alone is not sufficient to account for a reader's processing of agreement errors. This shows that while surprisal predicts the correct directionality as shown by Ryu and Lewis (2021), a model that includes exclusively lexical probabilistic relations such as GPT2 is not sufficient to account for the comprehension of subject-verb agreement.

This finding is in line with a study by van Schijndel and Linzen (2021). There, participants read temporarily ambiguous but grammatically correct sentences. The study showed that while surprisal values predicted the location of a slowdown in an ambiguous sentence, they severely underestimated the size of the effect. Crucially, their findings and ours are incompatible with some instantiations of surprisal theory (Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018), which attempt to explain measures of linguistic processing difficulty (e.g., reading times) from a single mechanism: lexical probability.

Instead, our results point toward accounts that model language comprehension as a process that is shaped by both grammatical constraints (such as the number match between subject and verb in subject-verb number agreement) and probabilistic constraints (such as the probability of a word in context). Importantly, our results suggest that the effect of grammaticality is not *overridden* by lexical predictability.[1] This is in line with models that include some bias towards grammatical input, such as the LPCI (Martin, 2016, 2020), cue-based retrieval models (Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006; Vasishth, 2001), feature percolation models (Eberhard, 1997; Eberhard, Cutting, & Bock, 2005), and the hybrid model of the latter two as well as a grammaticality bias model from the computational study by Yadav and colleagues (2023). In these models, the language processing system is biased towards grammatical input in different ways. In cue-based retrieval models and feature percolation models, grammatical knowledge guides language comprehension by supplying abstract features that cue retrieval of previous material from memory or trigger "feature checking" mechanisms. In LPCI, the (grammatical) knowledge of language is the stored information that is probabilistically cued by input, or internal repre-

[1] Importantly, that does not mean that the effects may not be affected in their magnitude.

sentations are cued by equally grammatical higher-level representations. In this framework, the grammaticality bias follows readily from the observation that a given person's acquired (abstract) knowledge of language is, indeed, grammatical.

What these models share, and proposals such as strong interpretations of surprisal theory lack, is that grammatical knowledge poses strong constraints on the expected input (see also: Greco, Cometa, Artoni, Frank, & Moro, 2023). Indeed, in the case of a strong grammatical constraint, lexical probability may be more influential for processing within the limits of grammaticality. In other words, lexical probability may exert a larger influence on the comprehension process when the sentence is grammatical. In the present study, this was indicated by the disappearance of the surprisal effect in ungrammatical conditions in the first spill-over region. Such an interpretation is also in line with previous findings that show attraction effects for ungrammatical, but not grammatical configurations (Tanner et al., 2014; Wagers et al., 2009): alternative bindings in the sentence are only sought for when the verb does not agree with noun that is in the correct syntactic position to be the head noun of the subject NP. This implies that the path that one needs to go down in order to interpret an incorrect sentence is not one that can readily be taken on the basis of the canonical mechanisms; to go off the beaten track, we might need to invoke additional mechanisms in order to reconcile the conflict between the expected (grammatical) features and the observed (ungrammatical) features.

At the same time, in contrast with our predictions on the basis of the LPCI, the present results do not provide evidence for modulation of magnitude of the ungrammaticality effect in subject-verb agreement by lexical probability. This is in contrast with the findings by Tung and Brennan (2023), who showed that a P600-effect as a result of incorrect ellipsis resolution could be reduced by higher lexical predictability. There are numerous differences between this study and ours, most notably that ellipsis resolution is very different from subject-verb agreement from a grammatical perspective. Ellipsis resolution concerns sentences in which information that has been supplied earlier, is omitted; consider 'one', which refers back to 'a shirt', in (4).

(4) The mother brought **one**$_{\text{[classifier]}}$ **shirt** that was next to the luggage, and the daughter also brought **one**$_{\text{[classifier]}}$ to go on a trip.

The manipulation of grammaticality was done on a classifier in Mandarin Chinese. This classifier exists in two types: one for individual objects, and one books

and pamphlets. The classifier occurs before 'shirt', and it is the referent 'one' in a context of ellipsis. Importantly, the classifier must be the same in both positions in order to refer back to the shirt. The authors show that a P600 is triggered when there is a mismatch between classifiers, and this P600 is modulated by the predictability of the target given the verb (Tung & Brennan, 2023).

Interestingly, the LPCI itself provides an explanation for the difference between these studies. Notice that the occurrence of ellipsis itself is much less predictable than the (obligatory) occurrence of subject-verb agreement. For that reason, the grammatical cue may not be as reliable in the case of ellipsis, as such giving room for lexical probability to play a role. Indeed, this view is fully consistent with the LPCI: knowledge of the grammar, just like the contextual probability, can cue other (lower-level) representations (Kaufeld, Ravenschlag, Meyer, Martin, & Bosker, 2020; Martin, 2016, 2020), and cues are weighted on the basis of their reliability (Martin, 2016, 2020). In the case of subject-verb agreement, the grammar very reliably cues a morphosyntactic representation of the verb (or the subject, depending on the order; more on this below). In the case of ellipsis, on the other hand, the grammar does provide constraints, but given the greater uncertainty about the rest of the sentence, the reliability of these constraints is lower.

As a logical consequence, this perspective suggests that the contextual probability cue used in this study was not strong enough to modulate the subject-verb agreement effect. This leaves us with several open questions. Firstly, we do not know whether it is possible to manipulate contextual lexical probability to such an extent (i.e., by having more extreme values of surprisal) that it is reliable enough to affect the ungrammaticality effect in its current form, or whether this can only be done by reducing the reliability of the grammatical cue, for example by introducing attractors. Secondly, it is unknown whether only lexical surprisal, i.e., the probability of a word given the context, plays a role in this process, or whether entropy, a metric of uncertainty about the surprisal values, is involved, as well; after all, entropy can be considered a metric of the reliability of the probabilistic cue.

4.4.2 Word order effects

An important difference between the present study and previous studies on subject-verb agreement is the order of the constituents. In most studies, the verb follows the subject (order subject-verb). In the present study, however, the subject followed the verb. This is a consequence of our manipulation of surpri-

sal. By starting the sentence with a prepositional phrase, we were able to create sufficient context for the target – in this case, the noun – to be manipulated for surprisal, while matching word frequency within a pair (see section 4.2.2 and Appendix I, 4.6). However, a prepositional phrase in first position reveals the *verb second* property of Dutch (and some other Germanic languages). Usually, the subject and verb move all the way up to the highest two positions in the structure. However, in our stimuli, the highest position is filled by the prepositional phrase. The verb moves to the second position, but the subject remains lower in the structure. Our results showed that this significant difference does not affect the directionality of ungrammaticality effect previously found (Nicol et al., 1997; Pearlmutter et al., 1999; Wagers et al., 2009), despite the effect now being measured on the subject, and not on the verb as is habitual: subjects that do not agree with the preceding verb are read more slowly than subjects that do agree with the preceding verb. In general, these findings indicate that the reversal of the subject and the verb do not qualitatively change the way different linguistic pressures shape reading times (i.e., the directionality is identical), though they may do so quantitatively (the slowdown may be larger or smaller). Importantly, this qualitative similarity of the data to those of earlier studies does not mean that the underlying mechanism that leads to a slowdown or speed-up in reading times is the same in the case of VS and SV sentences.

According to (psycho)linguistic theory, the verb is inflected to establish a relationship with the subject, and not the other way around. When the comprehender encounters an inflected verb, then, the expectations on the morphology of the subject are very strong: after all, the subject must be hypothesized for the verb to agree with it. This is especially true in some frameworks from theoretical linguistics (e.g., the minimalist tradition), in which the subject and verb are hypothesized to both originate from a verb phrase as the specifier and the head, respectively. The verb moves out of this structure to establish the agreement relation with the subject (see: Franck, Lassi, Frauenfelder, & Rizzi, 2006). Following this perspective, encountering an inflected verb *first* means that the parser will have to build a verb phrase with a hypothetical subject in the specifier for the verb to agree with. Encountering the subject first does not require the parser to hypothesize an inflected verb, however; the derivation process may happen in its natural order. From a psycholinguistic perspective, hypothesizing or prespecifying a number on the upcoming verb may not be beneficial: problems would arise when the noun appears to be part of a compound NP (Nicol et al., 1997, p. 585).

This suggests that the grammatical cue during subject-verb agreement may be stronger when the verb precedes the subject. With respect to the present study, this means that stronger probabilistic cues may be required to find an alteration of the agreement effect in the case of verb-subject agreement than in the case of subject-verb agreement. This hypothesis could be tested by repeating the present study using a language without a verb second property, such as English, Spanish or Mandarin Chinese.

4.4.3 Number effects

In potential contradiction with the hypotheses laid out in the previous section, there is a factor that can affect the agreement effect in the verb-subject structure: number. As is often found in studies of agreement with classical SVO structures as well, we observed an imbalance of the number of the verb on the subject: singular subjects incorrectly following plural verbs were read faster than plural subjects correctly following plural verbs at the target word, while plural subjects incorrectly following singular verbs are read more slowly than singular subjects correctly following singular verbs. This reversal of the ungrammaticality effect disappears after the target word. When we rank the reading times on the target word from fastest to slowest for the different configurations, we find the following: $V_{sin}N_{sin} < V_{pl}N_{sin} < V_{sin}N_{pl} < V_{pl}N_{pl}$. The reading times are ordered along (1) the number of the subject and (2) the number of the verb. In essence, this means that the slowdown associated with plural verbs and plural nouns relative to singular verbs and singular nouns can override any effects of ungrammaticality. At the first spill-over word, we again find an interaction between verb number and agreement: the difference between reading times for words following plural verbs and singular verbs is significant only in the grammatical condition. We will get back to this below. Notice that these findings are not associated with word length or word frequency; the reading times were residualized for these features prior to analysis.

The results are in line with the theory that plural number is marked relative to singular number (Bock & Eberhard, 1993; Eberhard, 1997; Pearlmutter et al., 1999). According to this theory, plurals contain a grammatical specification for number, while singulars do not. At the target, number plays a greater role than agreement, as can be seen in the ordering of the reading times above: the plural verb – plural noun combination receives the longest reading times at the target. In later regions (the spill-over words) the pattern changes slightly (more on this below). That this feature can override the effect of number in early regions is not

necessarily a surprise: after all, the number features on the noun and verb must be determined *before* the agreement relation can be established (by retrieval or feature checking).

To provide a full picture of the data, we performed the model-comparison analysis using the number on the *subject* as a predictor in the linear mixed effects model, instead of the number on the verb. Full output of the models is provided in Appendix III, 4.8. This showed that plural subjects lead to slower reading times than singular subjects at the target word ($\beta = 3.75{\cdot}10^{-2}$, SE $= 8.10{\cdot}10^{-3}$, t(3310) $= 4.63$, $p < 0.01$) and the first spill-over word ($\beta = 2.62{\cdot}10^{-2}$, SE $= 8.70{\cdot}10^{-3}$, t(3300) $= 3.01$, $p < 0.01$), irrespective of the grammaticality of the subject. At the second spill-over word, there was an interaction between the number on the subject and agreement ($\beta = -4.03{\cdot}10^{-2}$, SE $= 1.63{\cdot}10^{-3}$, t(3380) $= -2.47$, $p < 0.05$). Here, we observe a difference between plural and singular subjects only in the *incorrect* sentences: singular subjects that incorrectly follow plural verbs lead to *longer* reading times at the second-spill over word than plural subjects that incorrectly follow singular verbs (F(1,3383.61) $= 6.06$, $p = 0.056$). This is not the case in grammatical sentences (F(1,3383.21) $= 1.06$; $p = 1$). Any effect of the number of the subject disappears at the third spill-over word. These findings are in line with findings by Wagers and colleagues (2009), who employed a canonical subject-verb word order: in the study on subject-verb agreement without attractors, the authors report that singular subjects lead to larger ungrammaticality effects in later regions (i.e., after the critical verb) than plural subjects do.

An interesting pattern within these results emerges in later regions. In the first spill-over region, the number of the preceding verb plays a role in the *grammatical* conditions; and in the second spill-over, the number of the following subject plays a role only in the *ungrammatical* conditions. A model that accounts for these effects is a model that contains both an early markedness effect for plural subjects and verbs, as well as a grammaticality bias that has preference over the markedness effect in later regions. When the preceding verb is plural, our knowledge of the grammar poses a strong constraint on the subject that is yet to appear: it reliably cues the marked plural form. The strength of this grammatical cue could dampen any effect the marked plural subject may have on reading times. When the constraint is violated, however, the markedness effect can be observed again: incorrect plural subjects lead to longer reading times than incorrect singular subjects (in later stages). This reading of the results is in line with the flexible weighting of cues in the LPCI model.

4.5 Conclusion

The present study contributes to a growing literature examining how probability and uncertainty shape language comprehension in close collaboration with grammatical knowledge. The literature suggests that lexical probabilistic information can affect how the comprehender leverages syntactic cues (Brehm et al., 2020; Campanelli et al., 2018; Loerts et al., 2013; Tung & Brennan, 2023). The results of the present study indicate that both lexical probabilistic information as well as grammatical information are needed to describe reading time data from a subject-verb agreement paradigm. This is in direct contrast with proposals that model this phenomenon and language comprehension more generally using exclusively lexical probabilistic information (Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018; Ryu & Lewis, 2021). At the same time, the results suggest that the morphosyntactic cue provided by the subject or the verb in subject-verb agreement in Dutch is stronger than the cue of contextual lexical probability as used here: the ungrammaticality effect was not altered by lexical probability. In addition, the data provide some evidence that lexical probability is leveraged more reliably when the constraints placed by the grammar are obeyed. Taken together with previous findings, the results paint a picture in which grammatical cues, as well as contextual probabilistic cues, are weighted on the basis of their reliability (Martin, 2016, 2020). In the case of adjacent subject-verb agreement, the grammatical cue is highly reliable. To study how the effects of lexical probability and the establishment of subject-verb agreement interact in more detail, further studies could employ larger contextual probability effects, and alter the reliability of the syntactic cue by adding attractors to the stimuli.

4.6 Appendix I. Stimuli

Table 4.7: Experimental items.

Set	Condition surprisal	Condition agreement	Context	Target	Spill-over	Target surprisal (bits)	Target word freq.	Verb number
2	Low	Correct	Vanwege de ziekte schrijft de	dokter	een briefje voor school	6,31	4,0283	singular
2	Low	Incorrect	Vanwege de ziekte schrijft de	dokters	een briefje voor school	10,46	2,8657	singular
2	High	Correct	Vanwege de ziekte schrijft de	papa	Een briefje voor school	17,42	3,9897	singular
2	High	Incorrect	Vanwege de ziekte schrijft de	papa's	een briefje voor school	22,24	2,5328	singular
4	Low	Correct	In de klas vertelt de	docent	een heel spannend verhaal	6,5	2,1761	singular
4	Low	Incorrect	In de klas vertelt de	docenten	een heel spannend verhaal	8,5	1,6335	singular
4	High	Correct	In de klas vertelt de	prof	een heel spannend verhaal	16,51	2,1818	singular
4	High	Incorrect	In de klas vertelt de	profs	een heel spannend verhaal	20,21	2,0294	singular
5	Low	Correct	Volgens de verslaggever lopen de	ministers	gelukkig geen gevaar meer	11,7	1,8633	plural
5	Low	Incorrect	Volgens de verslaggever lopen de	ministers	gelukkig geen gevaar meer	11,68	3,0523	plural
5	High	Correct	Volgens de verslaggever lopen de	pony's	gelukkig geen gevaar meer	15,18	1,8633	plural
5	High	Incorrect	Volgens de verslaggever lopen de	pony	gelukkig geen gevaar meer	14,94	2,3345	plural
6	Low	Correct	Bij het monument houden de	bewakers	zich aan de regels	11,62	2,9987	plural
6	Low	Incorrect	Bij het monument houden de	bewaker	zich aan de regels	14,71	2,9513	plural
6	High	Correct	Bij het monument houden de	idioten	zich aan de regels	18,25	3,0314	plural
6	High	Incorrect	Bij het monument houden de	idioot	zich aan de regels	19,17	3,7101	plural
7	Low	Correct	Tijdens het eten bespreken de	zussen	hun plannen voor morgen	11,69	2,7474	plural
7	Low	Incorrect	Tijdens het eten bespreken de	zus	hun plannen voor morgen	13,67	3,7582	plural
7	High	Correct	Tijdens het eten bespreken de	dieven	hun plannen voor morgen	14,91	2,7474	plural
7	High	Incorrect	Tijdens het eten bespreken de	dief	hun plannen voor morgen	16,03	3,1176	plural
8	Low	Correct	Op het schoolplein spelen de	kleuters	een heel lawaaiïg spel	7,7	1,716	plural
8	Low	Incorrect	Op het schoolplein spelen de	kleuter	een heel lawaaiïg spel	10,26	1,8633	plural

8	High	Correct	Op het schoolplein spelen de	teamgenoten	een heel lawaaiïg spel	17,8	1,716	plural
8	High	Incorrect	Op het schoolplein spelen de	teamgenoot	een heel lawaaiïg spel	21,46	1,4771	plural
9	Low	Correct	Bij de groothandel koopt de	winkelier	een enorme hoeveelheid bier	11,34	1,8195	singular
9	Low	Incorrect	Bij de groothandel koopt de	winkeliers	een enorme hoeveelheid bier	12,9	1,5798	singular
9	High	Correct	Bij de groothandel koopt de	mafketel	een enorme hoeveelheid bier	27,02	1,8195	singular
9	High	Incorrect	Bij de groothandel koopt de	mafketels	een enorme hoeveelheid bier	29,63	1,5563	singular
12	Low	Correct	In de herfst controleert de	beheerder	de bomen op ziektes	11,2	2	singular
12	Low	Incorrect	In de herfst controleert de	beheerders	de bomen op ziektes	13,4	1,2553	singular
12	High	Correct	In de herfst controleert de	blondine	de bomen op ziektes	19,16	2,0043	singular
12	High	Incorrect	In de herfst controleert de	blondines	de bomen op ziektes	22,21	1,3979	singular
13	Low	Correct	In de kantine kopen de	jongeren	alleen maar ongezond eten	11,08	2,6096	plural
13	Low	Incorrect	In de kantine kopen de	jongere	alleen maar ongezond eten	14,88	2,5289	plural
13	High	Correct	In de kantine kopen de	tieners	alleen maar ongezond eten	15,45	2,4997	plural
13	High	Incorrect	In de kantine kopen de	tiener	alleen maar ongezond eten	15,63	2,567	plural
14	Low	Correct	In dat ziekenhuis liggen de	patiënten	na het ernstige ongeluk	5,23	2,9523	plural
14	Low	Incorrect	In dat ziekenhuis liggen de	patiënt	na het ernstige ongeluk	9,85	3,1477	plural
14	High	Correct	In dat ziekenhuis liggen de	vriendinnen	na het ernstige ongeluk	15,23	2,9523	plural
14	High	Incorrect	In dat ziekenhuis liggen de	vriendin	na het ernstige ongeluk	14,88	3,8804	plural
15	Low	Correct	Na de vergadering moeten de	advocaten	snel naar een afspraak	10,72	2,9609	plural
15	Low	Incorrect	Na de vergadering moeten de	advocaat	snel naar een afspraak	12,16	3,6532	plural
15	High	Correct	Na de vergadering moeten de	broeders	snel naar een afspraak	14,09	2,9763	plural
15	High	Incorrect	Na de vergadering moeten de	broeder	snel naar een afspraak	15,49	3,1962	plural
17	Low	Correct	In de studio schildert de	kunstenaar	van vroeg tot laat	7	2,7076	singular
17	Low	Incorrect	In de studio schildert de	kunstenaars	van vroeg tot laat	9,66	2,1903	singular
17	High	Correct	In de studio schildert de	kameraad	van vroeg tot laat	16,72	2,6964	singular
17	High	Incorrect	In de studio schildert de	kameraden	van vroeg tot laat	17,68	2,4914	singular
18	Low	Correct	Door de muziek hoort de	zanger	het luide alarm niet	7,73	2,4362	singular
18	Low	Incorrect	Door de muziek hoort de	zangers	het luide alarm niet	10,67	1,8865	singular
18	High	Correct	Door de muziek hoort de	boef	het luide alarm niet	16,4	2,4546	singular

18	High	Incorrect	Door de muziek hoort de	boeven	het luide alarm niet	16,92	2,5145	singular
19	Low	Correct	In de winter sjokt de	beer	door de verse sneeuw	11,58	3,0469	singular
19	Low	Incorrect	In de winter sjokt de	beren	door de verse sneeuw	11,94	2,6138	singular
19	High	Correct	In de winter sjokt de	kampioen	door de verse sneeuw	15,71	3,0554	singular
19	High	Incorrect	In de winter sjokt de	kampioenen	door de verse sneeuw	17,8	2,1335	singular
20	Low	Correct	In de parkeergarage stopt de	auto	plots met slippende banden	3,22	4,3017	singular
20	Low	Incorrect	In de parkeergarage stopt de	auto's	plots met slippende banden	5,23	3,2923	singular
20	High	Correct	In de parkeergarage stopt de	vriend	plots met slippende banden	15,31	4,3323	singular
20	High	Incorrect	In de parkeergarage stopt de	vrienden	plots met slippende banden	16	4,1627	singular
22	Low	Correct	In de bergen lopen de	geiten	altijd in de schaduw	11,12	2,2504	plural
22	Low	Incorrect	In de bergen lopen de	geit	altijd in de schaduw	14,82	2,549	plural
22	High	Correct	In de bergen lopen de	schutters	altijd in de schaduw	14,93	2,2455	plural
22	High	Incorrect	In de bergen lopen de	schutter	altijd in de schaduw	15,88	2,7931	plural
23	Low	Correct	Direct na aardrijkskunde moeten de	leerlingen	hun strafwerk gaan maken	4,87	2,7882	plural
23	Low	Incorrect	Direct na aardrijkskunde moeten de	leerling	hun strafwerk gaan maken	8,97	2,7551	plural
23	High	Correct	Direct na aardrijkskunde moeten de	sukkels	hun strafwerk gaan maken	17,82	2,7796	plural
23	High	Incorrect	Direct na aardrijkskunde moeten de	sukkel	hun strafwerk gaan maken	18,94	3,2541	plural
24	Low	Correct	Tijdens de ceremonie moeten de	aanwezigen	één voor één opstaan	5,95	1,9823	plural
24	Low	Incorrect	Tijdens de ceremonie moeten de	aanwezige	één voor één opstaan	9,63	1,6812	plural
24	High	Correct	Tijdens de ceremonie moeten de	immigranten	één voor één opstaan	16,24	1,9823	plural
24	High	Incorrect	Tijdens de ceremonie moeten de	immigrant	één voor één opstaan	19,46	1,716	plural
25	Low	Correct	Op de camping slaapt de	gast	van de grote bruiloft	11,41	3,3911	singular
25	Low	Incorrect	Op de camping slaapt de	gasten	van de grote bruiloft	8,75	3,3649	singular
25	High	Correct	Op de camping slaapt de	getuige	van de grote bruiloft	15,17	3,3948	singular
25	High	Incorrect	Op de camping slaapt de	getuigen	van de grote bruiloft	16,49	3,33	singular
26	Low	Correct	Met die schapen heeft de	herder	nog helemaal geen ervaring	8,97	2,415	singular
26	Low	Incorrect	Met die schapen heeft de	herders	nog helemaal geen ervaring	9,68	1,7404	singular
26	High	Correct	Met die schapen heeft de	specialist	nog helemaal geen ervaring	15,99	2,4099	singular
26	High	Incorrect	Met die schapen heeft de	specialisten	nog helemaal geen ervaring	16,59	2,0294	singular

27	Low	Correct	Op het terras rent de	ober	van tafel naar tafel	6,85	2,6424	singular
27	Low	Incorrect	Op het terras rent de	obers	van tafel naar tafel	11,78	1,7559	singular
27	High	Correct	Op het terras rent de	vrijgezel	van tafel naar tafel	15,32	2,6454	singular
27	High	Incorrect	Op het terras rent de	vrijgezellen	van tafel naar tafel	15,76	1,9494	singular
28	Low	Correct	Op de zeebodem ligt de	vis	die vorige week stierf	9,75	3,3406	singular
28	Low	Incorrect	Op de zeebodem ligt de	vissen	die vorige week stierf	11,81	3,246	singular
28	High	Correct	Op de zeebodem ligt de	soldaat	die vorige week stierf	15,41	3,3655	singular
28	High	Incorrect	Op de zeebodem ligt de	soldaten	die vorige week stierf	14,59	3,3549	singular
29	Low	Correct	Vangwege het luchtalarm zoeken de	bewoners	direct een veilige plaats	7,1	2,4771	plural
29	Low	Incorrect	Vangwege het luchtalarm zoeken de	bewoner	direct een veilige plaats	13,61	1,9823	plural
29	High	Correct	Vangwege het luchtalarm zoeken de	zwervers	direct een veilige plaats	15,15	2,2279	plural
29	High	Incorrect	Vangwege het luchtalarm zoeken de	zwerver	direct een veilige plaats	15,31	2,4728	plural
30	Low	Correct	In de vrachtwagen liggen de	pakketten	voor de grote drogisterij	11,16	1,3802	plural
30	Low	Incorrect	In de vrachtwagen liggen de	pakket	voor de grote drogisterij	17,71	2,3284	plural
30	High	Correct	In de vrachtwagen liggen de	parfums	voor de grote drogisterij	21,29	1,3802	plural
30	High	Incorrect	In de vrachtwagen liggen de	parfum	voor de grote drogisterij	18,95	2,6785	plural
31	Low	Correct	In de kratten zitten de	appels	die afgeleverd moeten worden	11,46	2,4983	plural
31	Low	Incorrect	In de kratten zitten de	appel	die afgeleverd moeten worden	13,54	2,6503	plural
31	High	Correct	In de kratten zitten de	kopieën	die afgeleverd moeten worden	16,2	2,4955	plural
31	High	Incorrect	In de kratten zitten de	kopie	die afgeleverd moeten worden	17,58	2,8681	plural
32	Low	Correct	Op de A27 staan de	vrachtwagens	in een lange file	9	2,2695	plural
32	Low	Incorrect	Op de A27 staan de	vrachtwagen	in een lange file	11,85	2,8768	plural
32	High	Correct	Op de A27 staan de	beroemdheden	in een lange file	17,99	2,2304	plural
32	High	Incorrect	Op de A27 staan de	beroemdheid	in een lange file	18,94	2,2718	plural
33	Low	Correct	Tijdens het protest beschermt de	held	de gevallen oude man	11,81	3,4165	singular
33	Low	Incorrect	Tijdens het protest beschermt de	helden	de gevallen oude man	12,17	2,8169	singular
33	High	Correct	Tijdens het protest beschermt de	tante	de gevallen oude man	16,64	3,4357	singular
33	High	Incorrect	Tijdens het protest beschermt de	tantes	de gevallen oude man	17,18	2,1644	singular
34	Low	Correct	In het televisieprogramma vertelt de	acteur	niet over het verleden	7,4	2,9633	singular

34	Low	Incorrect	In het televisieprogramma vertelt de	acteurs	niet over het verleden	9,89	2,7597	singular
34	High	Correct	In het televisieprogramma vertelt de	lafaard	niet over het verleden	19,45	2,9542	singular
34	High	Incorrect	In het televisieprogramma vertelt de	lafaards	niet over het verleden	21,14	2,4065	singular
35	Low	Correct	De volgende morgen vertrekt de	directeur	naar de zonnige bestemming	7,79	3,1998	singular
35	Low	Incorrect	De volgende morgen vertrekt de	directeuren	naar de zonnige bestemming	15,7	1,6335	singular
35	High	Correct	De volgende morgen vertrekt de	leugenaar	naar de zonnige bestemming	20,69	3,194	singular
35	High	Incorrect	De volgende morgen vertrekt de	leugenaars	naar de zonnige bestemming	20,45	2,2856	singular
36	Low	Correct	Onder mijn bureau ligt de	pen	die ik kwijt was	9,31	2,9768	singular
36	Low	Incorrect	Onder mijn bureau ligt de	pennen	die ik kwijt was	12,19	2,0569	singular
36	High	Correct	Onder mijn bureau ligt de	rat	die ik kwijt was	15,1	2,9978	singular
36	High	Incorrect	Onder mijn bureau ligt de	ratten	die ik kwijt was	16,95	2,8854	singular
37	Low	Correct	Op de velden werken de	boeren	met een grote ploeg	7,36	2,8254	plural
37	Low	Incorrect	Op de velden werken de	boer	met een grote ploeg	9,06	2,8109	plural
37	High	Correct	Op de velden werken de	personen	met een grote ploeg	14,69	2,8325	plural
37	High	Incorrect	Op de velden werken de	persoon	met een grote ploeg	15,76	3,6146	plural
39	Low	Correct	Op het strand vliegen de	meeuwen	je om de oren	4,79	1,6128	plural
39	Low	Incorrect	Op het strand vliegen de	meeuw	je om de oren	16,39	1,4624	plural
39	High	Correct	Op het strand vliegen de	volleyballen	je om de oren	19,39	1,1461	plural
39	High	Incorrect	Op het strand vliegen de	volleybal	je om de oren	15,82	1,6128	plural
40	Low	Correct	In dat pretpark zijn de	attracties	eigenlijk niet in orde	8,43	1,699	plural
40	Low	Incorrect	In dat pretpark zijn de	attractie	eigenlijk niet in orde	10,22	1,9685	plural
40	High	Correct	In dat pretpark zijn de	kwalificaties	eigenlijk niet in orde	15,28	1,699	plural
40	High	Incorrect	In dat pretpark zijn de	kwalificatie	eigenlijk niet in orde	15,17	1,301	plural
41	Low	Correct	Naar die sportschool gaat de	trainer	nooit om te sporten	9,15	2,5527	singular
41	Low	Incorrect	Naar die sportschool gaat de	trainers	nooit om te sporten	12,22	1,6902	singular
41	High	Correct	Naar die sportschool gaat de	kanjer	nooit om te sporten	18,98	2,5514	singular
41	High	Incorrect	Naar die sportschool gaat de	kanjers	nooit om te sporten	19,47	1,5185	singular
43	Low	Correct	Vanwege de afmeldingen moet de	familie	het feest helaas afblazen	9,72	4,1239	singular
43	Low	Incorrect	Vanwege de afmeldingen moet de	families	het feest helaas afblazen	15,05	2,8439	singular

43	High	Correct	Vanwege de afmeldingen moet de	mama	het feest helaas afblazen	18,67	3,9556	singular
43	High	Incorrect	Vanwege de afmeldingen moet de	mama's	het feest helaas afblazen	21,12	2,412	singular
44	Low	Correct	Op deze zeilboot vaart de	schipper	een rondje door Nederland	11,39	2,1584	singular
44	Low	Incorrect	Op deze zeilboot vaart de	schippers	een rondje door Nederland	16,86	1,1461	singular
44	High	Correct	Op deze zeilboot vaart de	patriot	een rondje door Nederland	20,3	2,1644	singular
44	High	Incorrect	Op deze zeilboot vaart de	patriotten	een rondje door Nederland	18,68	1,6812	singular
45	Low	Correct	Op die kinderboerderij werken de	vrijwilligers	aan de nieuwe hokken	9,8	2,4518	plural
45	Low	Incorrect	Op die kinderboerderij werken de	vrijwilliger	aan de nieuwe hokken	15,17	2,3997	plural
45	High	Correct	Op die kinderboerderij werken de	experts	aan de nieuwe hokken	15,41	2,444	plural
45	High	Incorrect	Op die kinderboerderij werken de	expert	aan de nieuwe hokken	18,4	2,8987	plural
46	Low	Correct	In dat vliegtuig veroorzaken de	passagiers	helaas een enorme chaos	9,32	2,8109	plural
46	Low	Incorrect	In dat vliegtuig veroorzaken de	passagier	helaas een enorme chaos	12,94	2,382	plural
46	High	Correct	In dat vliegtuig veroorzaken de	ouderen	helaas een enorme chaos	14,94	2,3181	plural
46	High	Incorrect	In dat vliegtuig veroorzaken de	oudere	helaas een enorme chaos	15,98	2,871	plural
47	Low	Correct	Met die auto rijden de	coureurs	naar hun beoogde startpunt	11,4	1,8692	plural
47	Low	Incorrect	Met die auto rijden de	coureur	naar hun beoogde startpunt	14,47	1,9494	plural
47	High	Correct	Met die auto rijden de	pelgrims	naar hun beoogde startpunt	14,08	1,8633	plural
47	High	Incorrect	Met die auto rijden de	pelgrim	naar hun beoogde startpunt	16,86	1,7709	plural
48	Low	Correct	In de achtertuin sluipen de	boeven	om ongezien te blijven	10,38	2,5145	plural
48	Low	Incorrect	In de achtertuin sluipen de	boef	om ongezien te blijven	13,7	2,4346	plural
48	High	Correct	In de achtertuin sluipen de	misdadigers	om ongezien te blijven	15,22	2,4609	plural
48	High	Incorrect	In de achtertuin sluipen de	misdadiger	om ongezien te blijven	17,12	2,5079	plural

Table 4.8: Filler items.

Filler number	Referent (if applicable)	Sentence
F1	Ambiguous	Abel leende Gijs het boek voor hij op vakantie ging
F2	Second	Lisanne lachte Jordy uit toen hij de grap begreep
F3	First	Twan praatte met Britt over de wedstrijd vlak voor hij een ongeluk kreeg
F4	Ambiguous	Bas vertelde Hendrik over het probleem vlak voor hij het gebouw verliet
F5	Second	Nora sprak met Ryan over de sollicitatie terwijl hij de koffers inpakte
F6	First	Benjamin vroeg Veerle naar het geld toen hij de straat overstak
F7	Ambiguous	Jasper knipoogde naar Felix toen hij het theater binnen liep
F8	Second	Ella gaf het briefje aan David voordat hij de woonkamer in liep
F9	First	Evert gaf het boek aan Maartje terug voordat hij naar Londen verhuisde
F10	Ambiguous	Tom gaf de microfoon aan Kasper voor hij het podium over rende
F11	Second	Nina knikte naar Florian terwijl hij de rechtszaal in liep
F12	First	Martijn gaf het script aan Esther toen hij de studio binnen kwam
F13	Ambiguous	Tim vertelde de grap aan Victor op het moment dat hij de ruimte in kwam lopen
F14	Second	Johanna glimlachte naar Matthijs toen hij naar Denemarken besloot te fietsen
F15	First	Fabian snauwde een bevel naar Tess toen hij de legerbasis op was gelopen
F16	Ambiguous	Samuel kreeg een bos bloemen van Dirk toen hij naar Duitsland vertrok
F17	Second	Rosalie ontving de brief van Pieter op het moment dat hij in Spanje aan het werk was
F18	First	Thomas miste Isa vanaf het moment dat hij op tournee was gegaan
F19	Ambiguous	Pim praatte met Thom over het tentamen terwijl hij een telefoontje kreeg
F20	Second	Jasmijn vertelde Daan over de toets toen hij de universiteit verliet
F21	First	Alexander sprak met Iris over de crisis vlak voor hij van Frankrijk naar Nederland reed
F22	Ambiguous	Kevin vroeg Stijn naar de uitslag van het onderzoek vlak voor hij door rood reed
F23	Second	Jennifer liet het boek aan Jochem zien toen hij naar cadeaus aan het zoeken was
F24	First	Koen belde met Anna vlak voor hij het instituut verlaten had
F25	Ambiguous	Lars vroeg Teun om een gunst voordat hij de geruchten te horen kreeg
F26	Second	Elise neigde Maurits de waarheid te zeggen voordat hij op vakantie zou gaan
F27	First	Sander liet Suzanne de muziek kiezen toen hij de vrachtwagen bestuurde
F28	Ambiguous	Justin besloot Mike te trakteren omdat hij met verlof zou gaan
F29	Second	Olivia beloofde Floris elke dag te bellen voordat hij de bus instapte
F30	First	Joris gaf Larissa een High five terwijl hij de finish over rende
F31	Ambiguous	Gerrit liet Mick wachten nadat hij de training had afgerond
F32	Second	Romy verschafte Johan een alibi toen hij het politiebureau betrad
F33	First	Oscar stuurde Nicole een email toen hij door ziekte thuis moest blijven
F34	Ambiguous	Tobias belde met Tristan terwijl hij een salade bereidde
F35	Second	Sanne fluisterde iets tegen Patrick voor hij het toneel betrad
F36	First	Stefan vroeg Roos boodschappen te doen nadat hij de presentatie had gegeven
F37	Ambiguous	Sebastiaan belde met Jens terwijl hij naar zee aan het fietsen was
F38	Second	Josephine maakte een foto met Luc nadat hij de camera had gepakt
F39	First	Wesley groette Karlijn op het moment dat hij het plein op wandelde
F40	Ambiguous	Jonas gaf de zonnebrand aan Simon vlak voordat hij de zee inging
F41	Second	Laura sprak met Willem af toen hij van vakantie terug was
F42	First	Niels gaf de bal aan Nienke nadat hij een punt scoorde
F43	Ambiguous	Joep zocht Berend op voordat hij op reis ging
F44	Second	Michelle gaf Olivier een knuffel voordat hij het kunstwerk presenteerde
F45	First	Thijs ging met Fenna naar de dierentuin omdat hij een ticket gewonnen had
F46	Ambiguous	Ruben bezocht Mohamed in Zuid-Limburg toen hij een weekend vrij was
F47	Second	Isabella verstond Max niet terwijl hij het gras aan het maaien was
F48	First	Ties kon Jill niet bereiken vlak voordat hij het vliegtuig instapte
F49	Ambiguous	Rutger zag Timo lopen nadat hij de supermarkt uitkwam
F50	Second	Lynn bevestigde de afspraak met Jurre op het moment dat hij naar Griekenland vertrok
F51	First	Rens bracht Maryam een kop koffie voordat hij op excursie ging

F52	Ambiguous	Luuk vroeg Johannes de deur af te sluiten toen hij het pand verliet
F53	Second	Mila belde Rick op nadat hij de uitslag had gekregen
F54	First	Maarten vloog Lizzy om de hals toen hij de trein uitstapte
F55	Ambiguous	Sjoerd overlegde met Jeffrey voor hij het restaurant reserveerde
F56	Second	Emily kreeg een uitnodiging van Dennis omdat hij de gastvrouw wilde helpen
F57	First	Adam overwoog het bedrijf aan Yara te verkopen omdat hij een uitdaging zocht
F58	Ambiguous	Bart stuurde Arthur een pakketje omdat hij naar Engeland was vertrokken
F59	Second	Daphne vroeg Leon het geld terug te betalen voor hij op reis zou gaan
F60	First	Jonathan fietste die nacht met Emma naar huis omdat hij de omgeving niet veilig vond
F61	N/A	Amber ging die avond naar het ziekenhuis omdat ze pijn in haar benen had
F62	N/A	Chantal nam een hond zodra ze met pensioen ging
F63	N/A	Marieke boekte een vlucht om haar moeder in Griekenland op te zoeken
F64	N/A	Femke haalde een onvoldoende voor de toets omdat ze niet geleerd had
F65	N/A	Isabel had een vaccinatie nodig toen ze naar Vietnam ging
F66	N/A	Anouk lachte uitbundig toen ze het goede nieuws te horen kreeg
F67	N/A	Milou begon aan een nieuwe studie toen ze geen baan kon vinden
F68	N/A	Myrthe ging naar de bouwmarkt zodra ze de sleutel van haar nieuwe huis had gekregen
F69	N/A	Mandy haalde haar rijbewijs makkelijk omdat ze zo veel lessen had gehad
F70	N/A	Manon belde de belastingdienst omdat ze geld terug zou krijgen
F71	N/A	Eva schreef een boek omdat ze haar ervaringen graag wilde delen
F72	N/A	Eline kreeg een groot cadeau van haar ouders toen ze dertig werd
F73	N/A	Joyce kocht een nieuwe fiets omdat ze haar oude niet meer kon vinden
F74	N/A	Lara luisterde naar de radio toen ze ontdekte dat haar tv niet meer werkte
F75	N/A	Madelief was haar portemonnee vergeten toen ze boodschappen ging doen
F76	N/A	Lisa ging op kraambezoek toen ze hoorde dat haar neefje was geboren
F77	N/A	Lieke bakte een taart omdat ze zich verveelde
F78	N/A	Maaike begon aan haar nieuwe baan zodra ze haar ontslag had ingediend
F79	N/A	Renske kleedde zich om toen ze naar het restaurant ging
F80	N/A	Pim ging naar de kapper toen zijn salaris gestort was
F81	N/A	Lena hoopte op beter nieuws maar werd erg teleurgesteld
F82	N/A	Naomi ging altijd op zaterdag naar yoga tot haar yogaschool failliet ging
F83	N/A	Simone las altijd de krant om op de hoogte van het nieuws te blijven
F84	N/A	Sophia beloofde beterschap maar ging daarna toch weer de fout in
F85	N/A	Ilse sprak met een zachte stem om haar kinderen niet wakker te maken
F86	N/A	Carmen probeerde een nieuw recept uit toen haar vrienden op bezoek kwamen
F87	N/A	Pepijn moest plotseling verhuizen toen hij werd overgeplaatst
F88	N/A	Jason verkocht zijn huis toen hij zijn hypotheek niet meer kon betalen
F89	N/A	Milan overnachtte bij een vriend omdat hij zijn huissleutels kwijt was
F90	N/A	Jack vermoedde dat er iets aan de hand was maar wist het niet zeker
F91	N/A	Toen de barkeeper die de irritante klant bediende wegliep gingen de deuren open
F92	N/A	Toen de manke dronkaard die de barkeeper betaalde lachte viel de kruk om
F93	N/A	De rector die de stoere puber strafte was erg onredelijk
F94	N/A	De overtreder die de smeris ontvlucht was is een kronkelig paadje ingerend
F95	N/A	Gisteren had de brede bodyguard die de filmster beschermde een vrije dag
F96	N/A	Jochem die altijd te hoog springt tijdens basketbal heeft zijn enkel bezeerd
F97	N/A	De politieman die de agressieve kraker arresteert gaat hardhandig te werk
F98	N/A	De geduldige dominee die de baby doopt begint te spreken
F99	N/A	De detective die criminelen opspoort krijgt een vette beloning van de staat
F100	N/A	De speurhond die de illegale drugs opspoort krijgt een koekje
F101	N/A	De gedetineerde die de trage bewaker afschudt krijgt hulp van zijn handlanger
F102	N/A	Niemand wist dat een gevaarlijke tornado ontstond die veel slachtoffers zou eisen
F103	N/A	De employee die de ontsnapte dieven beschrijft is erg nauwkeurig
F104	N/A	Het derde getal dat de oplettende toehoorder signaleert is zes
F105	N/A	Mijn vriendin Mona is gek op jongens die gemixte drankjes kopen voor haar
F106	N/A	Het vrouwtje dat verlaten poezen opvangt heeft ruzie met de buren
F107	N/A	Onlangs gaf de jongeman die de populaire portier inhuurde een groot feest
F108	N/A	Het elfje dat de sterke beren betoverde was erg vriendelijk

F109	N/A	De schilder die de knappe prinses tekent zit onder de verf
F110	N/A	De kokkin die pap voor dakloze burgers kookt heeft geen zin meer
F111	N/A	Nora die warme dekens voor arme mensen weeft is gelukkig
F112	N/A	De raadsman die de wrede delinquent bijstond was op de radio
F113	N/A	Joost grijpt de stoere kidnapper die de bange directrice ontglipt in zijn kraag
F114	N/A	Ik sprak gisteren nog met Henk die de scheve torens ontworpen heeft
F115	N/A	Toen de heilsoldaat die de zieke dakloze huisvestte de kamer binnenkwam schrok hij
F116	N/A	De grootvader die de slinkse deugniet doorhad voelde zich bedrogen
F117	N/A	Ik schrok toen de goedkope beunhaas die het achterste bordes repareerde lachte
F118	N/A	De judoka die de dwaze belager vloerde was niet te houden
F119	N/A	De bedrieger die de simpele taxateur misleidde voelde zich schuldig
F120	N/A	De ontwerper die de koningin kleding verkoopt is erg creatief

4.7 Appendix II. Results of the analysis using categorical surprisal

The analyses described below follow the same structure as those in the main text, with the only difference being that the surprisal factor used here is categorical with two levels ('low' and 'high'). This was the original design of the study, but due to an interactive relation between the agreement and surprisal manipulation (the effect of agreement on the surprisal values was larger in high surprisal than in low surprisal stimuli), we decided to use continuous surprisal values as a predictor instead. As can be seen in the results below, the pattern these analyses reveal is similar, although the effect of the categorical surprisal variable is smaller than that of the continuous surprisal variable.

4.7.1 Target word

Reducing the random effects structure revealed that the largest random effects structure that lead to non-singular fit included random slopes for agreement and a random intercept. Model comparison for the fixed effects structure showed that the interaction between surprisal and agreement did not improve model fit ($\chi^2(1) = 1.58$; $p = 0.21$), and surprisal did not do so either ($\chi^2(1) = 2.37$; $p = 0.12$). The interaction between agreement and number did contribute to model fit ($\chi^2(1) = 20.84$; $p < 0.01$). The interpreted model included mains effect of agreement and number as well as their interaction, random slopes for agreement, and a random intercept. This model revealed a main effect of agreement ($\beta = 4.37 \cdot 10^{-2}$, SE $= 1.25 \cdot 10^{-2}$, $t(304.26) = 3.51$, $p < 0.01$), with longer reading times for words directly following an incorrect target word than for words following a correct target word. In addition, there was a main effect of number, with longer reading times of the target word in sentences in which the preceding verb was plural compared to singular ($\beta = 4.58 \cdot 10^{-2}$, SE $= 1.15 \cdot 10^{-2}$, $t(3297.89) = 3.99$, $p < 0.01$). There was also an interaction between these two factors ($\beta = -7.43 \cdot 10^{-2}$, SE $= 1.62 \cdot 10^{-2}$, $t(3294.04) = -4.57$, $p < 0.01$). For a summary, see Table 4.9 below.

Simple effects comparisons revealed that the effect of agreement was in fact reversed for sentences with a plural preceding verb: the incorrect singular nouns were read faster than the correct plural nouns ($F(1,260.1)=6.56$, $p < 0.05$). In sentences with singular preceding verbs, the ungrammaticality effect had the expected directionality ($F(1)=12.31$, $p < 0.01$). This interaction is clearly seen

in figure 4.4B (word position 5). Comparison in the opposite direction (i.e., the effect of number per agreement condition) revealed that the effect of number was significant in grammatical sentences, with plural nouns following a plural verb being read more slowly than singular nouns following singular verbs ($F(1, 3300.30) = 15.94$, $p < 0.01$); in the incorrect condition, there was a trend for singular nouns following plural verbs to be read faster than plural nouns following singular verbs ($F(1, 3299.45) = 6.12$, $p = 0.054$). This can be seen in Figure 4.4D in the main text.

Table 4.9: The output of the interpreted linear mixed effects model of the residual log reading times at the target word (the subject). Model: residual log RT ~ surprisal + correct_number * agreement + (1 + agreement | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	$-8.48e^{-02}$	$1.04e^{-02}$	$1.98e^{02}$	-8.18	$3.45e^{-14}$	***
Agreement	$4.37e^{-02}$	$1.25e^{-02}$	$3.04e^{02}$	3.51	$5.17e^{-04}$	***
Correct number	$4.58e^{-02}$	$1.15e^{-02}$	$3.30e^{03}$	3.99	$6.66e^{-05}$	***
Agreement * correct number	$7.43e^{-02}$	$1.62e^{-02}$	$3.29e^{03}$	-4.57	$5.01e^{-06}$	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.7.2 Spill-over 1

In the first word following the target, the largest random effects structure that converged and did not yield singular fit included random slopes for agreement and a random intercept. Model comparison for the fixed effects structure showed that the interaction between surprisal and agreement did not improve model fit ($\chi^2(1) = 2.23$; $p = 0.14$), and surprisal alone did not, either ($\chi^2(1) = 1.75$; $p = 0.19$). The interaction between agreement and number did contribute to model fit ($\chi^2(1) = 8.27$; $p < 0.01$). The interpreted model included mains effect of agreement and number as well as their interaction, random slopes for agreement, and a random intercept. This model revealed a main effect of agreement ($\beta = 1.01{\cdot}10^{-1}$, $SE = 1.56{\cdot}10^{-2}$, $t(199.90) = 6.44$, $p < 0.01$), with longer reading times for words directly following an incorrect target word than words following a correct target word. This effect is visible in Figure 4.3. In addition, there was a main effect of number, with longer reading times of the target word in sentences in which the preceding verb was plural ($\beta = 3.11{\cdot}10^{-2}$, $SE = 1.23{\cdot}10^{-2}$, $t(3308) = 2.53$, $p < 0.05$). There was also an interaction between these two factors ($\beta = -5.01{\cdot}10^{-2}$, $SE = 1.74{\cdot}10^{-2}$, $t(3301) = -2.88$, $p < 0.01$). See the output of the model in Table 4.10 below.

Simple-effects analyses showed that the ungrammaticality effect had the correct directionality for both numbers (singular verbs: F(1)=41.43, $p_{corr} < 0.01$; plural verbs: F(1,179.92)=11.18, $p_{corr} < 0.01$). A comparison in the opposite direction revealed that the effect of number was only significant in the grammatical conditions (correct agreement: F(1,3311.51)=6.42, $p_{corr} < 0.05$; F(1,3300.99) = 2.37, $p_{corr} = 0.50$). The interaction is shown in Figure 4.4D in the main text (word position 6).

Table 4.10: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the first spill-over word. Model: residual log RT ~ agreement * correct number + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-4.2910^{-2}	$9.40e^{-03}$	$2.99e^{02}$	-4.57	$7.27e^{-06}$	***
Agreement	$1.01e^{-01}$	$1.56e^{-02}$	$2.00e^{02}$	6.44	$8.51e^{-06}$	***
Correct number	$3.11e^{-02}$	$1.23e^{-02}$	$3.31e^{03}$	2.54	$1.13e^{-02}$	*
Agreement * correct number	$-5.01e^{-02}$	$1.74e^{-02}$	$3.30e^{03}$	-2.88	$4.03e^{-03}$	**

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.7.3 Spill-over 2

In the second word following the target word, the only random effects structure that converged was a random intercept. Reduction of the fixed effects showed that the interaction between agreement and surprisal did not contribute to model fit ($\chi^2(1) = 2.16$; $p = 0.14$), whereas the main effect of surprisal ($\chi^2(1) = 12.71$; $p < 0.01$) did. The interaction between number and agreement did not improve model fit ($\chi^2(1) = 1.15$; $p = 0.28$). The main effects of number ($\chi^2(1) = 4.75$; $p < 0.05$) and agreement ($\chi^2(1) = 81.70$; $p < 0.01$) both did contribute to model fit. The interpreted model therefore contained main effects of surprisal, agreement and number and a random intercept. The model showed a main effect of agreement ($\beta = 7.39{\cdot}10^{-2}$, $SE = 8.13{\cdot}10^{-3}$, $t(3760) = 9.09$, $p < 0.01$), revealing longer reading times after an agreement error relative to correct agreement; and a main effect of surprisal ($\beta = 2.90{\cdot}10^{-2}$, $SE = 8.13{\cdot}10^{-3}$, $t(3760) = 3.57$, $p < 0.01$), showing that higher surprisal values lead to longer reading times. These effects are both visible in Figure 4.3 in the main text, at word position 7. In addition, there was an effect of number, again showing longer reading times are associated with nouns that follow a plural verb ($\beta = 1.77{\cdot}10^{-2}$, $SE = 8.14{\cdot}10^{-3}$, $t(3770) = 2.18$, $p < 0.05$).

Table 4.11: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the second spill-over word. Model: residual log RT ~ agreement + surprisal + correct number + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-4.53e^{-02}	9.55e^{-03}	4.30e ^{02}	-4.74	2.89e^{-06}	***
Surprisal	2.90e^{-03}	8.13e^{-03}	3.38e ^{03}	3.57	3.61e^{-04}	***
Agreement	7.39e^{-02}	8.13e^{-03}	3.38e ^{03}	9.09	<2e^{-16}	***
Correct number	1.77e^{-02}	8.14e^{-03}	3.38e ^{03}	2.18	0.029	*

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.7.4 Spill-over 3

At the third word after the target word the effect of surprisal appeared to subside. Using categorical surprisal, the largest random effects structure that yielded converged models was a random intercept only. Model comparison showed that the interaction did not contribute to model fit ($\chi^2(1) = 2.33$; $p = 0.13$), nor did surprisal ($\chi^2(1) = 2.09$; $p = 0.15$). The interaction between number and agreement did not contribute to model fit ($\chi^2(1) = 0.86$; $p = 0.35$), and the main effect of number did not, either ($\chi^2(1) = 0.48$; $p = 0.49$). Agreement did contribute to model fit ($\chi^2(1) = 34.48$; $p < 0.01$). The interpreted model included a main fixed effect of agreement and a random intercept. This model showed an effect of agreement ($\beta = 4.62{\cdot}10^{-2}$, SE $= 7.84{\cdot}10^{-3}$, $t(3390) = 5.89$, $p < 0.01$), indicating that slower reading times after an incorrect target word persists (at least) until the third word after the agreement error. This can be seen in Figure 4.3 in the main text. The results are summarized in table 4.12 below.

Table 4.12: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the third spill-over word. Model: residual log RT ~ agreement + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-2.86e^{-03}	7.77e^{-03}	1.57e ^{02}	-0.368	0.713	
Agreement	4.62e^{-02}	7.84e^{-03}	3.39e ^{03}	5.887	4.32e^{-09}	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.8 Appendix III. Results of the analysis using the number on the noun

The analyses described below follow the same structure as those in the main text with the only difference being that the 'number' factor refers to the number of the subject noun rather than the number of the verb. Notice that these two factors could not be included in the same model with an 'agreement' factor, since congruent number factors would indicate correct agreement, and incongruent number factors would capture incorrect agreement. The surprisal factor used is continuous.

4.8.1 Target word

The largest random effects structure that converged and was non-singular had random slopes for agreement. Model comparison revealed that the interaction between surprisal and agreement did not contribute to model fit ($\chi^2(1) = 0.051$; $p = 0.82$), but the main effect of surprisal did ($\chi^2(1) = 11.55$; $p < 0.01$). The interaction between agreement and the number of the noun did not contribute to model fit ($\chi^2(1) = 1.60$; $p = 0.21$), but the number of the noun itself did ($\chi^2(1) = 21.27$; $p < 0.01$). Agreement did not contribute to model fit ($\chi^2(1) < 0.01$; $p = 0.99$). The interpreted model contained main effects of surprisal and the number of the noun. The model output is displayed in table 4.13 below.

Table 4.13: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the target word. Model: residual log RT ~ surprisal + noun_number + (1 + agreement | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-1.21e^{-01}	1.53e^{-02}	1.38e ^{03}	-7.94	4.17e^{-15}	***
Surprisal	3.15e^{-03}	9.23e^{-04}	3.29e ^{03}	3.41	6.55e^{-04}	***
Noun_number	3.75e^{-02}	8.10e^{-03}	3.31e ^{03}	4.63	3.90^{e-06}	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.8.2 Spill-over 1

The largest random effect structure that converged and was non-singular had random slopes for surprisal and the number of the noun. In addition, three other models converged and were non-singular, namely a model with random slopes for surprisal, a model with random slopes for agreement, and a random intercept only. Comparison of the AIC values revealed that the model with random

slopes for agreement was best. Model comparison then revealed a marginal interaction between surprisal and agreement ($\chi^2(1) = 3.41$; $p = 0.065$). Further reduction from the model that included this interaction showed that the interaction between the number of the noun and agreement did not contribute to model fit ($\chi^2(1) = 0.67$; $p = 0.41$). The interpreted model contained an interaction between agreement and surprisal and main effects of agreement, surprisal, and the number of the noun. The model output is shown in table 4.14 below.

Table 4.14: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the first spill-over word. Model: residual log RT ~ agreement + surprisal + noun_number + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-8.79e^{-02}	1.93e^{-02}	2.39e ^{03}	-4.555	5.50e^{-06}	***
Agreement	1.25e^{-01}	3.19e^{-02}	1.98e ^{03}	3.918	9.22e^{-05}	***
Surprisal	3.67e^{-03}	1.33e^{-03}	3.31e ^{03}	2.757	5.86e^{-03}	**
Noun_number	2.62e^{-02}	8.70e^{-03}	3.30e ^{03}	3.014	2.60e^{-03}	**
Agreement * surprisal	-3.78e^{-03}	2.06e^{-03}	3.31e ^{03}	-1.840	0.066	.

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.8.3 Spill-over 2

The two models that converged and were non-singular were the one with random slopes for surprisal and agreement, and the one with a random intercept only. We initially performed model comparison with the large random slopes model, but this led to singular fit in the other models. We therefore continued with a random intercept only. Here, model comparison revealed that the interaction between agreement and surprisal did not contribute to model fit ($\chi^2(1) = 0.14$; $p = 0.71$). The main effect of surprisal did ($\chi^2(1) = 23.19$; $p < 0.01$), as did the interaction between the number of the noun and agreement ($\chi^2(1) = 6.11$, $p < 0.05$). The interpreted model contained the interaction between agreement and the number of the noun, as well as main effects of surprisal, agreement, and the number of the noun. The model output is provided in table 4.15 below.

To investigate the interaction between agreement and the number on the noun, we looked into simple effects. This revealed that the effect of agreement was significant for both singular and plural nouns (singular: $F(1,3380.66) = 52.33$, $p < 0.01$; plural: $F(1,3380.34) = 14.88$; $p < 0.01$), but the effect of the number of the noun was marginally significant after correction in ungrammati-

Table 4.15: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the second spill-over word. Model: residual log RT ~ agreement * noun_number + surprisal + (1 | participant).

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-8.67e^{-02}	1.58e^{-02}	1.96e^{03}	-5.49	4.60e^{-08}	***
Surprisal	4.56e^{-03}	9.45e^{-04}	3.38e^{03}	4.82	1.47e^{-06}	***
Agreement	8.40e^{-02}	1.16e^{-02}	3.38e^{03}	7.24	5.61e^{-13}	***
Noun number	1.18e^{-02}	1.15e^{-02}	3.38e^{03}	1.03	0.30	
Agreement * noun_number	-4.03e^{-02}	1.63e^{-02}	3.38e^{03}	-2.47	0.01	*

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

cal sentences only ($F(1,3383.61) = 6.06$, $p = 0.056$) with a total absence of the effect in grammatical sentences ($F(1,3383.21) = 1.06$; $p = 1$).

4.8.4 Spill-over 3

The only model that converged and was non-singular was the model with a random intercept only. Model comparison revealed that the interaction between agreement and surprisal did not contribute to model fit ($\chi^2(1) = 0.57$; $p = 0.45$); neither did the main effect of surprisal ($\chi^2(1) = 1.53$; $p = 0.22$), the interaction between agreement and number($\chi^2(1) = 0.49$; $p = 0.49$), and the main effect of number ($\chi^2(1) = 0.85$; $p = 0.36$). The main effect of agreement did contribute to model fit ($\chi^2(1) = 34.48$; $p < 0.01$). The interpreted model contained a main effect of agreement. The output is shown in table 4.16 below.

Table 4.16: The fixed effects from the interpreted linear mixed effects model of the residual log reading times at the third and final spill-over word. Model: residual log RT ~ agreement + (1 | participant)

	Estimate	Std. Error	df	t value	p value	
(Intercept)	-2.86e^{-03}	7.77e^{-03}	1.57e^{05}	-0.37	0.71	
Agreement	4.62e^{-02}	7.84e^{-03}	3.39e^{06}	5.89	4.32e^{-09}	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

5 | Lexical surprisal shapes the time course of syntactic structure building[1]

Abstract

When we understand language, we recognize words and combine them into sentences. How do we do this? In this Chapter, we explore the hypothesis that listeners use probabilistic information about words to build syntactic structure. Recent work has shown that lexical probability and syntactic structure both modulate the delta-band (0-4 Hz) neural signal. Here, we investigated whether the neural encoding of syntactic structure changes as a function of the distributional properties of a word. To this end, we analyzed MEG data of 24 native speakers of Dutch who listened to three fairytales with a total duration of 49 minutes. Using temporal response functions and a cumulative model-comparison approach, we evaluated the contributions of syntactic and distributional features to the variance in the delta-band neural signal. This revealed that lexical surprisal values (a distributional feature), as well as bottom-up node counts (a syntactic feature) positively contributed to the model of the delta-band neural signal. Subsequently, we compared responses to the syntactic feature between words with high- and low surprisal values. This revealed a delay in the response to the syntactic feature as a consequence of the surprisal value of the word: high surprisal values were associated with a delayed response to the syntactic feature by 150 to 190 milliseconds. The delay was not affected by word duration, and did not have a lexical origin. These findings suggest that the brain uses probabilistic information to infer syntactic structure, and highlight an importance for the role of *time* in this process.

[1]Adapted from Slaats, S., Meyer, A. S., & Martin, A. E. (in press). Lexical surprisal shapes the time course of syntactic structure building. *Neurobiology of Language*.

5.1 Introduction

In order to understand language, we must recognize words and combine them into larger linguistic units like phrases and sentences. This process is complicated by the fact that as the sensory input unfolds, be it speech, sign, or text, we must settle on an interpretation of the input (viz., perception and recognition) in addition to transforming or combining that input into larger meaning units. At least two types of information can help us in this process, knowledge about what we are perceiving (e.g., which linguistic unit, how that unit fits with others) and knowledge about how likely it is to occur. These two types of information can be roughly described as the structure of language and knowledge of its statistical distribution. Over the past several decades, much psycholinguistic research has focused on accounting for syntactic phenomena either as a form of transitional probabilities between different linguistic units (e.g., Frank & Bod, 2011; Frank & Christiansen, 2018; Frost et al., 2019; McCauley & Christiansen, 2019), or as a separate level of representation that is hierarchically structured and abstracts away from the lexical items itself (e.g., Brennan & Hale, 2019; Lo et al., 2022; Matchin & Hickok, 2020), without much integration between the two types of knowledge. Nevertheless, recent work in psycho- and neurolinguistics has provided evidence that both types matter (Maheu, Meyniel, & Dehaene, 2022; Nelson, El Karoui, et al., 2017; Weissbart & Martin, 2023). We know from perception and cognition that brains, both human and non-human, are incredible probabilistic engines (e.g., Santolin & Saffran, 2018), and that they are capable of producing abstract, generalizable representations (e.g., Coopmans, Kaushik, & Martin, 2023; Deacon, 1997; Doumas, Hummel, & Sandhofer, 2008; Martin & Doumas, 2019a). Here, therefore, we test a framework where humans *use* lexical distributional information to build abstract, hierarchical representations that give rise to meaning. It is an instantiation of cue integration (viz. Ernst & Bülthoff, 2004; Marslen-Wilson & Tyler, 2007; Martin, 2016, 2020): word-by-word statistics are cues for linguistic rules.

5.1.1 Statistical patterns in learning and comprehension

Psycholinguistic experiments have taught us that humans are capable of extracting statistical regularities very quickly: infants and adults alike are able to extract words, simple rules, and even non-adjacent dependencies from a continuous stream of syllables after as little as two minutes of exposure using nothing more than transitional probabilities (Aslin et al., 1998; Batterink & Paller, 2017, 2019;

Gervain, 2014; Gómez, 2002; Isbilen, Frost, Monaghan, & Christiansen, 2022; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Vouloumanos & Werker, 2009). It has since been hypothesized that this capacity underlies our extraction of syntactic rules: we use distributional cues to infer the structure underlying the input (Rowland, Chang, Ambridge, Pine, & Lieven, 2012; Saffran, 2001; Thompson & Newport, 2007). Early modeling work revealed exactly these statistical patterns that language follows are a direct consequence of the syntactic structure of the input (e.g., Elman, 1991, 1993). More recently, corpus studies and computational models suggest that (backward) surprisal contains information about the phrase structure of sentences (McCauley & Christiansen, 2019).

These findings culminated in models using those statistical patterns not just in a theory of language acquisition, but also in a theory of language comprehension. An influential example of such a theory is surprisal theory (Hale, 2001, 2006, 2016; Levy, 2008a, 2008b; Levy & Gibson, 2013). Surprisal theory broadly aims to predict when comprehension difficulties arise. The underlying assumption is that comprehenders make use of probabilistic information to predict both the structure of the input they have just heard or seen, and what they might encounter next. The extent to which these predictions are correct is hypothesized to determine the difficulty of processing. The model uses surprisal, the negative log probability of a word (or other linguistic unit) given the context, as a quantification of the validity of the predictions made. If surprisal is high on a given word, this word was unexpected given the context, and processing difficulty (often indexed by slower reaction times in, for example, self-paced reading tasks) is predicted to occur. Since surprisal can be calculated over any representation, be it phonemic, lexical, or even structural, surprisal theory does not commit to a representation of language. It is agnostic about the representations and mechanisms that lead to structure-dependent interpretation (see Slaats & Martin, 2023 and Chapter 2).

Since the introduction of surprisal theory, distributional information has been shown to account for much variance in models of behavior and neural activity *after* learning as well (Armeni et al., 2019; Brennan & Hale, 2019; Hale, 2006, 2016; Hasson, 2017; Hasson & Tremblay, 2015; Heilbron et al., 2022; Levy & Gibson, 2013; Smith & Levy, 2013), a trend that continues rapidly with the introduction of large language models. For example, higher surprisal values and a larger decrease in entropy are both associated with slower reading times (Aurnhammer & Frank, 2019; Frank, 2013; Linzen & Jaeger, 2016) – the process-

ing difficulty from surprisal theory. More recently, neuroimaging experiments have shown that oscillations at delta-, beta- and gamma bands track surprisal (Weissbart et al., 2019), entropy reduction correlates with temporal lobe activity (Nelson, Dehaene, et al., 2017), and that surprisal and word frequency are tracked over and above acoustic and speech segmentation representations (Gillis et al., 2021). In other words, probability at the word level is a good predictor for behavioral and neurophysiological measurements (Slaats et al., 2023).

But the power of distributional information does not stop there: some effects that are attributed to linguistic structure can be evoked by statistical regularities as well. In a seminal work, Ding and colleagues (2016) showed that the rate of occurrence of linguistic structures (syllables, phrases and sentences) is reflected in power in the neural signal at the corresponding frequencies (4Hz, 2Hz, and 1Hz, respectively). This effect was widely adopted in the literature as reflecting the construction of linguistic units: the brain encodes abstract linguistic information. However, since its publication several studies have shown that the low-frequency frequency tagging effects can be induced by transitional probability information alone (Bai, 2022; Batterink & Paller, 2017).

This overwhelming evidence for the importance of distributional information has reignited the debate on whether language acquisition and language comprehension alike are both rooted in sequential, statistical information (Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018; Frank & Yang, 2018), rather than the hierarchical tree structures that are part of linguistic theory (Chomsky, 1956, 1965; Everaert et al., 2015; Pollock, 1989; Rizzi, 1997, i.a.). This is quite the departure from the early hypothesis in the statistical learning literature that statistics function as a *cue* rather than the instantiation of the structure itself.

5.1.2 Syntactic structure in neural dynamics

Logically, however, statistics-only accounts struggle to explain language behavior (Fodor & Pylyshyn, 1988; Hale et al., 2022; Martin, 2016, 2020; Slaats & Martin, 2023). For one, listeners are able to understand sentences that include words or combinations of words that they never encountered before. In line with this observation, there is ample evidence for the use of abstract structure in language learning and comprehension. For example, learners privilege abstract knowledge of scope-taking over transitional probabilities when presented with a structurally altered version of English (Culbertson & Adger, 2014), and are able to infer abstract structure in the input after a single exposure (Marcus, Vijayan,

Bandi Rao, & Vishton, 1999). Beyond that, everyday language production shows that both children and adults produce utterances that they have not heard before (Conwell & Demuth, 2007; Valian, 1986).

And indeed, neuroimaging studies have shown repeatedly that inferred structural information modulates activity in low frequency bands, particularly the delta (<4Hz) band (Bai et al., 2022; Brennan & Martin, 2020; Kaufeld, Bosker, et al., 2020; Lo et al., 2022; Meyer et al., 2017; Tavano et al., 2022; Ten Oever, Carta, et al., 2022) and gamma band (Nelson, El Karoui, et al., 2017; Peña & Melloni, 2012) – even when there are no acoustic markers of this structural information. For example, Bai and colleagues (2022) presented participants with two different structures: phrases (*de blauwe bal,* 'the blue ball') and sentences (*de bal is blauw*, 'the ball is blue'). These two types of stimuli had the same number of syllables and indistinguishable power spectra, but the neural response differed between the conditions in various ways: low-frequency (1-8 Hz) phase coherence, <2Hz phase connectivity, and theta(4-10Hz)-beta(15-40Hz) phase-amplitude coupling. These findings suggest that even small changes of syntactic structure have large consequences for the (low-frequency) neural dynamics. Similarly, Tavano and colleagues (2022) show that those syntactic categories, phrases and sentences, generate a neural rhythm as reflected in inter-trial phase coherence that is mathematically independent of the presentation rate of the words.

Evidence for delta-band involvement in the process of structure building also comes from studies comparing word lists and sentences (Lo et al., 2022; Lu, Jin, Pan, & Ding, 2022; Slaats et al., 2023). Lu and colleagues (2022) presented participants with sentences and word lists of animate and inanimate nouns that both repeated at 1Hz to assess whether delta-band dynamics track semantic properties of words, or whether the changes are related to structural properties of the stimulus. A lexical distributional approach as those advocated by Frank and colleagues (Frank & Bod, 2011; Frank et al., 2012; Frank & Christiansen, 2018; Frank & Yang, 2018) would predict stronger 1- and 2Hz response peaks in the word list condition than in the sentence condition; the opposite is predicted by model that assumes a role for syntactic structure in delta-band activity. The study showed that the 1Hz response peak was larger for sentences than for word lists, suggesting again that low-frequency activity is modulated by or causal for structure building. In a similar vein, Lo and colleagues (2022) showed that synchronization to the sentential rhythm in the delta band only occurs when the sentences are syntactically well-formed. Finally, Slaats and colleagues (2023)

compared delta-band responses to individual words between word lists and sentences, while controlling for effects of surprisal. This study showed that responses to words were affected in their temporal and spatial organization when embedded in a sentence structure: the responses appeared earlier and activity was propagated to left inferior frontal areas in the sentence condition only.

5.1.3 A time and space for both

Some studies pit the importance of structure or sequential probabilities against each other (Brennan & Hale, 2019; Christiansen & Chater, 2015; Frank et al., 2012; Frank & Yang, 2018). Frank and Bod (2011), for example, used probabilistic language models that were trained to predict the next part-of-speech (POS) with a hierarchical and sequential architecture to model reading time data. They found that the hierarchical models did not account for variance over and above sequential probability estimates, and suggested that human sentence processing relies more on sequential than on hierarchical structure. Brennan and Hale (2019), on the other hand, suggest that hierarchical structure is important during language comprehension. They use several sequential models and a context-free grammar to obtain surprisal for part-of-speech and use those to model EEG data from naturalistic listening. In contrast to Frank and Bod (2011), they find that the context-free grammar estimates predict EEG data over and above the sequential models.

When we consider all of these findings together, we must conclude that both distributional and abstract, hierarchical syntactic information play a role in language comprehension – and that both shape the neural signal. Indeed, a study by Roark and colleagues (2009) showed that models with separate lexical and syntactic surprisal/entropy features are better at modelling RT data than models that do not make this distinction. Similarly, Nelson, El Karoui, et al. (2017) found that intracranial EEG signals differentially encode responses to probabilistic and syntactic information. In line with these findings, several (statistical) learning experiments suggest that the brain represents the statistical biases as well as abstract rules (Maheu et al., 2022; Monte-Ordoño & Toro, 2017; Saffran, 2001; Toro et al., 2011). A model in which probability plays a role, while structure does too, is in line with one of the brain's main features: it can map probabilistic information onto deterministic representations (a relatively undisputed example is categorical perception; Harnad, 2003) (Martin, 2020; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Given this background, rather than contrast distributional information with syntactic information, we investigate how these factors *jointly* shape the neural signal. As in the statistical learning literature (e.g. Saffran, 2001; Thompson & Newport, 2007), we ask if distributional information can serve as a *cue* for syntactic structure during comprehension: contextual lexical distributional information should affect the quality of the neural signature of structure building. Lexical probability thus should interact with abstract representations of sentence structure. We ask (1) whether the neural encoding of linguistic structure changes as a function of the distributional properties of a word, and (2) whether this influence can be linked to probabilities in the immediate context (two preceding words) or rather to probabilities in the larger context. Following findings from statistical learning and models of transitional probabilities (McCauley & Christiansen, 2019; Thompson & Newport, 2007) and in line with a model of language comprehension proposed by Martin (2016; 2020) we hypothesize that the neural encoding of linguistic structure is affected by the probability of a word in the context.

While this may appear a relatively straightforward question to answer, several issues of both methodological and theoretical origin arise. The first issue concerns the operationalization of contrasting a distributional factor with a latent structural one. Problems arise because of the nature of lexical distributional information such as surprisal and entropy (discussed in depth in Chapter 2): these values are affected by of *any* change that is made in the underlying structure. If one wants to manipulate the latent syntactic structure underlying a sentence, the surprisal values of the words will also change. Because of this issue, the potential effects at hand are not easily captured in a factorial design. We solve this problem by making use of variance of lexical distributional information and syntactic structures that occurs naturally in continuous speech: we use temporal response functions (TRFs) to model responses to latent linguistic variables, such as syntactic structure and lexical surprisal, in MEG data obtained using a naturalistic listening paradigm.

The second issue concerns the broader theoretical questions concerning the nature of distributional information in the brain: over which representation does the brain store distributional information, and how is this implemented mechanistically? While this particular study will not provide an answer to this larger question, it will speak to two questions that follow from it. Specifically, (1) which type of distributional information most faithfully captures the information available to the brain, and (2) which of those plays a role in the process

of the inference of latent linguistic structure? This question concerns estimates derived from large language models like the GPT-family, those derived from simpler models like long-short term memory networks, or the even simpler trigram models. In this study we look into a version of GPT2 and a trigram model.

Separate predictions may be derived with regards to the two questions posed above. With respect to question (1), one can expect models with a larger context and potentially enhanced sensitivity to the latent factors driving the statistical patterns to perform better when it comes to describing the neural signal generally as a consequence of capturing more sources of variance (and indeed, this appears to be true (Heilbron, Ehinger, Hagoort, & de Lange, 2019) until a certain point (Kuribayashi, Oseki, Brassard, & Inui, 2022); (see also Futrell et al., 2020)): a larger context and more parameters theoretically allow for capturing more fine-grained sources of variance. Predictions concerning question (2), on the other hand, are not so easy to derive. While we hypothesize generally that lexical distributional information can affect the process of syntactic structure building as described above, it is unknown whether one would need long-context, fine-grained variability to capture this hypothesized effect, or whether the local context provides enough distributional information to capture it. The statistical learning literature suggests that short-context probability, such as bigram and trigram frequencies, can function as a cue for linguistic structure (Aslin & Newport, 2012; Aslin et al., 1998; Frost et al., 2019; Gómez, 2002; Isbilen et al., 2022; Knowlton & Squire, 1996; McCauley & Christiansen, 2019; Thompson & Newport, 2007). For this reason, we hypothesize that any effect of lexical distributional information on the inference of syntactic should be observable using a short-context metric such as trigram probability.

In summary, in the present study, we address the following questions: (1) whether the neural encoding of linguistic structure changes as a function of the distributional properties of a word, and (2) whether this influence can be linked to probabilities in the immediate context (two preceding words) or rather to probabilities in the larger context. We do this by analyzing MEG data of participants who listened to fairytales. Using TRFs, we model the neural signatures of syntactic structure building in the delta band, and compare those between different distributional contexts (i.e., high versus low surprisal). In order to characterize the lexical distributional information that is available to the brain, we estimate lexical distributional information with two different language models: a trigram model, which uses only two words to estimate the predictability

of the current word, and a Dutch version of GPT2, a large transformers model that uses a very large context window.

5.2 Methods

5.2.1 Participants

24 right-handed native speakers of Dutch (18 female, 20-58 years old (mean = 33.4)) were recruited from the participant pool at Radboud University Nijmegen, the Netherlands. All participants reported normal hearing, had normal or corrected-to-normal vision, and reported no history of language-related impairments. Participants gave written informed consent. The experiment was approved by the Ethics Committee for human research Arnhem/Nijmegen (project number CMO2014/288).

5.2.2 Materials

The stimuli consisted of three fairytales (one by Hans Christian Andersen, two by the Brothers Grimm) read out at comfortable pace by female native speakers of Dutch. Each story was divided into segments of approximately 5.5 minutes (range 4:58 – 6:40), leading to 9 segments and a total duration of 49 minutes and 17 seconds. Each segment was normalized for loudness using the FFmpeg software (EBU R128 standard). The transcripts of the stories were checked for consistency with the recordings, adjusted for spelling where necessary and subsequently automatically aligned with the audio using the WebMAUS segmentation software to extract word onset time-points (Kisler, Reichel, & Schiel, 2017). All stories contained a natural variation of words and sentence structures. In total, the stories contained a total of 8551 words in 791 sentences, with an average length of 10.8 words (range 1-35, sd. 5.95).

5.2.3 Procedure and data acquisition

Participants were tested individually in a magnetically shielded room. They were instructed to sit still and look at a fixation cross that was presented in the middle of a screen while they listened passively to the fairytales. Each block started with a 10-second period during which resting state data were recorded. After each story segment, five multiple-choice comprehension questions were asked. On average, participants' accuracy was 88.1% (sd. 7.52%), indicating that they

were paying attention to the content of the stories. The stimuli were presented via plastic tubes and ear pieces to both ears. The experiment was run using Psychtoolbox in Matlab (Brainard, 1997).

MEG data were recorded continuously with a 275-channel axial gradiometer (CTF) system at a sampling frequency of 1200Hz. Three head localizer coils were attached to the participant's head (nasion, left- and right ear canals through fitted ear molds) to determine the position of the head relative to the MEG sensors. The head position was monitored throughout measurement and, if necessary, corrected during breaks. In addition, eye movements and heartbeat were recorded with additional EOG and ECG electrodes.

5.2.4 MEG preprocessing

Preprocessing was done with MNE-Python (version 0.23.1, Gramfort et al., 2013). The MEG data were down sampled to 600 Hz and band-pass filtered at 0.5-40 Hz using a one-pass zero-phase, non-causal FIR filter. We interpolated bad channels using Maxwell filtering, and used ICA to eliminate artifacts resulting from eye movements (EOG) and heartbeats (ECG). The data were segmented into nine epochs time-locked to the onset and offset of the story audio recordings. At some point in data collection, some channels of the scanner failed due to technical issues. We interpolated these channels for those participants to ensure the same number of channels for all participants. In continuation, the epochs were resampled to 200Hz and band-pass filtered between 0.5 and 4 Hz (the delta band).

5.2.5 Temporal response functions

We modeled the neural signal using temporal response functions (TRFs) with different acoustic and linguistic features. This approach has been used to distinguish between responses to different linguistic features, ranging from the speech envelope and phonemic information (Di Liberto et al., 2015; Donhauser & Baillet, 2020; Tezcan et al., 2023), to lexical information (Slaats et al., 2023; Weissbart et al., 2019) and even syntactic embedding (Nelson, El Karoui, et al., 2017). In essence, the method is a multivariate multiple linear regression, where we used lagged time series of different annotations of the stimulus as features. In this way, it is possible to distinguish between variability in the signal that stems from acoustic processing, lexical processing, and many others.

The equation of the model reads as follows:

$$y_c(t) = \sum\sum x_f(t)\beta_f(t-\tau_k) + \eta(t) \tag{5.1}$$

Where $\{y_c\}_t, \{x_f\}_t, \{\beta_f\}_t$ represent the recorded MEG signal of a given channel c, the input feature f and its temporal response function, respectively. $\{\eta\}_t$ is a gaussian noise process which accounts for aspects of the stimulus that are not captured by the coefficients attributed to the features in the model. We solved this equation using ridge regression (as opposed to, for example, boosting; Brodbeck et al., 2023). This means that we estimated the coefficients of the TRFs $\hat{\beta}_f$ by minimizing the squared error between the measured MEG signals and the reconstructed signal obtained from equation (1) while keeping the norm of the TRFs coefficients$||\beta||_2$ low to avoid overfitting. This minimization problem is solved in a closed form by:

$$\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T Y \tag{5.2}$$

Where $Y \in \mathbb{R}^{N\times C}$ is the matrix representation of the measured MEG signal (for C channels arranged column-wise, each with N data samples); $\hat{\beta} \in \mathbb{R}^{(K.F)\times C}$ contains the estimated TRFs with K lags, F features for all C channels; $X \in \mathbb{R}^{N\times(K.F)}$ is a matrix containing all lagged feature time series of length N; λ is a regularization coefficient and I_d the identity matrix. The regularisation coefficient is needed to avoid overfitting, which in this case translates to the square matrix $X^T X$ not being full rank. Numerically small eigenvalues or simply ill-conditioned matrices can make the inversion unstable and therefore require regularization. In TRF-models, this happens when features present some amount of autocorrelation, as is the case in our models (e.g., the acoustic envelope is strongly autocorrelated).

In equation (1), the vector of weights $\beta_f(t)$ represents the coefficients parameterizing the temporal response functions. They form a time course reminiscent of an event related potential that tells us at which point in time (and, potentially, where) a feature modulates the neural signal. Thus, an increase at a certain lag for a given feature reflects an increase in the associated brain response to this feature at that given sensor and at the given time lag after stimulus onset.

To evaluate how our models perform at reconstructing the neural data, we computed the Pearson's correlation coefficient between the true data and data reconstructed using the estimated TRFs. The correlation between the reconstruction and the original MEG indicates how much of the variance in the neural signal is explained by the features. The TRFs were not estimated on the same

portion of data used to score the model. As further explained in section 5.2.7 "Model fitting & Statistical analysis", we used a nested cross-validation procedure to tune the regularization parameter, estimate the TRF coefficients and finally score the resulting model. Unless specified otherwise, all analyses described below were done with custom made Python scripts using MNE-Python (Gramfort et al., 2013).

5.2.6 Stimulus representations

To characterize the speech signal and latent linguistic features, we constructed eight features that belong either to the base features or to the set of experimental features. The base features are present in every model, and are used to remove variance from factors that could potentially influence the results.

Base features The base features are *speech envelope*, *word onset*, and *word frequency*.

The *speech envelope feature* was computed for each stimulus by taking the absolute value of the Hilbert transform and down sampling it to 200 Hz to match the MEG sampling rate. The envelope feature was added to represent the acoustic response and as such separate acoustic processing from linguistic processes of interest: structure building.

The *word onset feature* was added to capture broadly any time-locked response to word onset for which the variance is not already explained by other features. To this end, we extracted word onset time-points using the WebMAUS segmentation software (Kisler et al., 2017). We used a train of unit impulses, where the feature signal is one at the word onset sample and zero otherwise:

$$x(t) = \sum_{words} \delta\left(t - t_{onset}\right) \tag{5.3}$$

These impulse trains were convolved with a Gaussian kernel with a standard deviation of 15ms. Such temporal smoothing has the effect of inflating the autocorrelation of the signal. We designed the width of this smoothing such that the smoothed impulses end up with energy spanning a comparable frequency band as to our continuous regressor (the speech envelope). The Fourier Transform of a gaussian is also a gaussian, and the 15ms standard deviation of the temporal smoothing kernel equates to a spectral standard deviation of 21.22Hz. This ensured that all features required a similar degree of regularization in the regression analysis, and made it possible to include impulse-like features such

as word onsets and the envelope in the same regularized regression. Notably, this also translates into some uncertainty about or knowledge of the exact word onset timings.

Like the word onset feature, the *word frequency feature* was constructed as an impulse train of zeros everywhere but at word onset. Here we used the respective word frequency value to modulate the height of the impulses. We used the log-transformed value of occurrence per million words, obtained from the SUBTLEX-NL corpus (Keuleers et al., 2010):

$$x_{wf}(t) = \sum_{words} -\log(p(w)) \times \delta(t - t_{onset}) \tag{5.4}$$

where P(w) represents the unigram probability estimated from occurrence per million words. If a word did not exist in the corpus, the fallback value of 0.301 (log/million) was used, corresponding to the lowest word frequency in the corpus. The values were scaled (divided by their standard deviation) across all stimuli. The resulting signal was convolved with the same Gaussian kernel as the word onset feature.

Experimental features We designed five experimental features to investigate the influence of contextual lexical distributional measures on structure building: *surprisal* and *entropy,* the distributional features; and *top-down, bottom-up* and *left-corner* node counts, the structural features.

The *surprisal feature* reflects how predictable a given word is in its context. It is the (traditionally two-based) log-transformation of the conditional probability of a word. If surprisal is low, the word was predictable given the context; if it is high, the word was not predictable given the context. See the equation in 5.5.

$$I(w_i|w_{i-n}...w_{i-1}) = -\log_{10}(p(w_i|w_{i-n}...w_{i-1})) \tag{5.5}$$

The *entropy feature* consists of lexical entropy, a weighted probability measure that quantifies the uncertainty about the upcoming word on the basis of the previous words. It provides a numeric answer to the following question: given the *n* previous words, with what degree of certainty can we predict the upcoming word? See the equation in 5.6.

$$H(w_i|w_{i-n}...w_{i-1}) = -\sum p(w_i|w_{i-n}...w_{i-1}) \log(p(w_i|w_{i-n}...w_{i-1})) \tag{5.6}$$

We generated these two metrics in two ways. For the long-context distributional information models, the values were derived from GPT2, a large-scale transformers language model that was fine-tuned for Dutch (de Vries & Nissim, 2021). This version of GPT2 was fine-tuned using a context window of 128 tokens. Tokens do not map onto words in a one-to-one fashion. Instead, a token can correspond to a word, but also to a morphological marker; for example, the Dutch plural 's' marker may be a token. As such, the context used for the surprisal and entropy estimates will roughly correspond to a little under 128 words. For the short-context distributional information models, the values were obtained from a trigram model created with SRILM (Stolcke, 2002) trained on ∼1.2M words from the Dutch corpus from OpenSubtitles (Lison & Tiedemann, 2016). This model takes the preceding two words to estimate the surprisal (and entropy) values of the target word. We used Kneser-Ney discounting with interpolation to estimate values for missing words or trigrams.

To extract neural signatures of structure building, we needed a feature that reflected the syntactic structure underlying the input. To this end, we manually parsed all sentences using a simplified version of X-bar theory (Carnie, 2013). This entailed that we created a full X-bar structure for all noun phrases (NPs) and verb phrases (VPs), but not for the other phrases unless intermediate projections were filled. Using the full X-bar structure for NPs and VPs ensured that each parse contained an explicit distinction between arguments and adjuncts, with arguments being attached as a sister of the head and adjuncts occupying the intermediate projection. An example of one of the resulting parses is displayed in Figure 5.1.

From these parses we extracted node count estimates to function as the syntactic features in our TRF models. Node counts have been found to effectively represent syntactic complexity in the neural signal (Brennan et al., 2016; Giglio, Ostarek, Sharoh, & Hagoort, 2024; Li & Hale, 2019; Nelson, El Karoui, et al., 2017). Node counts can be computed in different ways, depending on the algorithm the parser is hypothesized to use to reach the structured representation. We calculated node counts according to three algorithms: a top-down algorithm (further: top-down), a bottom-up algorithm (further: bottom-up) and a left-corner algorithm (further: left-corner). The top-down parsing method is maximally predictive. Upon encountering a word, all nodes governing this word to the right are assumed to be built. For example, if the parser encounters the determiner 'the' in the sentence 'the train arrived' (see tree structure in Figure 5.2), the parser will build not only the determiner, but also the NP and the VP,

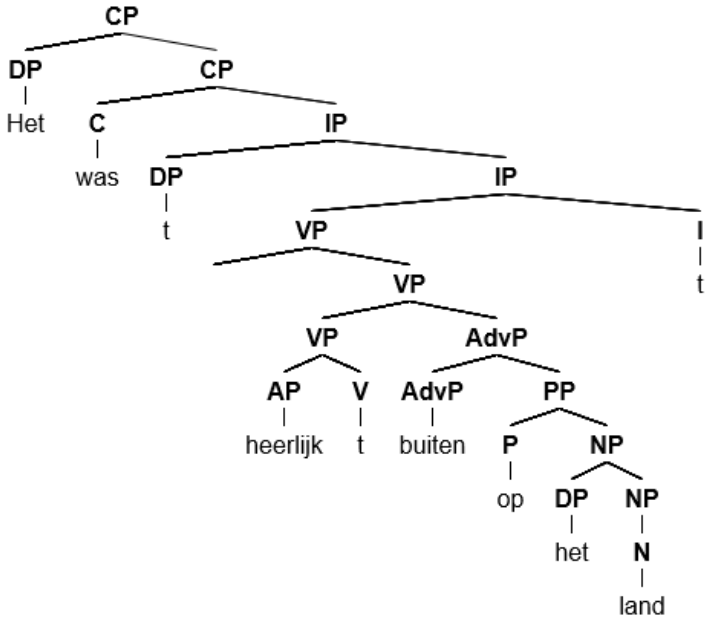
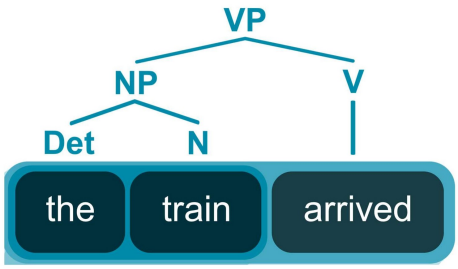


Figure 5.1: A parsed sentence from the stimuli. The sentences were parsed according to an adapted minimalist paradigm. The sentence reads (non-literally): 'the weather was lovely on the countryside'. This is the first sentence of one of the stories by Anderson.

hence a node-count of 3. The bottom-up method is completely non-predictive: in this method, the parser will build only the nodes it has seen all evidence for. That means that the NP from our example will not be built until the noun 'train' has been seen. The left-corner algorithm is a mixture of these two. This mildly predictive parsing method will project a constituent as soon as the first item is found, but no constituents above this are built. In the case of the train, this means that the NP is built when the determiner has been seen, but the VP will only be built once the whole NP has been seen.

As can be seen in Figure 5.1 above, the parses we created contained traces of moved elements. These traces do not have acoustic correlates in the signal. In order to represent the structure they are part of, we assigned their node counts to other words within the sentence. Specifically, we added the node count of the trace to the node count of the word following it. This strategy was chosen because we reasoned that the location of these traces can be inferred after their position.

Since the linguistic features (frequency, entropy, surprisal, bottom-up node count, top-down node count, left-corner node count) might be correlated to some extent, we need to assert that the degree of multicollinearity present in our stimulus representation will not hinder the TRF coefficient interpretation. We



Top-down	3	1	1
Bottom-up	1	2	2
Left-corner	2	1	2

Figure 5.2: Node counts per word according to top-down, bottom-up and left-corner parsing algorithms.

checked whether the Variance Inflation Factor (VIF) was below 5 (considered a relatively conservative measure of multicollinearity; Sheather, 2009; Tomaschek et al., 2018). The VIF was computed by correlating the z-scored entropy, surprisal, word frequency, and node count values, and taking the diagonal of the inverted correlation matrix. This was done for all the stimuli. The VIF was higher than 9.7 for left-corner due to high positive correlations with both bottom-up and top-down, we did not include left-corner in our models. After removal of this feature, all VIF-values were lower than 3.5 (see Appendix 5.6 for the correlation matrices of the used features). Like the word frequency feature, the other features were scaled and inserted in a stick function, after which the stick function was convolved with the same Gaussian window.

5.2.7 Model fitting & Statistical analysis

Any TRF analysis has two deliverables: firstly, the TRF (the development of the estimated coefficients across time), which is an ERP-like waveform that captures how the neural signal changes as a function of a feature of interest (in our case, the features of interest are the node-count responses); and, secondly, the reconstruction accuracy. This is a metric of model fit (Pearson's correlation, as explained in section 5.2.5 above). We use the second deliverable, the reconstruction accuracy, to assess whether our used features are relevant for a description of the neural signal, and the first deliverable to evaluate whether syntactic structure building processes are affected by the lexical distributional context.

The current analysis consists of two parts: the "main effects"-analyses, and the "interaction"-analyses. In the main effects-part, we conducted an analysis on the reconstruction accuracy across the whole scalp to assess the contribution of each of the features individually, as well as a comparison between the effects of surprisal and entropy from our different language models (trigram and GPT2). We did this to ensure all the effects that were included in the interaction analysis were relevant for the neural signal, i.e., to ensure that main effects were present before we investigated interactions. In the interaction-part, we conducted analyses on the TRFs from models that consider the interaction between estimates of lexical probability and syntactic structure building. The two parts are described in more detail below.

In the main effects-analyses, we estimated TRF models for all combinations of the features of interest *over and above* a 'null' model that included the envelope, word onset and word frequency features. The models and their features are summarized in Table 5.1 below. All models were fitted with surprisal and entropy features estimated from a trigram model and GPT2.

Table 5.1: The fitted encoding models in the main effects-analyses.

Model name	Envelope	Word onset	Feature Word freq.	Surprisal	Entropy	Bottom-up	Top-down
main_null	×	×	×				
main_surprisal	×	×	×	×			
main_entropy	×	×	×		×		
main_distributional	×	×	×	×	×		
main_topdown	×	×	×				×
main_bottomup	×	×	×			×	
main_topdown_bottomup	×	×	×			×	×
main_surprisal_topdown	×	×	×	×			×
main_surprisal_bottomup	×	×	×	×		×	
main_surprisal_topdown_bottomup	×	×	×	×		×	×
main_entropy_topdown	×	×	×		×		×
main_entropy_bottomup	×	×	×		×	×	
main_entropy_topdown_bottomup	×	×	×		×	×	×
main_distributional_topdown	×	×	×	×	×		×
main_distributional_bottomup	×	×	×	×	×	×	
main_distributional_topdown_bottomup	×	×	×	×	×	×	×

Note. **An × indicates that a feature was included in the model.**

Estimating all feature combinations allowed us to estimate a slope of the given feature *irrespective* of the presence of other features. We did this by averaging the reconstruction accuracies of the resulting model across sensors (i.e., one value per participant per model) and submitting these averages to linear mixed models using the lme4 in R (Bates et al., 2015). These models contained a binomial factor for each of the features of interest (surprisal, entropy, bottom-up, and top-down), indicating whether or not a feature was present in the model.

We estimated large linear mixed effects models in which all factors interacted with each other. A model with a full random effects structure was not possible (because there were not enough observations), so we fit this large model four times with each time three out of four factors in the random effects structure. On each large model we performed model comparison using the *step* function from the *LmerTest* package (Kuznetsova et al., 2017). This function reduces the random- and fixed effects structure of a model in a maximal-to-minimal fashion. We then compared the resulting best models for their AIC value, and report the model with the lowest AIC value below.

Because the effects of every feature may differ across the scalp, we also estimated the slope of every feature by averaging the per-sensor reconstruction accuracy values over all the models that did- or did not include a given feature. For example, to examine the effect of entropy, we averaged per participant, per sensor over all the models that include entropy to obtain one 'with-entropy'-value for every sensor for every participant, and we averaged per participant, per sensor over all the models that do not include entropy to obtain one 'without-entropy'-value for every sensor for every participant (see table 5.1 above). Per feature, that means we obtained two values for every sensor: one with the feature, and one without. We then evaluated any difference between these using a cluster-based permutation test between these values using *permutation_cluster_test* from the MNE-Python library.

Cluster-based permutation tests address the null hypothesis of exchangeability across conditions by a Monte Carlo estimate of the randomization distribution of a cluster-based test statistic, optimizing statistical sensitivity while controlling the false alarm rate. Here, we used the T-statistic as the test statistic. In these tests, we create matrices of all sensors (and, in the case of TRF-waveforms) samples. Then, we compute the difference between two conditions and express it as a T-statistic for each of these data points. The T-values are thresholded at an a priori threshold, and the thresholded T-values are summed across clusters on the basis of spatial (and temporal) adjacency. The significance of the resulting largest cluster's test statistic is compared to a pre-defined number of similarly obtained test statistics, after random permutation of the condition labels. Throughout this study, we permuted the values 10.000 times using a t-test as the test-statistic with a threshold of 1.714 (based on 24 participants).

In the interaction-analyses, we estimated interactions between the features that were chosen on the basis of the main-effects analysis. To foreshadow this, the chosen features were *surprisal* and *bottom-up*. All models contained the

speech envelope, word onsets, word frequency, and surprisal features. To evaluate the effect of lexical surprisal on the process of structure building, we split the bottom-up feature by the median of surprisal (derived from the trigram model or GPT2). Doing so in one model yielded two separable responses (the model 'bottomup_split_surprisal' from Table 5.2 below): a node count TRF for low-surprisal words, and a node count TRF for high-surprisal words. We then compared these resulting TRFs using a cluster-based permutation test implemented as *spatio_temporal_cluster_test* from the MNE-Python library. Any differences between the TRF waveforms can be interpreted as differences in the low-frequency neural readout of structure building between words with low- or high surprisal values.

Table 5.2: The fitted encoding models in the interaction effects-analyses.

Model name	Feature					
	Envelope / Word onset / Word freq.	Surprisal	Bottom-up / high surprisal	Bottom-up / low surprisal	Bottom-up / random 1	Bottom-up / random 2
bottomup_low_surprisal	×	×		×		
bottomup_high_surprisal	×	×	×			
bottomup_split_surprisal	×	×	×	×		
bottomup_split_random	×	×			×	×

Note. **An × indicates that a feature was included in the model.**

Further, to assess the variance explained by the low- versus high-surprisal response to structure building, we fit two additional models: a model with only the bottom-up values for high surprisal words, and a model with only the bottom-up values for low-surprisal words. The reconstruction accuracy values from these models were compared to the model 'main_surprisal' from the main effects analysis: this model is identical to those computed here, except for the presence of (half of) the bottom-up node count feature. The difference in reconstruction accuracy between these models – i.e., the increase in reconstruction accuracy as a result of the addition of the bottom-up node count feature – was subsequently compared between the low- and high surprisal models using a cluster-based permutation test.

In continuation, we wanted to evaluate the reliability of the effects on the TRF waveform (any differences between the low- and high surprisal node count TRFs) using the reconstruction accuracy values. Because dichotomizing the node-count feature on the basis of a continuous variable is likely far from the true interaction in the neural signal (the brain probably does not divide words into low- or high surprisal categories), a direct comparison of the reconstruction

accuracy values from the split feature to an intact feature did not seem a fair comparison. Therefore, we decided to perform evaluation of the effect by comparing the model 'bottomup_split_surprisal' to an equivalent model in which the split was performed randomly (i.e., the words were randomly distributed over two sets).

After obtaining the differences using the cluster-based permutation test and confirming them through the reconstruction accuracy values, we wanted to evaluate whether there was a latency difference between the responses to bottom-up node count for low- or high surprisal words. To do this, we compared the TRFs for bottom-up node count for the low- and high surprisal words in a cross-correlation. This cross-correlation was performed on the grand-average TRF waveforms of the sensors that were part of the significant clusters resulting from the cluster-based permutation test that compared the two responses. In other words, the sensors were the ones that contributed to the significant difference between the two distributions. We sequentially cross-correlated each sensor, and normalized the values by dividing them by the maximal value from the cross-correlation for that sensor. We then obtained the positive peaks for every sensor. The peak corresponds to the "lag" at which the two signals had the highest correlation, and shows how different the two responses are in time. Subsequently, we took the most frequently occurring peak value, and shifted one of the two TRF waveforms to match the other one, and computed the correlation. To check for significance, the same procedure was repeated for randomly selected channels and time-lags 10.000 times.

5.3 Results

5.3.1 Main effects: Whole-brain averages

The model comparison approach on the whole-brain average reconstruction accuracies of the *trigram models* (shown left in Figure 5.3) showed that a model with several interactions between the factors surprisal, entropy, top-down and bottom up was the best descriptor of the data. Specifically, there were interactions between entropy and surprisal, entropy and top-down, surprisal and top-down, and the two syntactic features. The model formula is shown in Equation (5.7) below. Full model-comparison statistics are provided in Appendix 5.7.

$$\begin{aligned} accuracies \sim & entropy + surprisal + topdown + bottomup \\ & + entropy \cdot surprisal + entropy \cdot topdown \\ & + surprisal \cdot topdown + topdown \cdot bottomup \\ & + (1 + topdown \cdot bottomup \cdot surprisal | subject) \end{aligned} \tag{5.7}$$

The results of this model showed a significant negative effect of entropy (β = -6.81·10^{-4}, SE = 5.37· 10^{-5}, t(237) = -12.69, p < 0.01), indicating that entropy decreased the reconstruction accuracy of the signal. There was a further negative effect of top-down (β = -8.88· 10^{-4}, SE = 1.51·10^{-4}, t(25.23) = -5.87, p < 0.01), similarly suggesting that this feature decreased the reconstruction accuracy of the signal. Bottom-up, on the other hand, had a positive effect on the reconstruction accuracy (β = 1.03·10^{-3}, SE = 3.25·10^{-4}, t(22.99) = 3.19, p < 0.01), as did surprisal (β = 3.65· 10^{-4}, SE = 1.31·10^{-4}, t(26.24) = 2.79, p < 0.01). In addition, there were several interactions between features. There was an interaction between entropy and surprisal (β = 5.02·10^{-4}, SE = 6.20·10^{-5}, t(237) = 8.10, p < 0.01) and between top-down and all the other features: entropy (β = 1.49·10^{-4}, SE = 6.20· 10^{-5}, t(237) = 2.40, p < 0.05), surprisal (β = 3.45·10^{-4}, SE = 8.33·10^{-5}, t(24.72) = 4.14, p < 0.01), and bottom-up (β = 7.77· 10^{-4}, SE = 7.85·10^{-5}, t(43.24) = 9.98 p < 0.01). This suggests that the respective benefit from adding entropy, surprisal or bottom-up may be affected by the presence of top-down. The full output of the linear mixed model is shown in table 5.3.

The model comparison approach on the whole-brain average reconstruction accuracies of the GPT2-models (shown on the right in Figure 5.3) revealed a similar pattern. Indeed, the same model fit best to the data. The full model is displayed in Equation 5.8 below. Again, model comparison statistics are provided in Appendix 5.7.

$$\begin{aligned} accuracies \sim & entropy + surprisal + topdown + bottomup \\ & + entropy \cdot surprisal + entropy \cdot topdown \\ & + surprisal \cdot topdown + topdown \cdot bottomup \\ & + (topdown \cdot bottomup \cdot surprisal | subject) \end{aligned} \tag{5.8}$$

Using GPT2, there was also positive effect of surprisal (β = 1.20· 10^{-3}, SE = 6.63·10^{-5}, t(28.20) = 6.46, p < 0.01) and of bottom-up (β = 7.95· 10^{-4}, SE

= 3.03·10^{-4}, t(23.82) = 2.63, p < 0.05). In addition, top-down significantly decreased average reconstruction accuracy of the signal (β = -9.03· 10^{-4}, SE = 1.37·10^{-4}, t(39.23) = -6.61, p < 0.01), as did entropy (β = -3.89· 10^{-4}, SE = 6.63·10^{-5}, t(282.80) = -5.87, p < 0.01). Here, too, there were interactions between surprisal and entropy (β = -2.61·10^{-4}, SE = 7.65· 10^{-5}, t(282.80) = 3.41, p < 0.01), and between top-down and the other features (entropy: β = 1.97· 10^{-4}, SE = 7.65·10^{-5}, t(282.80) = 2.57, p < 0.05; surprisal: β = 1.53·10^{-4}, SE = 7.65·10^{-5}, t(282.80) = 1.99, p < 0.05; bottom-up: β = 8.43·10^{-4}, SE = 7.65· 10^{-5}, t(282.80) = 11.01, p < 0.01), suggesting that the effect of top-down may be less negative when bottom-up is part of the model. The full output of the linear mixed model is displayed in table 5.4 below.

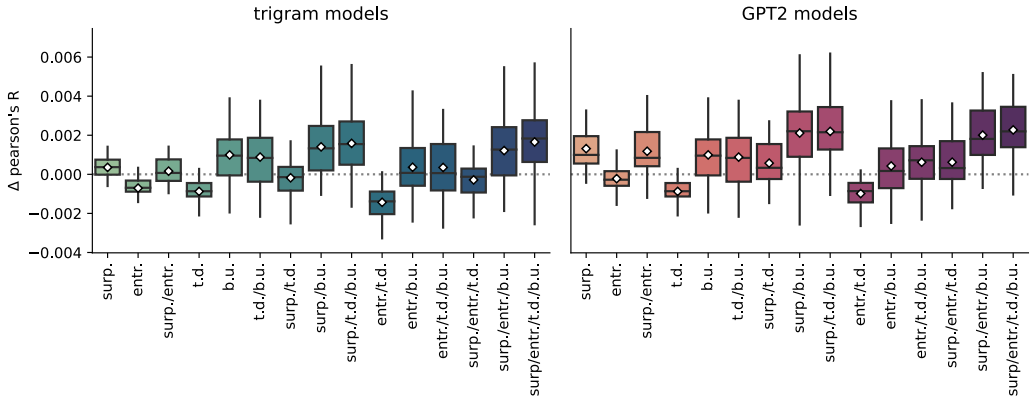


Figure 5.3: The difference in reconstruction accuracy (delta Pearson's R) between the base model (speech envelope, word onsets, word frequency) and the other fitted models (see Table 5.1) for trigram models and GPT2-models.) Abbreviations: surp.: surprisal; entr.: entropy; t.d.: top-down; b.u.: bottom-up.

Table 5.3: Results from the linear mixed effects model evaluating the main effects on trigram models.

	Coeff.	Std. Error	df	t value	p value	
(Intercept)	1.14e^{-01}	4.70e^{-03}	23.08	24.25	< 2e-16	***
Entropy	-6.81e^{-04}	5.37e^{-05}	237.00	-12.69	< 2e-16	***
Surprisal	3.64e^{-04}	1.31e^{-04}	26.24	2.79	9.70e^{-03}	**
Top-down	-8.88e^{-04}	1.51e^{-04}	25.23	-5.87	3.86e^{-06}	***
Bottom-up	1.03e^{-03}	3.25e^{-04}	22.99	3.19	4.12e^{-03}	**
Entropy * Surprisal	5.02e^{-04}	6.20e^{-05}	237.00	8.10	2.92e^{-14}	***
Entropy * Top-down	1.49e^{-04}	6.20e^{-05}	237.00	2.40	0.02	*
Surprisal * Top-down	3.45e^{-04}	8.33e^{-05}	24.72	4.15	3.47e^{-04}	***
Top-down * Bottom-up	7.77e^{-04}	7.85e^{-05}	43.24	9.89	1.12e^{-12}	***

Note. Signif. codes: * p < 0.05; ** p < 0.01; *** p < 0.001

Table 5.4: Results from the linear mixed effects model evaluating the main effects on GPT2 models.

	Coeff.	Std. Error	df	t value	p value	
(Intercept)	1.14e ^{01}	4.63e^{-03}	23.14	24.45	< 2e-16	***
Entropy	-3.89e^{-04}	6.63e^{-05}	282.80	-5.87	1.22e^{-08}	***
Surprisal	1.20e^{-03}	1.86e^{-04}	28.20	6.46	5.25e^{-07}	***
Top-down	-9.03e^{-04}	1.37e^{-04}	39.23	-6.61	7.31e^{-08}	***
Bottom-up	7.95e^{-04}	3.03e^{-04}	23.82	2.63	0.01	*
Entropy * Surprisal	2.61e^{-04}	7.65e^{-05}	282.80	3.41	7.54e^{-04}	***
Entropy * Top-down	1.97e^{-04}	7.65e^{-05}	282.80	2.57	0.01	*
Surprisal * Top-down	1.53e^{-04}	7.65e^{-05}	282.80	1.99	0.05	*
Top-down * Bottom-up	8.43e^{-04}	7.65e^{-05}	282.80	11.01	< 2e-16	***

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

To assess any differences between the trigram and GPT2-estimates of surprisal and entropy, we selected the models that included only these estimates (the top four models from table 5.1: *main_null, main_entropy, main_surprisal* and *main_distributional*) and subjected the average reconstruction accuracy for each participant, model and language model to a linear mixed model with the factors entropy, surprisal, language model (trigram or GPT2), and their interactions. We performed model comparison on fixed and random effects in the same way as above. Model comparison statistics are provided in Appendix 5.7.

The best model had the following structure:

$$\begin{aligned} accuracies \sim & model \cdot entropy \cdot surprisal \\ & + (1 + model \cdot surprisal | subject) \end{aligned} \tag{5.9}$$

This model revealed the same effects on surprisal and entropy as shown above; a positive effect of surprisal ($\beta = 1.32 \cdot 10^{-3}$, SE $= 1.10 \cdot 10^{-4}$, $t(28.08) = 5.53$, $p < 0.01$) and a negative effect of entropy ($\beta =$ -2.22$\cdot 10^{-4}$, SE $= 1.04 \cdot 10^{-6}$, $t(115) =$ -2.15, $p < 0.05$). In addition, there was an interaction between language model and surprisal ($\beta = 9.61 \cdot 10^{-4}$, SE $= 2.35 \cdot 10^{-4}$, $t(35.76) = 4.09$, $p < 0.01$), suggesting that the increase of reconstruction accuracy as a result of surprisal was larger in the GPT2-models than in the trigram models, and an interaction between language model and entropy ($\beta = 4.85 * 10^{-4}$, SE $= 1.47 \cdot 10^{-4}$, $t(115.00) = 3.31$, $p < 0.01$), indicating that the effect of entropy was more negative for the trigram models than the GPT2-models. Finally, there was a three-way interaction between language model, entropy and surprisal ($\beta = 4.28 \cdot 10^{-4}$, SE $=$ 2.07$\cdot 10^{-4}$, $t(115.00) = 2.07$, $p < 0.05$). Post-hoc comparisons revealed that this was because the interaction between entropy and surprisal was significant

for the trigram models, but not for the GPT2-models (GPT2-models: F(1,92) = 0.35, $p_{Bonferroni} = 1$; trigram models: F(1,92)=12.34, $p_{Bonferroni} < 0.01$). The interaction between entropy and surprisal in the trigram models was driven by an effect for entropy only when surprisal was not in the model (no surprisal: F(1,92) = 46.59, $p_{Bonferroni} < 0.01$; surprisal: F(1,92) = 3.45, $p_{Bonferroni} = 0.27$). The full model output is shown in table 5.5.

Table 5.5: Results from the linear mixed effects model comparing GPT and trigram models.

	Coeff.	Std. Error	df	t value	p value	
(Intercept)	$1.14e^{01}$	$4.86e^{-03}$	23.01	23.38	$< 2e^{-16}$	***
Language model	$-6.58e^{-15}$	$1.10e^{-04}$	85.52	0.00	1.00	
Entropy	$-2.22e^{-04}$	$1.04e^{-06}$	115.00	-2.15	0.03	*
Surprisal	$1.32e^{-03}$	$2.38e^{-04}$	28.08	5.53	$6.55e^{-06}$	***
Language model * entropy	$-4.85e^{-04}$	$1.47e^{-04}$	115.00	-3.31	$1.25e^{-03}$	**
Language model * surprisal	$-9.61e^{-04}$	$2.35e^{-04}$	35.76	-4.09	$2.33e^{-04}$	***
Entropy * surprisal	$8.61e^{-05}$	$1.47e^{-04}$	115.00	0.59	0.56	
Language model * entropy * surprisal	$4.28e^{-04}$	$2.07e^{-04}$	115.00	2.07	0.04	*

Note. Signif. codes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

In summary, the features surprisal and bottom-up had positive effects on the whole-brain average reconstruction accuracy. The features entropy and top-down appear to bring down the whole-brain average reconstruction accuracy values. In addition, the presence of the top-down feature affects the relative benefit (or detriment) of the other features. Furthermore, the GPT2-derived surprisal features are a better predictor for the delta-band neural signal than the trigram-derived surprisal features.

5.3.2 Main effects: Cluster-based permutation tests

To gain some insight into the spatial distribution of the effects described above, we computed the per-participant per-sensor averages for the models that did- or did not include a given feature.[2] As can be observed in Figures 5.4 and 5.5

[2]This was done after attempting to fit a linear mixed model on each sensor separately with maximal per-participant random effects, in order to extract a coefficient for every feature (per subject, per sensor) that was independent from the other features rather than averaging across all models. This approach has failed so far, on the one hand due to the limited number of observations per participant (after all, there is only one reconstruction accuracy value per participant), and on the other hand due to almost non-existing between-participant variability for some features (most notably, entropy), leading to infinite values in the model. This approach is to be explored further: fitting the models separately on each story part separately could be a good option for this, as such obtaining a TRF and a reconstruction accuracy value for every story (per participant, per sensor). However, the exploration is beyond the scope of this thesis Chapter.

below, the general pattern is the same as in the analysis on the whole-brain average reconstruction accuracies: bottom-up and surprisal contribute positively to model fit, while entropy and top-down do not. Both bottom-up node counts and surprisal values contribute to model fit in large, bilateral clusters, although the bottom-up node count feature notably contributes more to model fit in the left hemisphere. As for surprisal, it is clear that both GPT2-derived and trigram surprisal values contribute to model fit in almost all sensors. The pattern of improvement is slightly different between the two, with less contribution around auditory areas from the trigram surprisal values. The difference was not tested statistically, so we do not draw conclusions on the basis of this. Both measures of entropy appear to decrease model fit bilaterally. The top-down node count feature is a curious case: there is no evidence for an improvement, but the decrease is significant only in the right hemisphere when the additional features are drawn from GPT2-models.

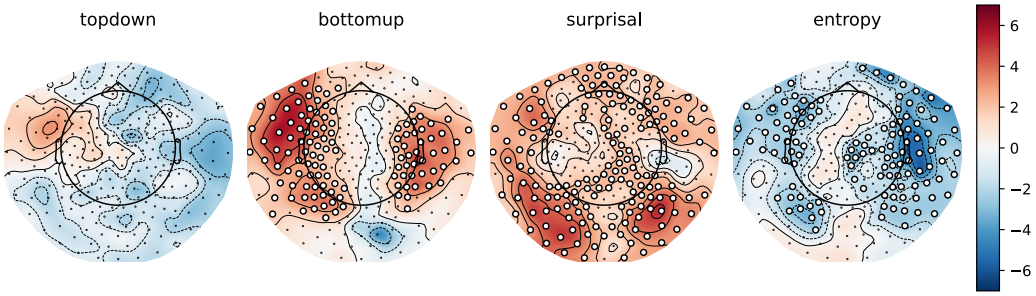


Figure 5.4: Trigram models. Scalp maps of the t-values resulting from the contrast between the averages of all models that contain a specific predictor (e.g., top-down) and all models that do not contain this predictor. Each scalp map represents this contrast for a different feature. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

With respect to our first question, these analyses revealed that as a general predictor of the delta-band neural signal, surprisal estimates from large models like GPT2 outperform the short-context trigram models. In addition, from a methodological perspective, the above analyses clearly show the non-trivial effect the addition of some features to the TRF-model can have on the reconstruction accuracy of other features using ridge regression specifically, even when these features are not dangerously correlated as indicated by the Variance Inflation Factor. For the purposes of the rest of the present study, these analyses have provided sufficient evidence for the positive contributions of bottom-up node count and lexical surprisal features in a model of the delta-band neural signal.

We will therefore continue further analyses on these two features, and remove top-down node count and lexical entropy from our models.

5.3.3 Interaction effects

To investigate whether the process of syntactic structure building is affected by lexical distributional information, we split the bottom-up node count feature into two features: bottom-up node count for low surprisal words (i.e., words that are statistically relatively predictable from the context) and high surprisal words (i.e., words that are statistically (i) relatively unpredictable or (ii) unpredicted from the context). We then compared the resulting TRFs, which capture the neural response to bottom-up structure building, between these surprisal-conditions. Differences between the TRFs provide insight into how the process of structure building may be mediated by distributional information.

Trigram models When using a simple trigram model for surprisal estimation and using those values to divide the words over 'high surprisal' and 'low surprisal' categories, the cluster-based permutation test reveals a widespread difference between the TRFs. The clusters that contributed to the difference between the distributions had a bilateral distribution across the scalp, although the effects were most pronounced in the left-frontal area. The clusters were spread out across the entire time-window, meaning that the effects were visible slightly before word onset and lasted approximately one second after word onset. As can be observed in Figure 5.6 below, the response to bottom-up node count appears

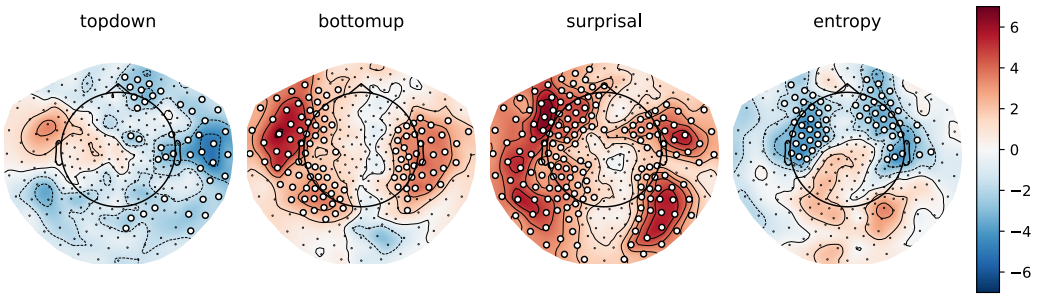


Figure 5.5: GPT2-models. Scalp maps of the t-values resulting from the contrast between the averages of all models that contain a specific predictor (e.g., top-down) and all models that do not contain this predictor. Each scalp map represents this contrast for a different feature. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

to be slowed down and increased in magnitude in high-surprisal words relative to the low-surprisal words.

The pattern on the reconstruction accuracy values was slightly different. Despite this large difference between the bottom-up node count response to high- and low surprisal words, the cluster-based permutation test revealed no difference between a model that split the bottom-up node count feature by surprisal and a model that split the bottom-up node count feature randomly. In addition, a comparison between the relative benefit of the high- and low-surprisal bottom-up node count features revealed that the high-surprisal bottom-up node count feature explained more variance in the delta-band neural signal than did the low-surprisal feature (see Figure 5.7C). Comparing the effects of these halves of the bottom-up feature separately (i.e., a model with only the high surprisal bottom-up feature versus a model that did not include this feature; a model with only the low surprisal bottom-up feature versus a model that did not include this feature) did not reveal any effects (Figure 5.9A and B).

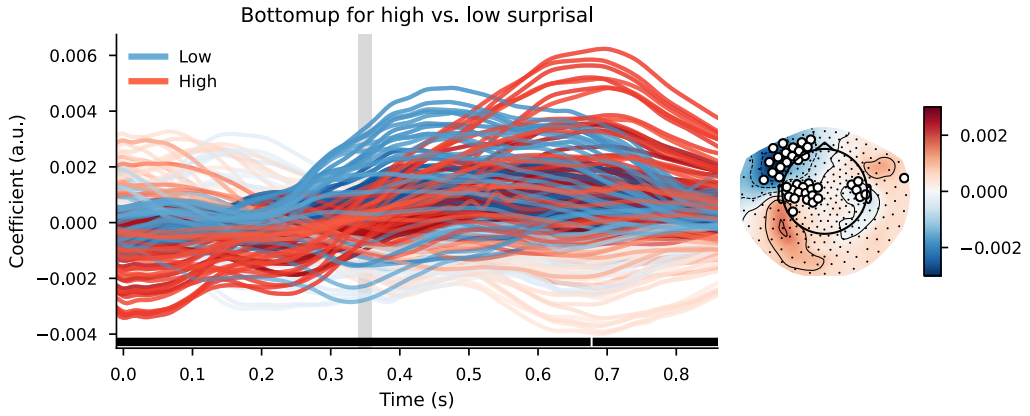


Figure 5.6: The bottom-up node count TRF for high surprisal (in red) and low surprisal (in blue) with surprisal from the trigram model as the dividing estimate. Individual lines represent sensors. The displayed sensors contributed to the clusters that allowed us to reject the null-hypothesis. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Vertical gray bar is the time-point of the scalp map displayed on the right. The scalp map shows the difference between the coefficients from the high- and low surprisal words. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null-hypothesis at the time-point of the gray bar.

GPT2-models When using a complex language model (GPT2) to divide words over 'high surprisal' and 'low surprisal' categories, the general pattern on the TRFs was similar. The cluster-based permutation test revealed a widespread difference between the bottom-up node count responses for low-surprisal and high-surprisal words. The clusters that contributed to this difference had a wide temporal distribution, with clusters between before word onset to 500 milliseconds after word onset, and a cluster at the end of the time window – approximately from 700 milliseconds onwards. Again, the most prominent difference was in left temporal/frontal sensors. Visual inspection of the TRFs (Figure 5.8) suggests, again, a later response to the bottom-up node count feature for high-surprisal words than for low surprisal words.

The pattern on the reconstruction accuracies was inconclusive. The cluster-based permutation test comparing the reconstruction accuracy values for models with a random split of the bottom-up node feature to the model with a surprisal-based split of the bottom-up node count feature revealed no effects. There were also no effects for the separate node-count predictors on the reconstruction of the neural signal relative to a base model, nor was there a difference between these differences.

Taken together, the results from the interaction analyses suggest that even a short-context surprisal value affects the timing of structure building (in a bottom-up fashion). The lack of effects on the reconstruction accuracy values make the

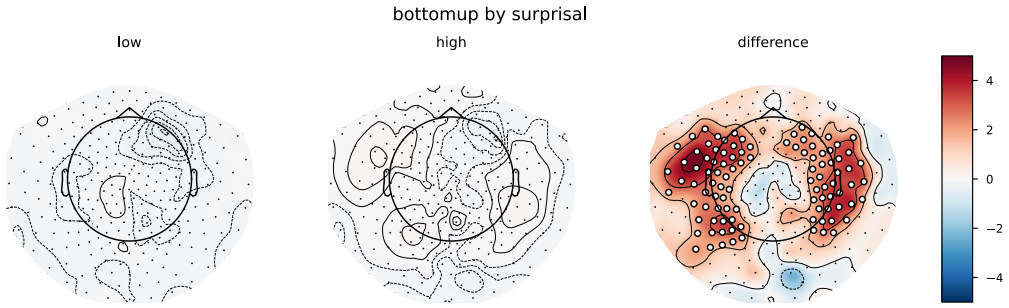


Figure 5.7: Trigram models. Scalp maps of the t-values resulting from the contrast between the model main_surprisal *and the model with only low surprisal bottom-up node counts (A; leftmost panel)* or *high surprisal bottom-up node counts (B; middle panel). The rightmost panel (C) depicts the contrast between those contrasts, i.e. the difference in increase of reconstruction accuracy between high- and low surprisal bottom-up node counts.* Red values show that the high surprisal bottom-up node counts explain more variance. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

results difficult to interpret: apparently, a systematic split for node count by surprisal does not lead to a significantly higher reconstruction accuracy than a random split for node count. This suggests that the effect is small – or there may be confounding factors that obscure the state of affairs.

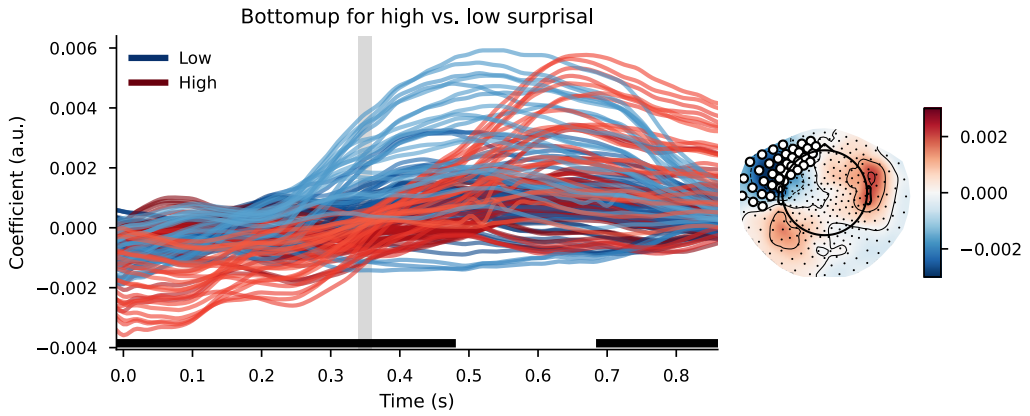


Figure 5.8: The bottom-up node count TRF for high surprisal (in red) and low surprisal (in blue) with surprisal from GPT2 as the dividing estimate. Individual lines represent sensors. The displayed sensors contributed to the clusters that allowed us to reject the null-hypothesis. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Vertical gray bar is the time-point of the scalp map displayed on the right. The scalp map shows the difference between the coefficients from the high- and low surprisal words. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null-hypothesis at the time-point of the gray bar.

5.3.4 Correction for word duration

Indeed, there are several factors that potentially correlate with (or are causal of) surprisal. One of the correlating factors that is not in itself causal of surprisal values in a way that, for example, word frequency might be, is particularly interesting for our current finding: the factor of word duration. Indeed, higher surprisal values tend to be associated with longer word durations (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 2011); this is also the case in our data (trigram-models: duration of high-surprisal words is significantly higher than the duration of low surprisal word; $t(8548) = 49.14$, $p < 0.01$, $mean_{high}(SD) = 0.32s$ (0.16); $mean_{low}(SD) = 0.17s$ (0.12); significant positive correlation between duration and surprisal ($\rho = 0.59$; $p < 0.01$); GPT2-models: duration of high-surprisal words is significantly higher than the

duration of low surprisal word; $t(8548) = 37.55$, $p < 0.01$, $mean_{high}(SD) = 0.31s$ (0.17); $mean_{low}(SD) = 0.19s$ (0.13); significant positive correlation between duration and surprisal ($\rho = 0.42$; $p < 0.01$). This suggests that the later response to node counts could – very simply – be an effect of slower processing of longer words (New, ferrand, pallier, & brysbaert, 2006; Tyler, Voice, & Moss, 2000).

To examine whether word duration could explain our effects, we extracted the TRFs for bottom-up node count only for words that were matched for duration. We did this by computing a histogram (100 bins) for word duration in both low- and high surprisal conditions, and extracting the overlap between these two distributions. We then took a random subsample of words from both the low- and high surprisal conditions such that the resulting distribution of word durations was the same in both conditions. This resulted in a subset of approximately half of the words (4556 out of 8550 words). The distributions and their overlap are displayed in Figure 5.9 below. All words selected in this analysis came from the yellow shaded distribution, creating two sets of words that differed in their surprisal value along the median, but that had near-identical distributions of word duration. We then used only these words to compute the bottom-up TRF in low- and high surprisal conditions, as well as a random split (like above).

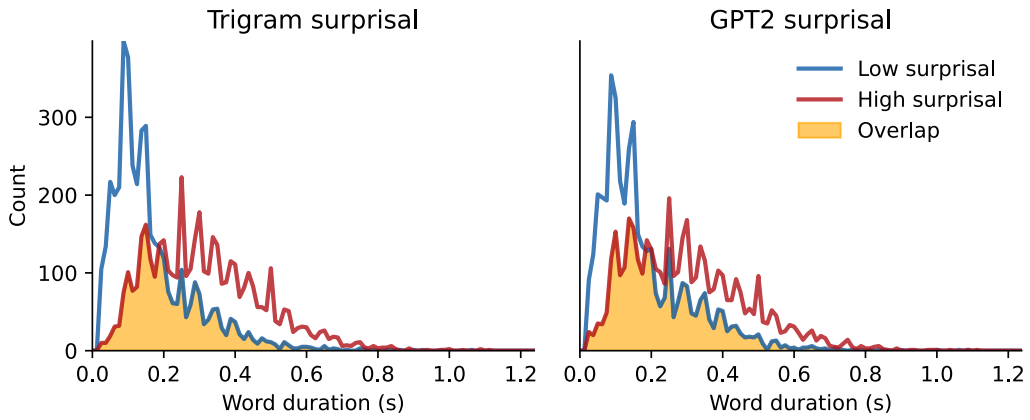


Figure 5.9: The histograms of word duration. (100 bins between 0.0 and 1.25 seconds) split along the median of surprisal (low surprisal in blue, high surprisal in red). The yellow shaded area is the overlapping distribution from which words were selected to correct for differences in word duration.

Trigram models Despite this extensive subsampling, the cluster-based permutation test revealed a significant difference between the responses to bottom-

up node count. The clusters that contributed to this difference had a left frontotemporal distribution and were constrained to a relatively early time window – between 50 and 450 milliseconds. Visual inspection of the waveforms again suggested a temporal shift in the response to bottom-up node count, with the bottom-up node count response delayed for high-surprisal words relative to low-surprisal words; see Figure 5.10.

The assessment of significance through the reconstruction accuracy values showed a clearer image this time: the cluster-based permutation test that assessed difference between a systematic split by surprisal and a random split of the bottom-up node count values came back significant, with reconstruction accuracies higher when the bottom-up node count values were split by the median surprisal values than when they were split randomly. See Figure 5.11 below. Despite this difference being significant, there were no differences between the contribution of the bottom-up node count feature for low or high surprisal words. This difference did exist before controlling for word duration (see 5.3.3).

The significantly different reconstruction accuracy values between models with randomly split bottom-up predictors and surprisal split bottom-up predictors confirms the effects found on the TRF-waveforms: a model that allows for variation between words on the basis of surprisal leads to a better description of the signal than a model that allows variation randomly. This suggests that the temporal shift we observe visually is indeed there. To look into this in more detail, we computed the cross-correlation between the high and low surprisal bottom-up node count responses for the sensors that contributed to the difference between the two distributions.

The cross-correlation on the sensors that were part of the clusters (depicted in Figure 5.10) revealed that the time-point at which the correlation was highest for most sensors was at 150 milliseconds post word-onset. The average correlation between the shifted low-surprisal (as shown in Figure 5.12C below) and the original high-surprisal response was 0.73 (sd = 0.28), with a maximum of 0.95. This high correlation for a large subset of channels, which did not exist for random selections of channels and time-shifts (see Figure 5.12D) suggests that the response to high surprisal words was delayed by 250 milliseconds relative to the low surprisal words. At the same time, the relatively high variance between channels may indicate either that the temporal shift is not uniform (i.e., not all readouts of structure-building (potentially with different neural sources) are temporally affected by contextual surprisal) or that some readouts of structure building from different neural sources respond qualitatively different as a

function of surprisal (i.e., shifting them in time in any direction will not increase the correlation).

GPT2-models Despite the subsampling, the difference between high- and low surprisal bottom-up node count responses remained also when the split was based on GPT2-extracted surprisal values. The cluster-based permutation test showed a difference between the two time-courses. Clusters that contributed to this difference were once again prominent in left-frontotemporal areas and spanned a wide time-window, with one cluster ranging from word onset to approximately 550 milliseconds, and a second cluster in a late time-window, from around 700 milliseconds onwards (Figure 5.13). As in the analysis on the split by trigram surprisal, excluding words on the basis of an extremely long or short duration clarified the picture: splitting the bottom-up node count feature on the basis of surprisal led to higher reconstruction accuracy than splitting the bottom-up node count feature randomly (Figure 5.14).

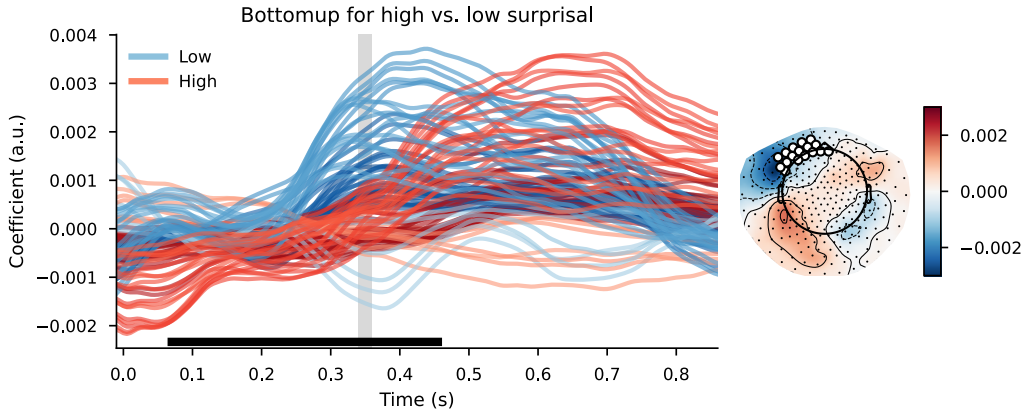


Figure 5.10: The bottom-up node count TRF for high surprisal (in red) and low surprisal (in blue) with surprisal from the trigram model as the dividing estimate, after correction for word duration. Individual lines represent sensors. The displayed sensors contributed to the clusters that allowed us to reject the null-hypothesis. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Vertical gray bar is the time-point of the scalp map displayed on the right. The scalp map shows the difference between the coefficients from the high- and low surprisal words. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null-hypothesis at the time-point of the gray bar.

While there was no difference between the high- and low surprisal bottom-up node count features when it comes to variance explained before correcting for word duration, there was a difference now: low-surprisal bottom-up node counts explained more variance than the high-surprisal counterpart (Figure 5.15). This is in stark contrast with the findings on the trigram models, where we observed a larger contribution to the reconstruction accuracy for the high-surprisal bottom-up node counts that disappeared after correcting for word duration.

To further investigate the potential temporal shift of the response as a function of surprisal, we performed a cross-correlation analysis. The cross-correlation revealed that the time-point at which the correlation was highest for most sensors was at 190 milliseconds post word-onset. The correlation was 0.71 on average (sd = 0.29), with a maximum of 0.93. As before, the high correlation – which did not exist for a random subset of channels and time-points, see Figure 5.16D – suggests that the response indeed varies in time as a function of lexical surprisal. Again, relatively large variance suggests that this may not be the case for all sensors in the selection: the temporal shift may not be the same for all sensors, or the response qualitatively differs between low- and high surprisal words for some sensors.

When we jointly consider all of the results above, a pattern emerges. Firstly, there is a clear indicator that temporal properties of a readout of structure build-

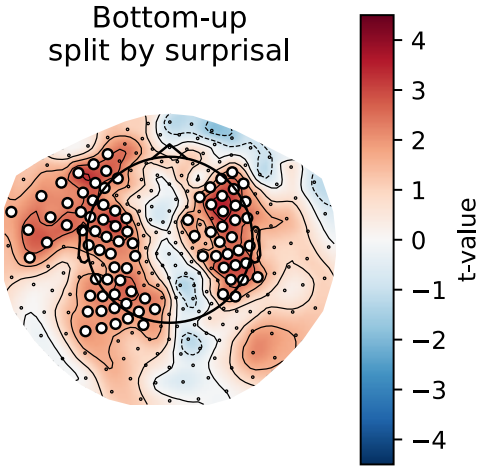


Figure 5.11: Scalp-map of the t-values resulting from the contrast 'Systematic split of the bottom-up predictor' vs. 'random split of the bottom-up predictor' using surprisal from the trigram model as the dividing estimate. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

ing, extracted with bottom-up node counts, are affected by lexical surprisal: structure building operations are performed later when the surprisal of the word to be integrated in the sentence is higher relative to when the surprisal of that word is lower. When we consider a split by the median, this temporal shift appears quite big: between 150 and 200 milliseconds. This is confirmed by the reconstruction accuracy values: reconstruction accuracy values are higher when the bottom-up node count feature is split systematically using surprisal, than when the feature is spit randomly.

Secondly, when it comes to the amount of variance explained by structure-building operations, the pattern is affected by word duration. When we do not

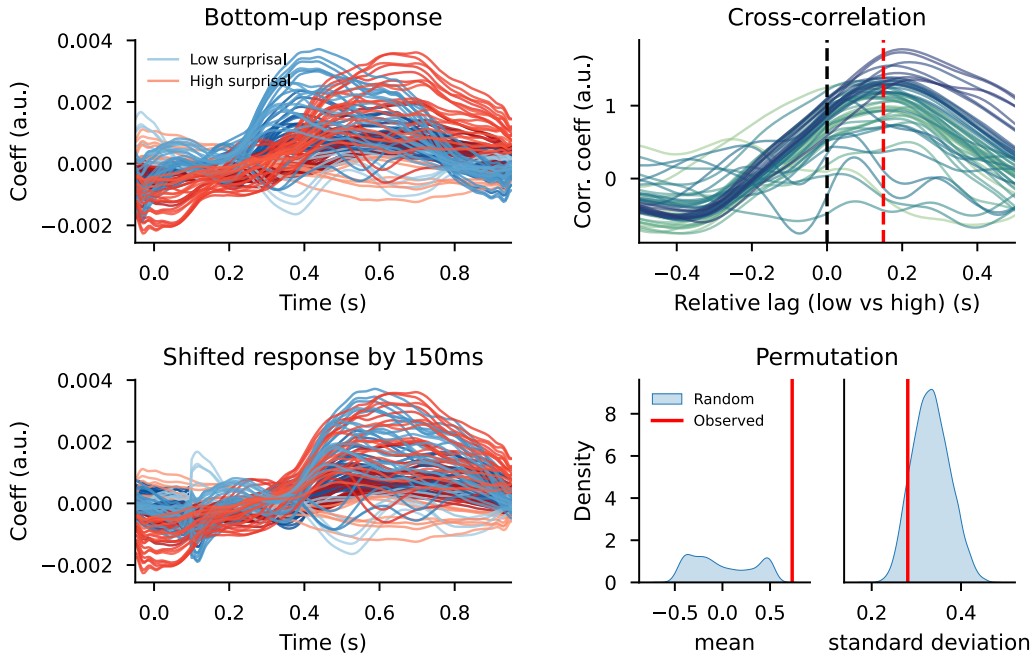


Figure 5.12: Cross-correlation results for the trigram models, after selection for word duration. (A) top left: bottom-up TRF time-courses for the sensors from the cluster-based permutation test between high surprisal (in red) and low surprisal (in blue). (B) top right: cross-correlation between the high- and low surprisal bottom-up responses for the sensors from the clusters (scaled). Colors indicate sensors. (C) bottom left: the shifted response from the low surprisal condition (in blue) to overlap with the high surprisal condition (in red). (D) bottom right: kernel density plots of means and standard deviations from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation.

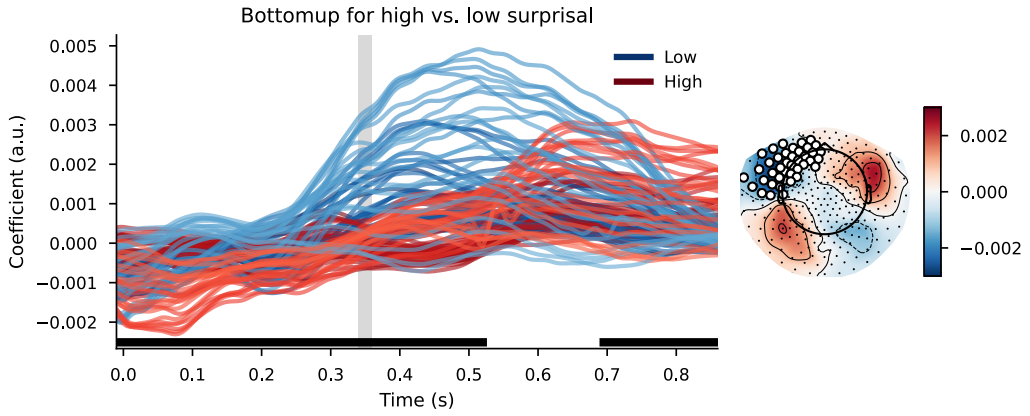


Figure 5.13: The bottom-up node count TRF for high surprisal (in red) and low surprisal (in blue) with surprisal from GPT2 as the dividing estimate, after correction for word duration. Individual lines represent sensors. The displayed sensors contributed to the clusters that allowed us to reject the null-hypothesis. Black bars indicate time points that contributed to clusters that allowed us to reject the null-hypothesis. Vertical gray bar is the time-point of the scalp map displayed on the right. The scalp map shows the difference between the coefficients from the high- and low surprisal words. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null-hypothesis at the time-point of the gray bar.

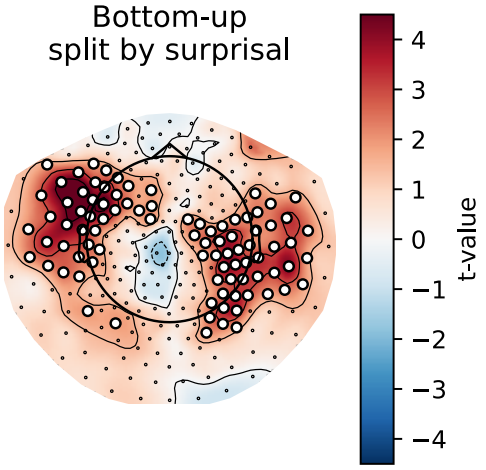


Figure 5.14: Scalp-map of the t-values resulting from the contrast 'Systematic split of the bottom-up predictor' vs. 'random split of the bottom-up predictor' using surprisal from GPT2 as the dividing estimate. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

correct for word duration, we observe a larger variance explained by bottom-up node count for high surprisal than for low surprisal, but only when surprisal is extracted from a trigram model; there is no difference when surprisal is extracted from GPT2. However, after correction for word duration, we observe that the difference found for the trigram models disappears – i.e., high surprisal bottom-up node count no longer explains more variance than low-surprisal bottom-up node count – and the *reverse* pattern is seen for the analysis using GPT2: here, *low* surprisal bottom-up node counts appear to explain more variance than the high surprisal bottom-up node counts.

These observations pattern with the amplitudes of the TRF waveforms. In the trigram model, before correcting for word duration, the amplitude of the node count-response to high-surprisal words is larger than the amplitude of the response to low-surprisal words (see Figure 5.6). This difference appears to disappear after correction for word duration (see Figure 5.10). At the same time, there is no obvious amplitude difference between the high- and low surprisal node-count responses from the GPT2-model (Figure 5.8), while the low-surprisal bottom-up node count has an obviously larger amplitude after correction for word duration (Figure 5.13). Taken together, this suggests a non-trivial relationship between word duration, language model for surprisal estimation, and response amplitude. We will return to this in the discussion.

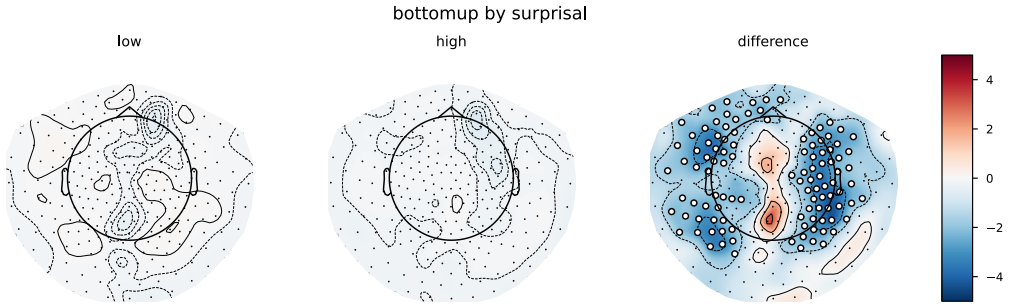


Figure 5.15: GPT2 models after correction for word duration. Scalp maps of the t-values resulting from the contrast between the model *main_surprisal* and the model with only low surprisal bottom-up node counts (A; leftmost panel) *or* high surprisal bottom-up node counts (B; middle panel). The rightmost panel (C) depicts the contrast between those contrasts, i.e. the difference in increase of reconstruction accuracy between high- and low surprisal bottom-up node counts. Red values show that the high surprisal bottom-up node counts explain more variance. White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

5.3.5 The role of word recognition

A 150 to 190 millisecond shift in response time begs the question to what extent the delay is driven by lexical contextual information directly affecting structure-building operations. After all, there is an important process that – in an interactive, cascaded model of language comprehension – occurs prior to or simultaneously with the generation of syntactic structure: word recognition. A word that is predictable from the context is recognized faster (Grosjean & Itzler, 1984) and read faster (Amenta, Hasenäcker, Crepaldi, & Marelli, 2023; Aurnhammer & Frank, 2019). In a cascaded architecture, then, an earlier completing or faster

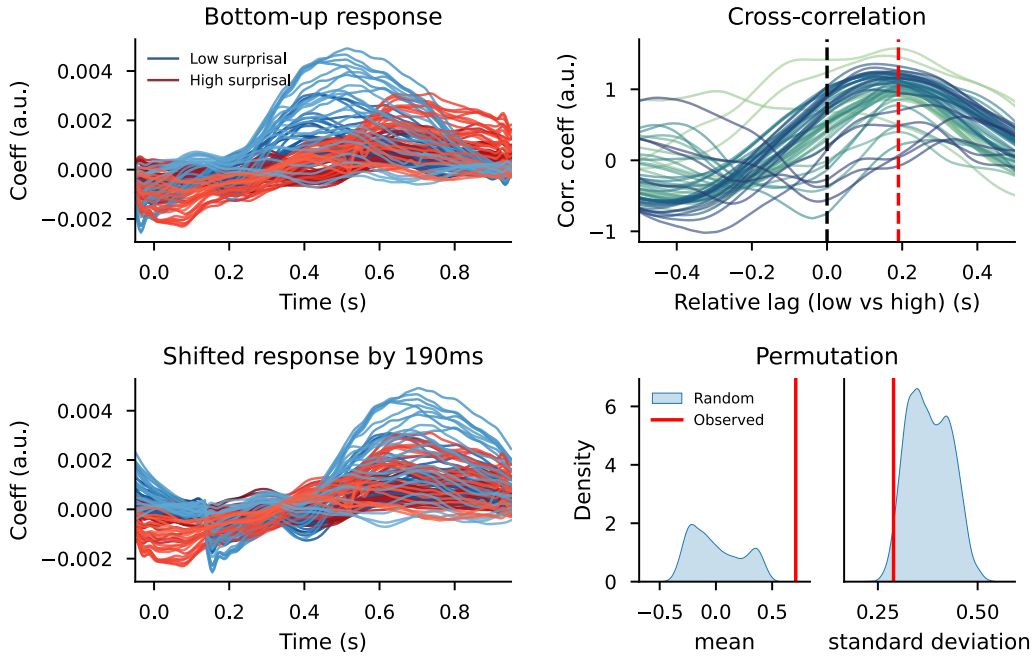


Figure 5.16: Cross-correlation results for the GPT2 models, after selection for word duration. (A) top left: bottom-up TRF time-courses for the sensors from the cluster-based permutation test between high surprisal (in red) and low surprisal (in blue). (B) top right: cross-correlation between the high- and low surprisal bottom-up responses for the sensors from the clusters (scaled). Colors indicate sensors. (C) bottom left: the shifted response from the low surprisal condition (in blue) to overlap with the high surprisal condition (in red). (D) bottom right: kernel density plots of means and standard deviations from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation.

process of word recognition could affect the time-course of the inference of syntactic structure.

To investigate this, we performed the same contrast for high- versus low surprisal on a feature that captures the presence of lexical information in the neural signal: word frequency (Slaats et al., 2023). That is, this time, we split the *word frequency* feature into two separate features on the basis of the surprisal values (high-surprisal word frequency, low-surprisal word frequency). Again, we only performed this analysis for the words obtained from the overlapping distributions of word length to exclude the possibility that word duration drives any of the effects.

Trigram models Interestingly, the cluster-based permutation test revealed that the word frequency response differed between high- and low surprisal words, suggesting that there is indeed a difference in lexical processing between high- and low surprisal words. This is further confirmed by a higher reconstruction accuracy for a split of word frequency by surprisal than a random split of word frequency (see Figure 5.17 below). However, this difference is crucially *not* temporal in nature. In fact, it appears to be one of amplitude: coefficients are higher for the low-surprisal words than for the high-surprisal words. The cross-correlation on the sensors that differed between conditions revealed that the correlation between the word frequency response to high- and low surprisal words was highest at a delay of zero milliseconds. This indicates that there is no detectable time-shift, which can be clearly observed in Figure 5.18A, B and C. The correlation between the two responses was not high on average (mean = 0.22), though there was considerable variance: with a standard deviation of 0.44, some channels had a Pearson's correlation coefficient of 0.97 between the two conditions, though more than 45% of the channels had a correlation coefficient lower than 0.2.

GPT2-models When using GPT2 to divide words over high- and low surprisal condition, we again observe a difference between the responses that appears mostly one of amplitude (see Figure 5.19A, left upper corner). However, the difference between a GPT2-split and a random split does not reach significance. If we use the sensors that are part of clusters that contribute to the significant difference between the two responses to perform a cross-correlation, we observe a similar pattern as for the trigram model. The highest correlation between the two responses was at a time-lag close to zero: the two responses were most simi-

lar at a delay of -30 milliseconds – which positions the response to high-surprisal words slightly *before* the response to low-surprisal words. In other words, despite there is a timing effect here, this effect is in the opposite direction. These analyses suggest therefore that temporal differences in the process of lexical retrieval are not the cause of the delayed response to bottom-up structure building – though qualitative differences between the processes can still play a role.

5.4 Discussion

In this study, we investigated how the delta-band neural signal represents and exploits lexical distributional information in the process of syntactic structure building during auditory language comprehension. We approached this question in two main sub questions. Firstly, we asked whether trigram- or GPT2-derived estimates of lexical surprisal are a better model of the delta-band neural signal during language comprehension. Secondly, we asked whether the delta-band neural readout of syntactic structure building changes as a function of the distributional properties of a word, and if this influence can be linked to probabilities based on the immediately preceding words (as reflected in surprisal and entropy

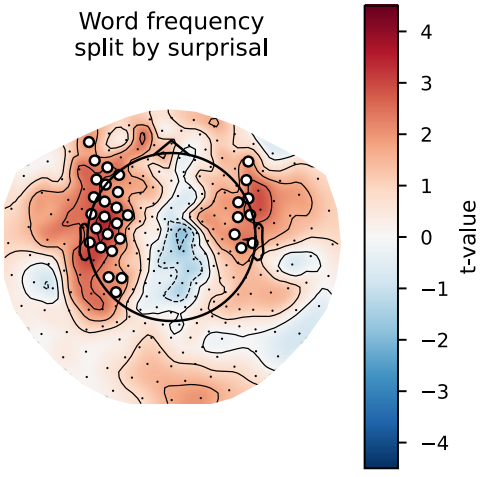


*Figure 5.17: Scalp-map of the t-values resulting from the contrast 'Systematic split of the word frequency predictor' vs. 'random split of the word frequency predictor' using surprisal from the trigram model as the dividing estimate.*White dots on the scalp map indicate the sensors that contributed to the clusters that allowed us to reject the null hypothesis (i.e., the difference is not 0).

estimates from a trigram model), or rather to the larger context (as reflected in GPT2-models).

To answer these questions, we used a modelling approach and a naturalistic listening paradigm. We presented participants with audiobooks while we recorded their MEG, and analyzed the resulting data using temporal response functions (TRFs). This linear regression approach allowed us to study high-level processes during language comprehension, while controlling for lower level processes like speech tracking. Our analysis consisted of two parts: a main-effects analysis to evaluate which features modeled the delta-band neural signal most

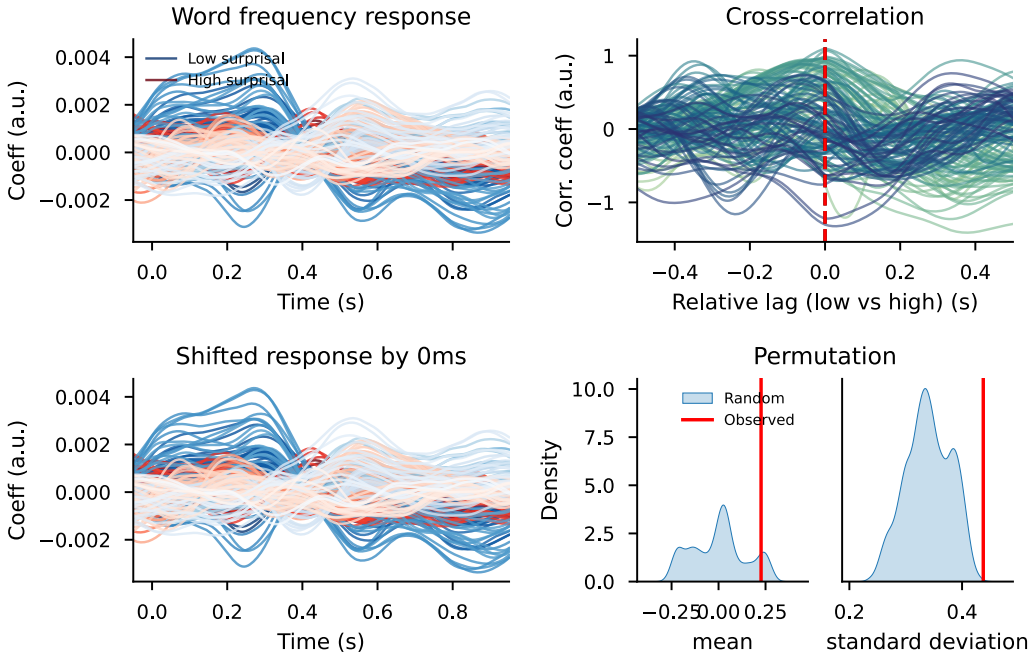


Figure 5.18: Cross-correlation results for the trigram models on the word frequency feature, after selection for word duration. (A) top left: word frequency TRF time-courses for the sensors from the cluster-based permutation test between high surprisal (in red) and low surprisal (in blue). (B) top right: cross-correlation between the high- and low surprisal word frequency responses for the sensors from the clusters (scaled). Colors indicate sensors. (C) bottom left: the shifted response from the low surprisal condition (in blue) to overlap with the high surprisal condition (in red). (D) bottom right: kernel density plots of means and standard deviations from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation.

accurately; and an interaction analysis, to evaluate whether lexical distributional information affects the process of syntactic structure building (the inference of syntactic structure).

5.4.1 Describing the delta band neural signal: Surprisal and bottom-up node counts

The main effects-analysis showed that the features that contributed positively to models of the data were bottom-up node counts and lexical surprisal. These features were used for further analysis in the interaction-analysis. This finding

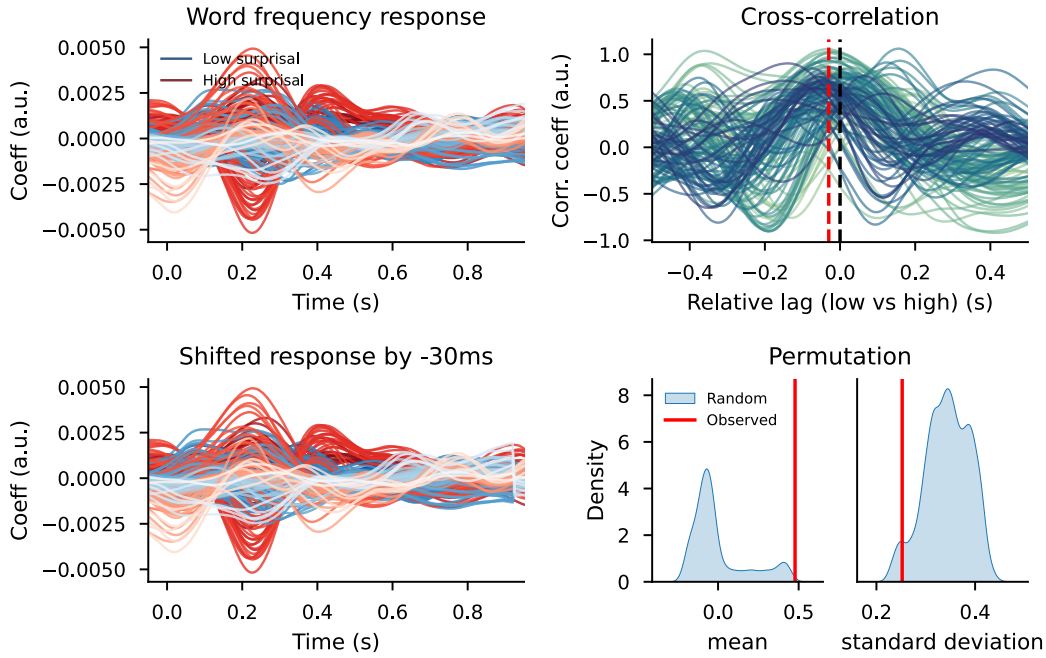


Figure 5.19: Cross-correlation results for the GPT2 models on the word frequency feature, after selection for word duration. (A) top left: word frequency TRF time-courses for the sensors from the cluster-based permutation test between high surprisal (in red) and low surprisal (in blue). (B) top right: cross-correlation between the high- and low surprisal word frequency responses for the sensors from the clusters (scaled). Colors indicate sensors. (C) bottom left: the shifted response from the low surprisal condition (in blue) to overlap with the high surprisal condition (in red). (D) bottom right: kernel density plots of means and standard deviations from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation.

is in accordance with studies that used other features to model the delta-band neural signal, such as phonemes: the delta-band signal is modelled better with surprisal than with entropy, while the opposite is true for the theta band (4-8/10 Hz) (Donhauser & Baillet, 2020; Mai & Wang, 2023). That the delta-band neural signal is influenced by syntactic structure is known (Kaufeld, Bosker, et al., 2020; Lo et al., 2022; Lu et al., 2022; Slaats et al., 2023), but which parsing strategy yields the best predictors for the delta-band neural signal is an open question. Some studies suggest that bottom-up parsing strategies are more predictive of the neural signal (Giglio et al., 2024; Nelson, El Karoui, et al., 2017), others found no difference (Brennan et al., 2016), and there is even evidence for importance of the top-down strategy in favor of bottom-up (Coopmans, 2023). It is likely that the exact paradigm (production, comprehension) methodology (EEG, fMRI, iEEG, MEG) and analysis choices (source localization, TRF-estimation algorithm, etc) influence the outcome of this comparison. In our study, bottom-up node counts had good performance, and for this reason, we continued with this feature.

Furthermore, the main-effects analysis revealed that surprisal extracted from GPT2, a large language model that was fine-tuned for Dutch using a context-window of 128 tokens ($\sim$ 128 words), performed better in our TRF model of the data than surprisal calculated using a trigram model, despite both of the metrics performing well. This finding is in line with previous findings by Heilbron and colleagues (2019), who compared TRFs and reconstruction accuracy for trigram- and GPT2-estimates in continuous listening in English. As in the present study, the authors showed that GPT2-derived surprisal estimates performed much better than trigram surprisal estimates.

An important open question is *why* GPT2-derived surprisal estimates perform better. From a psychological perspective, a possibility is that the brain represents both long- and short context distributional information during language comprehension (potentially independently from each other; Goodkind & Bicknell, 2021). Surprisal estimates from the fine-tuned GPT2-model are sensitive to variability at a distance of 128 tokens. This means that a word's relation to the overall discourse is represented in those probability estimates, while this relation is hardly captured by probability at a short distance of two words. At the same time, though, GPT2-estimates do not exclude the probability of a word given the immediate context, as the two previous words are obviously part of the input to estimate surprisal for the current word. This means that GPT2-estimates of surprisal capture some of the same regularities as the trigram model. In that sense,

GPT2 captures not only long-context effects, but also short-context effects. We must not overinterpret this result: this difference does not mean that the model architecture of GPT2 is "more human" than the trigram model. In fact, we know that GPT2 can represent long-context effects that are beyond what humans can maintain in memory (see Guest & Martin, 2023). Within reason, therefore, we can conclude that the results from the trigram- and GPT2-models show that the delta-band neural signal covaries with surprisal estimates that find their origin in both short- and long contexts during language comprehension.

5.4.2 Computation of structure in time

The aim of this study was to assess whether the neural encoding of linguistic structure changes as a function of the distributional properties of a word, and whether this influence can be linked to probabilities in the immediate context (two preceding words) or rather to probabilities in the larger context (operationalized using GPT2). To this end, we extracted responses to annotations of syntactic structure, and we evaluated whether these responses differed between words that were statistically predictable (low surprisal) and words that were statistically relatively unpredictable (high surprisal).

The analysis revealed that distributional properties of a word affected the process of syntactic structure building in the temporal domain. Even after correcting for word duration, the response to a metric of syntactic structure – bottom-up node count – occurred earlier for words that were statistically predictable given the context (low surprisal) than for words that were unpredictable given the context (high surprisal). This effect was clearly visible using a simple, short-context metric of lexical distributional information: trigram surprisal. A cross-correlation on the grand average waveforms indicated that the neural signature of structure building occurred ∼150 milliseconds earlier for low-surprisal words than for high-surprisal words. The temporal effect was slightly larger when using surprisal from GPT2, the operationalization of long-context surprisal: in this case, the neural signature of structure building was observed ∼190 milliseconds earlier for low-surprisal words relative to high-surprisal words.

In an interactive, cascaded model of language comprehension, word recognition is hypothesized to occur prior to or simultaneously with the inference of syntactic structure (Marslen-Wilson & Welsh, 1978; Martin, 2016, 2020). Because words that are predictable from the context are recognized and read faster than words that are not predictable (Amenta et al., 2023; Aurnhammer & Frank, 2019; Grosjean & Itzler, 1984), it was deemed necessary to evaluate whether

there is a difference in the time course of lexical processing between the high- and low-surprisal words: if such a difference existed in the data – or, more specifically, if signatures of lexical processing appeared earlier for low-surprisal than for high-surprisal words -, it is possible that the effects observed for structure building do not reflect modulation of the structure building process by contextual distributional information directly. Instead, such a finding would open the possibility that contextual distributional information affects lexical processing in time, which could in turn affect structure building. However, a comparison between the high- and low-surprisal alternates of a response that has been related to lexical processing (word frequency; Slaats et al., 2023) revealed no temporal differences. In other words, the present analysis provided no evidence for temporal modulation of lexical processing as a consequence of contextual distributional information. This suggests that the temporal dynamics of lexical processing do not directly affect the process of structure building, and makes it more likely that the contextual distributional information directly affects the process of structure building. However, that is not to say that lexical processing does not play a role: it is possible – and even likely – that other differences between processes at the lexical level that are not visible as delays will affect the process of structure building.

Taken together, these results indicate that the contextual probability of a word affects the computation of linguistic structure in time, with structural information being inferred either earlier or faster when a word is expected in a given statistical context. The last two words appear to be quite informative for this process, although longer context distributional information also plays a role.

5.4.3 What & when are not independent

The temporal effects shown in this study are in line with a model proposed by Ten Oever and Martin (2021; 2024). The model situates itself in the framework of neural oscillations serving a functional role in language comprehension. Besides activity in the delta- and gamma bands outlined above, oscillatory activity in other frequency ranges has been suggested to play a key role in language processing, most notably the theta band (Doelling et al., 2014; Ghitza, 2013; Ghitza, Giraud, & Poeppel, 2012), but also the alpha and beta bands (Lam et al., 2016; Zioga, Weissbart, Lewis, Haegens, & Martin, 2023), 2023), giving rise to various theories of the mechanisms underlying oscillations for language (e.g., Brennan & Martin, 2020; Meyer, 2018; Rimmele, Morillon, Poeppel, & Arnal, 2018). An important open question in the formation of these theories is how ongoing oscil-

lations can track language – a signal that is pseudorhythmic rather than purely rhythmic. Ten Oever and Martin (2021) propose that the pseudorhythmicity in speech carries information about the linguistic content. This works as follows. Imagine we are concerned with tracking the word rate. An ongoing oscillator tracks the average word rate. Now, the *phase* of the ongoing oscillation at which a word arrives, carries information about its predictability: if the word arrives early, i.e., before the most excitatory moment in the cycle, the input is likely predictable from the context. On the other hand, if the input arrives relatively late – i.e., after the most excitatory moment in the cycle – the word is likely to be less predictable from the context. This allows the language system of the comprehender to anticipate unpredictable input.

Obviously, the current study does not speak to this directly, as our readout does not provide information about phase, and our study concerns high-level linguistic operations, which are not (yet) explicitly embedded in the model the authors proposed. What the present results do indicate is that contextual information does not only affect the timing of word production, it also affects the timing of higher-level operations. This is much in line with what Ten Oever & Martin suggest; an extension of their proposal, perhaps. Ten Oever & Martin (2021) suggest that a neural population that corresponds to a linguistic unit in the internal language model of an individual (their individually acquired structural and statistical knowledge of language) may be sensitized if that exact linguistic unit is predictable from the context. By consequence, this population may be active earlier, on a less excitable phase of the ongoing oscillation. According to the present results, lexical distributional information does not necessarily activate neural populations that represent lexical information earlier (relative to word onset). A higher lexical probability does, however, more quickly activate neural populations that play a role in representing the syntactic structure underlying the input.

5.4.4 Long- and short-context effects on structure building

Interestingly, most of the temporal delay or shift that we observe in the high- and low surprisal node count responses is captured by the simple trigram models (150 milliseconds). This suggests that local statistical relations between words have a large impact on the process of syntactic structure building. However, not all of the temporal effect is captured by simple trigram surprisal estimates: GPT2-based models suggest that the temporal difference in the response to bottom-up node count can be as large as 190 milliseconds. Why does this difference

exist? We propose that short-context statistical relations are the strongest cue for structure building. At the same time, the short-context statistical relations may be affected by probability in the discourse context. We hypothesize that this causes the larger difference captured by the GPT2-based models: words that are predictable given the larger discourse context. The fine-tuning of GPT2 for Dutch used a context of 128 tokens, which means that surprisal estimates are sensitive to words that appeared less than 128 words ago. Importantly, our stimuli were *fairytales*. This means that they contained words and word sequences that are locally unpredictable, but globally predictable. For example, in one of the stories, the main character is a duckling that can speak ("[...], zei het eendje", which translates to "[...], said the duckling"). We situate these findings with those from Nieuwland & Van Berkum (2006), who show that the discourse context can eliminate N400-effects in sentences with anomalies of animacy (e.g., "the peanut was in love"). Importantly, the fact that GPT2 captures regularity in the global context and humans do, too (and trigram models do not), does not mean that the mechanism underlying this representation is shared or even necessarily similar between GPT2 and humans (Guest & Martin, 2023).

5.4.5 Word duration, surprisal estimate, and response amplitude

Besides the clear finding of a temporal delay as a function of surprisal that does not depend on word duration, and is not a direct consequence of temporal delays at the lexical level as a function of surprisal, we observe a pattern in the response amplitude and explained variance that appears to depend on word duration. The pattern is as follows: before correction for word duration, we observed a larger amplitude *and* larger variance explained for high surprisal than for low surprisal in the trigram models. These differences did not (clearly) exist in the GPT2-models. After correction for word duration, the pattern shifts: there are no clear differences between high- and low surprisal response amplitude and variance explained for the trigram models. In the GPT2-models, however, we find larger response ampitude for *low* surprisal than for high surprisal bottom-up node count responses, and a similar effect on the variance explained. If we group these findings for simplicity, we can conclude that correcting for word duration *decreases* the amplitude for the high-surprisal words. The parallel between response amplitude and variance explained suggests that they are connected: it

is possible that a response explains more of the variance in the signal, if it has a larger amplitude.

These findings confirm that there is a relationship between surprisal and word duration, which has been known for a while (Mahowald et al., 2013; Piantadosi et al., 2011). Beyond this, however, it also suggests that word duration and surprisal *together* drive response amplitudes to higher-level features like syntactic structure building. The present data do not allow us to draw conclusions about this relation, though there are several possibilities for how the factors relate to each other. It is important to keep in mind that lexical surprisal contains influences from different latent factors, one being syntactic structure (Slaats & Martin, 2023). Syntactic predictability has been found to affect the duration of utterances, with less predictable structures yielding longer utterances (Kuperman & Bresnan, 2012; Moore-Cantwell, 2013). It is possible, then, that the duration of the word is itself a cue towards the syntactic structure, and by proxy, it is possible that we have affected the syntactic predictability of the words and constituents in the high- and low surprisal categories. What effect this variable itself should have on the neural response to bottom-up node counts is unclear, although our results suggest that the effect is mostly one of response amplitude, with less predictable, longer words receiving larger amplitudes. Studies with highly controlled stimuli may provide further insight into these relationships. Here, we wish to suggest only that many different aspects of the stimulus, even its duration, likely play a role even in high-level stages of the process of language comprehension (see also Martin, 2016).

5.5 Conclusion

Over the past several decades, much psycholinguistic research has focused on accounting for syntactic phenomena either as a form of transitional probabilities between different linguistic units (e.g., Frank & Bod, 2011; Frank & Christiansen, 2018; Frost et al., 2019; McCauley & Christiansen, 2019), or as a separate level of representation that is hierarchically structured and abstracts away from the lexical items itself (e.g. Brennan & Hale, 2019; Lo et al., 2022; Matchin & Hickok, 2020), without much integration between the two types of linguistic knowledge. In this study, we aimed to test a framework where humans *use* lexical distributional information to build abstract, hierarchical representations that give rise to meaning as an instance of cue-integration. Specifically, we asked whether the low-frequency neural encoding of linguistic structure changes as a

function of the distributional properties of a word, and whether this influence can be linked to probabilities in the immediate context (two preceding words) or rather to probabilities in the larger context (operationalized using GPT2). We did this by extracting delta-band responses to syntactic node count using temporal response functions, and comparing these responses between high- and low surprisal words. Our results showed that lexical distributional information indeed affects the process of syntactic structure building as indexed by delta-band neural responses to node count, and that it did so in the temporal domain: the delta-band response to structure building was delayed by 150 to 190 milliseconds for words that are statistically unpredictable given the context (high surprisal) relative to words that are statistically predictable given the context. This delay appeared not to be driven by temporal changes in lexical processing as indexed by word frequency. In addition, we have shown that most of this effect is captured when using trigram surprisal (150 out of a maximum 190 milliseconds). Our findings speak to theories that model language comprehension as a cascaded process in which cues at different levels are used to infer higher-level representations (Marslen-Wilson & Welsh, 1978; Martin, 2016, 2020), and theories that link abstract linguistic knowledge to the temporal properties of speech (Ten Oever & Martin, 2021, 2024).

5.6 Appendix I. Correlation matrices for feature values

5.6.1 Trigram models

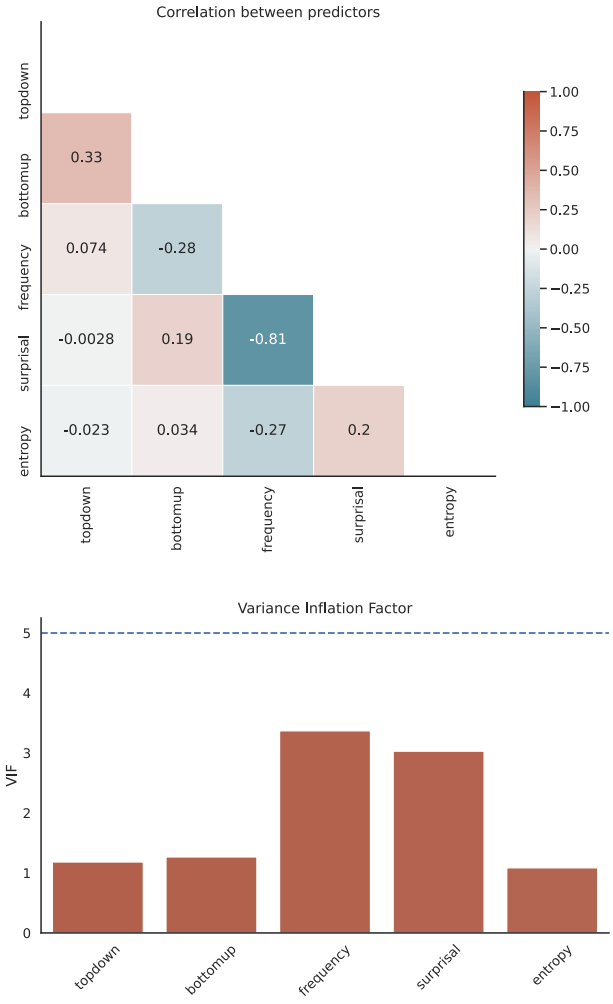


Figure 5.20: Correlation matrix and VIF-values for the trigram values.

5.6.2 GPT2-models

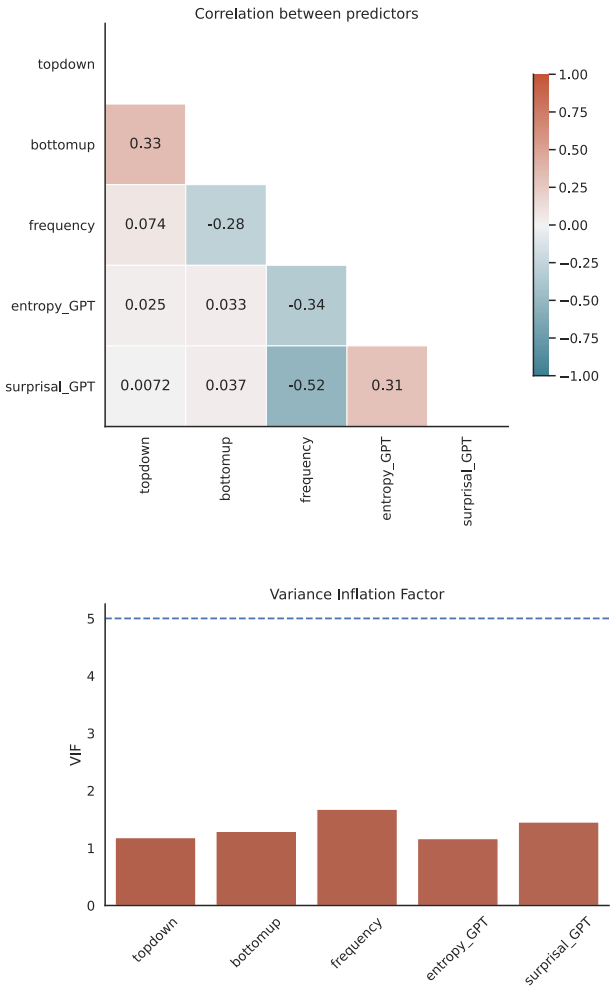


Figure 5.21: Correlation matrix and VIF-values for the GPT2-values.

5.7 Appendix II. Model comparison statistics as output by *step* (LmerTest) from the 'Main effects' analysis

5.7.1 Trigram models

5.7.1.1 Random effects sttructure: 1 + top down * bottom up * surprisal | participant

*Table 5.6: Random effects for structure 1 + top down * bottom up * surprisal | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	2177.04	-4248.07			
topdown * bottomup * surprisal in (1 + topdown * bottomup * surprisal | subject)	0	45	2163.49	-4236.99	27.08	8	$6.84e^{-04}$

*Table 5.7: Fixed effects for random effects structure 1 + top down * bottom up * surprisal | participant.*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy surprisal * topdown * bottomup	1	$7.52e^{-08}$	1	226.28	0.80	0.37
entropy * surprisal * bottomup	2	$1.00e^{-08}$	1	253.98	0.11	0.74
entropy * surprisal * topdown	3	$1.33e^{-08}$	1	232.02	0.14	0.71
surprisal * topdown * bottomup	4	$1.44e^{-08}$	1	40.41	0.15	0.70
entropy * topdown * bottomup	5	$2.43e^{-08}$	1	233.10	0.26	0.61
surprisal * bottomup	6	$6.15e^{-08}$	1	50.27	0.67	0.42
entropy * bottomup	7	$1.20e^{-07}$	1	235.96	1.30	0.26
entropy * surprisal	0	$6.06e^{-06}$	1	236.97	65.60	$2.92e^{-14}$
entropy * topdown	0	$5.33e^{-07}$	1	236.97	5.77	0.02
surprisal * topdown	0	$1.59e^{-06}$	1	24.72	17.19	$3.47e^{-04}$
topdown * bottomup	0	$9.03e^{-06}$	1	43.25	97.85	$1.12e^{-12}$

5.7.1.2 Random effects structure: 1 + top down * bottom up * entropy | participant

*Table 5.8: Random effects for structure: 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	2097.24	-4088.48			
topdown * bottomup * entropy in (1 + topdown * bottomup * entropy | subject)	1	45	2097.08	-4104.16	0.32	8	1.00
topdown * entropy in (topdown + bottomup + entropy + topdown * bottomup + topdown * entropy + bottomup * entropy | subject)	2	38	20967.00	-4118.00	0.16	7	1.00
bottomup * entropy in (topdown + bottomup + entropy + topdown * bottomup + bottomup * entropy | subject)	3	32	2096.57	-4129.14	0.85	6	0.99
topdown * bottomup in (topdown + bottomup + entropy + topdown * bottomup | subject)	4	27	2094.84	-4135.68	3.47	5	0.63
topdown in (topdown + bottomup + entropy | subject)	0	23	2050.64	-4055.27	88.41	4	$2.87e^{-18}$
bottomup in (topdown + bottomup + entropy | subject)	0	23	1893.74	-3741.49	402.19	4	$9.35e^{-86}$
entropy in (topdown + bottomup + entropy | subject)	0	23	2088.14	-4130.27	13.40	4	$9.46e^{-03}$

*Table 5.9: Fixed effects for random effects structure 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$7.52e^{-08}$	1	276.00	0.38	0.54
entropy * surprisal * bottomup	2	$1.00e^{-08}$	1	277.00	0.05	0.82
entropy * surprisal * topdown	3	$1.33e^{-08}$	1	298.59	0.05	0.82
surprisal * topdown * bottomup	4	$2.03e^{-08}$	1	279.00	0.10	0.75
entropy * topdown * bottomup	5	$2.43e^{-08}$	1	298.53	0.09	0.76
surprisal * bottomup	6	$6.15e^{-08}$	1	281.00	0.32	0.57
entropy * bottomup	7	$1.20e^{-07}$	1	282.00	0.62	0.43
entropy * topdown	8	$5.33e^{-07}$	1	283.00	2.77	0.10
entropy * surprisal	0	$6.06e^{-06}$	1	284.00	31.31	$5.19e^{-08}$
surprisal * topdown	0	$2.78e^{-06}$	1	284.00	14.36	$1.84e^{-04}$
topdown * bottomup	0	$1.42e^{-05}$	1	284.00	73.22	$7.39e^{-16}$

5.7.1.3 Random effects structure: 1 + top down * surprisal * entropy | participant

*Table 5.10: Random effects for structure: 1 + top down * surprisal * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	1898.36	-3690.72			
topdown * surprisal * entropy in (1 + topdown * surprisal * entropy | subject)	1	45	1898.31	-3706.63	0.09	8	1
surprisal * entropy in (topdown + surprisal + entropy + topdown * surprisal + topdown * entropy + surprisal * entropy | subject)	2	38	1898.29	-3720.59	0.04	7	1
topdown * entropy in (topdown + surprisal + entropy + topdown * surprisal + topdown * entropy | subject)	3	32	1898.28	-3732.56	0.02	6	1
topdown * surprisal in (topdown + surprisal + entropy + topdown * surprisal | subject)	4	27	1898.17	-3742.34	0.22	5	1
entropy in (topdown + surprisal + entropy | subject)	5	23	1897.68	-3749.35	0.99	4	0.91
topdown in (topdown + surprisal | subject)	0	20	1889.76	-3739.52	15.83	3	$1.23e^{-03}$
surprisal in (topdown + surprisal | subject)	0	20	1893.43	-3746.85	8.50	3	0.04

*Table 5.11: Fixed effects for random effects structure 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$7.52e^{-08}$	1	322.00	0.08	0.77
entropy * surprisal * bottomup	2	$1.00e^{-08}$	1	263.16	0.01	0.92
entropy * surprisal * topdown	3	$1.33e^{-08}$	1	324.00	0.01	0.90
surprisal * topdown * bottomup	4	$2.03e^{-08}$	1	325.00	0.02	0.88
entropy * topdown * bottomup	5	$2.43e^{-08}$	1	326.00	0.03	0.87
surprisal * bottomup	6	$6.15e^{-08}$	1	327.00	0.07	0.79
entropy * bottomup	7	$1.20e^{-07}$	1	328.00	0.14	0.71
entropy * topdown	8	$5.33e^{-07}$	1	329.00	0.60	0.44
surprisal * topdown	9	$2.78e^{-06}$	1	330.00	3.14	0.08
entropy * surprisal	0	$6.06e^{-06}$	1	287.85	6.53	0.01
topdown * bottomup	0	$1.42e^{-05}$	1	287.85	15.27	$1.16e^{-04}$

5.7.1.4 Random effects structure: 1 + bottom up * surprisal * entropy | participant

*Table 5.12: Random effects for structure: 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	2091.65	-4077.30			
bottomup * surprisal * entropy in (1 + bottomup * surprisal * entropy | subject)	0	45	1967.07	-3844.15	249.15	8	2.61e^{-49}

*Table 5.13: Fixed effects for random effects structure 1 + bottom up * surprisal * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	7.52e^{-08}	1	275.92	0.38	0.54
entropy * surprisal * bottomup	2	9.90e^{-09}	1	224.52	0.05	0.82
entropy * surprisal * topdown	3	1.33e^{-08}	1	277.97	0.07	0.79
surprisal * topdown * bottomup	4	2.03e^{-08}	1	278.98	0.10	0.75
entropy * topdown * bottomup	5	2.43e^{-08}	1	279.99	0.13	0.72
surprisal * bottomup	6	5.78e^{-08}	1	110.22	0.30	0.59
entropy * bottomup	7	1.13e^{-07}	1	157.34	0.59	0.44
entropy * topdown	8	5.33e^{-07}	1	282.84	2.77	0.10
entropy * surprisal	0	5.04e^{-06}	1	62.02	26.06	3.39e^{-06}
surprisal * topdown	0	2.78e^{-06}	1	283.81	14.37	1.83e^{-04}
topdown * bottomup	0	1.42e^{-05}	1	283.81	73.27	7.24e^{-16}

5.7.1.5 Best models for every random effects structure configuration & their AIC-value

Table 5.14: Best models for every random effects structure configuration & their AIC-value

Largest model	Chosen model	df	AIC
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * bottomup * surprisal | subject)	r_values ~ entropy + surprisal + topdown + bottomup + entropy * surprisal + surprisal * topdown + topdown * bottomup + (1 + topdown * bottomup * surprisal | subject)	46	-4372.50
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * bottomup * entropy | subject)	r_values ~ entropy + surprisal + topdown + bottomup + entropy * surprisal + surprisal * topdown + topdown * bottomup + (topdown + bottomup + entropy | subject)	19	-4273.29
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * surprisal * entropy | subject)	r_values ~ entropy * surprisal + topdown * bottomup + (topdown + surprisal | subject)	14	-3885.84
r_values ~ entropy * surprisal * topdown * bottomup + (1 + bottomup * surprisal * entropy | subject)	r_values ~ entropy * surprisal + topdown * bottomup + surprisal * topdown (1 + bottomup * surprisal * entropy | subject)	45	-4214.86

5.7.2 GPT2 models

5.7.2.1 Random effects sttructure: 1 + top down * bottom up * surprisal | participant

*Table 5.15: Random effects for structure 1 + top down * bottom up * surprisal | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	2122.21	-4138.41			
topdown * bottomup * surprisal in (1 + topdown * bottomup * surprisal | subject)	1	45	2121.76	-4153.52	0.89	8	1.00
topdown * bottomup in (topdown + bottomup + surprisal + topdown * bottomup + topdown * surprisal + bottomup * surprisal | subject)	2	38	2120.75	-4165.49	2.03	7	0.96
topdown * surprisal in (topdown + bottomup + surprisal + topdown * surprisal + bottomup * surprisal | subject)	3	32	2118.98	-4173.96	3.53	6	0.74
topdown in (topdown + bottomup + surprisal + bottomup * surprisal | subject)	0	27	2062.52	-4071.05	112.91	5	$9.94e^{-23}$
bottomup * surprisal in (topdown + bottomup + surprisal + bottomup * surprisal | subject)	0	27	2113.00	-4172.00	11.96	5	0.04

*Table 5.16: Fixed effects for random effects structure 1 + top down * bottom up * surprisal | participant.*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$5.46e^{-08}$	1	276.00	0.39	0.53
entropy * surprisal * topdown	2	$3.29e^{-09}$	1	295.37	0.01	0.90
surprisal * topdown * bottomup	3	$5.97e^{-09}$	1	300.20	0.03	0.87
entropy * topdown * bottomup	4	$5.48e^{-08}$	1	278.99	0.39	0.53
entropy * surprisal * bottomup	5	$4.59e^{-07}$	1	302.44	2.08	0.15
surprisal * bottomup	6	$1.03e^{-08}$	1	37.23	0.07	0.79
entropy * bottomup	7	$2.98e^{-07}$	1	281.83	2.13	0.15
entropy * surprisal	0	$1.63e^{-06}$	1	282.83	11.60	$7.54e^{-04}$
entropy * topdown	0	$9.27e^{-07}$	1	282.83	6.59	0.01
surprisal * topdown	0	$5.58e^{-07}$	1	282.83	3.97	0.05
topdown * bottomup	0	$1.71e^{-05}$	1	282.83	121.30	$1.04e^{-23}$

5.7.2.2 Random effects structure: 1 + top down * bottom up * entropy | participant

*Table 5.17: Random effects for structure: 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	1999.12	-3892.23			
topdown * bottomup * entropy in (1 + topdown * bottomup * entropy | subject)	0	45	1990.44	-3890.87	17.36	8	0.03

*Table 5.18: Fixed effects for random effects structure 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$5.46e^{-08}$	1	275.97	0.15	0.70
entropy * surprisal * topdown	2	$3.29e^{-09}$	1	272.02	$8.28e^{-02}$	0.93
surprisal * topdown * bottomup	3	$5.97e^{-09}$	1	277.99	0.02	0.90
entropy * topdown * bottomup	4	$5.18e^{-08}$	1	110.16	0.14	0.71
entropy * surprisal * bottomup	5	$4.59e^{-07}$	1	279.02	1.23	0.27
surprisal * bottomup	6	$1.59e^{-08}$	1	280.96	0.04	0.84
entropy * bottomup	7	$2.42e^{-07}$	1	98.84	0.62	0.43
surprisal * topdown	8	$5.58e^{-07}$	1	282.71	1.51	0.22
entropy * topdown	9	$9.31e^{-07}$	1	127.57	2.37	0.13
entropy * surprisal	0	$1.63e^{-06}$	1	283.96	4.40	0.04
topdown * bottomup	0	$1.63e^{-05}$	1	171.87	44.00	$4.10e^{-10}$

5.7.2.3 Random effects structure: 1 + top down * surprisal * entropy | participant

*Table 5.19: Random effects for structure: 1 + top down * surprisal * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	1899.59	-3693.18			
topdown * surprisal * entropy in (1 + topdown * surprisal * entropy | subject)	1	45	1899.48	-3708.96	0.23	8	1.00
surprisal * entropy in (topdown + surprisal + entropy + topdown * surprisal + topdown * entropy + surprisal * entropy | subject)	2	38	1899.47	-3722.94	0.02	7	1.00
topdown * entropy in (topdown + surprisal + entropy + topdown * surprisal + topdown * entropy | subject)	3	32	1899.41	-3734.83	0.11	6	1.00
topdown * surprisal in (topdown + surprisal + entropy + topdown * surprisal | subject)	4	27	1899.15	-3744.3	0.53	5	0.99
entropy in (topdown + surprisal + entropy | subject)	5	23	1897.15	-3748.29	4.00	4	0.41
topdown in (topdown + surprisal | subject)	6	20	1894.13	-3748.27	6.02	3	0.11
surprisal in (surprisal | subject)	0	18	1876.60	-3717.20	35.07	2	$2.42e^{-08}$

*Table 5.20: Fixed effects for random effects structure 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$5.46e^{-08}$	1	322.00	0.06	0.80
entropy * surprisal * topdown	2	$3.29e^{-09}$	1	323.00	$3.72e^{-02}$	0.95
surprisal * topdown * bottomup	3	$5.97e^{-09}$	1	324.00	$6.75e^{-02}$	0.93
entropy * topdown * bottomup	4	$5.48e^{-08}$	1	325.00	0.06	0.80
entropy * surprisal * bottomup	5	$4.59e^{-07}$	1	326.00	0.52	0.47
surprisal * bottomup	6	$1.59e^{-08}$	1	327.00	0.02	0.89
entropy * bottomup	7	$2.98e^{-07}$	1	328.00	0.34	0.56
surprisal * topdown	8	$5.58e^{-07}$	1	329.00	0.64	0.42
entropy * topdown	9	$9.27e^{-07}$	1	330.00	1.06	0.30
entropy * surprisal	10	$1.63e^{-06}$	1	331.00	1.87	0.17
entropy	11	$2.47e^{-06}$	1	332.00	2.83	0.09
surprisal	0	$4.65e^{-05}$	1	23.00	52.93	$2{,}1e^{-07}$
topdown * bottomup	0	$1.71e^{-05}$	1	333.00	19.41	$1{,}43e^{-05}$

5.7.2.4 Random effects structure: 1 + bottom up * surprisal * entropy | participant

*Table 5.21: Random effects for structure: 1 + top down * bottom up * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		53	2085.14	-4064.29			
bottomup * surprisal * entropy in (1 + bottomup * surprisal * entropy | subject)	1	45	2084.99	-4079.98	0.30	8	1.00
surprisal * entropy in (bottomup + surprisal + entropy + bottomup * surprisal + bottomup * entropy + surprisal * entropy | subject)	2	38	2083.37	-4090.75	3.24	7	0.86
bottomup * entropy in (bottomup + surprisal + entropy + bottomup * surprisal + bottomup * entropy | subject)	3	32	2081.03	-4098.06	4.68	6	0.58
bottomup * surprisal in (bottomup + surprisal + entropy + bottomup * surprisal | subject)	4	27	2077.69	-4101.38	6.68	5	0.25
bottomup in (bottomup + surprisal + entropy | subject)	0	23	1895.11	-3744.23	365.15	4	$9.38e^{-78}$
surprisal in (bottomup + surprisal + entropy | subject)	0	23	1980.33	-3914.66	194.72	4	$5.12e^{-41}$
entropy in (bottomup + surprisal + entropy | subject)	0	23	2059.91	-4073.82	35.56	4	$3.57e^{-07}$

*Table 5.22: Fixed effects for random effects structure 1 + bottom up * surprisal * entropy | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
entropy * surprisal * topdown * bottomup	1	$5.46e^{-08}$	1	276.00	0.28	0.60
entropy * surprisal * topdown	2	$3.29e^{-09}$	1	277.00	0.02	0.90
surprisal * topdown * bottomup	3	$5.97e^{-09}$	1	278.00	0.03	0.86
entropy * topdown * bottomup	4	$5.48e^{-08}$	1	279.00	0.28	0.60
entropy * surprisal * bottomup	5	$4.59e^{-07}$	1	280.00	2.36	0.13
surprisal * bottomup	6	$1.59e^{-08}$	1	281.00	0.08	0.78
entropy * bottomup	7	$2.98e^{-07}$	1	282.00	1.53	0.22
surprisal * topdown	8	$5.58e^{-07}$	1	283.00	2.86	0.09
entropy * surprisal	0	$1.63e^{-06}$	1	284.00	8.29	$4.28e^{-03}$
entropy * topdown	0	$9.27e^{-07}$	1	284.00	4.71	0.03
topdown * bottomup	0	$1.71e^{-05}$	1	284.00	86.70	$3.6e^{-18}$

5.7.2.5 Best models for every random effects structure configuration & their AIC-value

Table 5.23: Best models for every random effects structure configuration & their AIC-value

Largest model	Chosen model	df	AIC
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * bottomup * surprisal | subject)	r_values ~ entropy + surprisal + topdown + bottomup + entropy * surprisal + entropy * topdown + surprisal * topdown + topdown * bottomup + (topdown + bottomup * surprisal | subject)	25	-4292.74
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * bottomup * entropy | subject)	r_values ~ entropy + surprisal + topdown + bottomup + entropy * surprisal + topdown * bottomup (1 + topdown * bottomup * entropy | subject)	44	-4041.09
r_values ~ entropy * surprisal * topdown * bottomup + (1 + topdown * surprisal * entropy | subject)	r_values ~ surprisal + topdown * bottomup + (surprisal | subject)	9	-3924.08
r_values ~ entropy * surprisal * topdown * bottomup + (1 + bottomup * surprisal * entropy | subject)	r_values ~ entropy + surprisal + topdown + bottomup + entropy * surprisal + entropy * topdown + topdown * bottomup + (bottomup + surprisal + entropy | subject)	19	-4235.94

5.7.3 Trigram vs GPT2

In the tables below, the variable 'lm' stands for 'language model'.

5.7.3.1 Random effects structure: 1 + entropy * surprisal | participant

*Table 5.24: Random effects for structure: 1 + entropy * surprisal | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		19	985.01	-1932.02			
entropy * surprisal in (1 + entropy * surprisal | subject)	1	15	984.90	-1939.80	0.22	4	0.99
entropy in (entropy + surprisal | subject)	2	12	984.31	-1944.61	1.19	3	0.76
surprisal in (surprisal | subject)	0	10	972.00	-1924.01	24.61	2	$4.54e^{-06}$

*Table 5.25: Fixed effects for random effects structure 1 + entropy * surprisal | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
lm * entropy * surprisal	1	5.49e^{-07}	1	138.00	1.84	0.18
lm * entropy	2	8.78e^{-07}	1	139.00	2.93	0.09
entropy * surprisal	3	1.08e^{-06}	1	140.00	3.56	0.06
entropy	0	4.74e^{-06}	1	141.00	15.34	1.40e^{-04}
lm * surprisal	0	6.69e^{-06}	1	141.00	21.63	7.55e^{-06}

5.7.3.2 Random effects structure: 1 + model * entropy | participant

*Table 5.26: Random effects for structure 1 + model * entropy | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		19	977.36	-1916.71			
lm * entropy in (1 + lm * entropy | subject)	1	15	975.83	-1921.66	3.06	4	0.55
entropy in (lm + entropy | subject)	2	12	974.77	-1925.55	2.11	3	0.55
lm in (lm | subject)	0	10	972.00	-1924.01	5.54	2	0.06

*Table 5.27: Fixed effects for random effects structure 1 + top down * bottom up * surprisal | participant.*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
lm * entropy * surprisal	1	5.49e^{-07}	1	138.00	1.52	0.22
lm * entropy	2	8.78e^{-07}	1	139.00	2.42	0.12
entropy * surprisal	3	1.08e^{-06}	1	140.00	2.96	0.09
entropy	0	4.74e^{-06}	1	141.00	12.77	4.82e^{-04}
lm * surprisal	0	6.69e^{-06}	1	141.00	18.01	3.96e^{-05}

5.7.3.3 Random effects structure: 1 + model * surprisal | participant

*Table 5.28: Random effects for structure: 1 + model * surprisal | participant*

Effect	Elimin. params	npar	logLik	AIC	LRT	df	p value (χ^2)
<none>		19	1023.36	-2008.73			
lm * surprisal in (1 + lm * surprisal | subject)	0	15	999.73	-1969.47	47.26	4	1.35e^{-09}

*Table 5.29: Fixed effects for random effects structure 1 + model * surprisal | participant*

Effect	Elimin. params	Sum.Sq	NumDF	DenDF	F value	p value
lm * entropy * surprisal	0	5.49e^{-07}	1	115.00	4.27	0.04

5.7.3.4 Best models for every random effects structure configuration & their AIC-value

Table 5.30: Best models for every random effects structure configuration & their AIC-value

Largest model	Chosen model	df	AIC
r_values ~ lm * entropy * surprisal + (1 + entropy * surprisal | subject)	r_values ~ lm * surprisal + entropy + (surprisal | subject)	9	-1987.92
r_values ~ lm * entropy * surprisal + (1 + lm * entropy | subject)	r_values ~ lm * surprisal + entropy + surprisal + (lm | subject)	9	-1969.71
r_values ~ lm * entropy * surprisal + (1 + lm * surprisal | subject)	r_values ~ lm * entropy * surprisal + (1 + lm * surprisal | subject)	19	-2007.73

6 | The limits of the Temporal Response Function

Abstract

This Chapter presents an overview of several sets of simulations that provide insight into the possibilities and the limits of the temporal response function as used in **Chapter 3** and **5** of this dissertation. The general aim of the simulations was to assess whether any effects found in the TRF-analyses could be attributable to properties of either the data or the linear model that were unrelated to the theoretical phenomenon under consideration. The simulations revealed the following. Firstly, comparing reconstruction accuracy values between conditions is most reliable when the signals from the two conditions have the same duration and share the same number of responses. Secondly, any differences between frequency bands resulting from band-specific TRF-models can be interpreted reliably. Thirdly, in the case of two different (linguistic) conditions with unbalanced feature values, the same TRF can be extracted if the true responses in the data are indeed identical in the two conditions. Fourthly, and finally, the TRF can capture effects in time (i.e., response delays or time-shifts), but only in a categorical fashion. TRF models with a categorical split can be reliably evaluated for reconstruction accuracy by comparing the systematic categorical split to a random categorical split. When creating such models, it is important to keep in mind that the TRF cannot directly model temporal interactions of a continuous nature as a consequence of being a time-invariant linear system. Instead, the TRF model will capture this temporal effect as noise, and potentially a separate response (e.g., surprisal).

6.1 Introduction

This Chapter serves as a prologue and/or epilogue to some of the Chapters in this dissertation – namely, those that use the forward linear model called the *Temporal Response Function* (TRF) (Chapters 3 and 5). The Chapter summarizes the results of simulations that explore some properties of the TRF which have knock-on effects for the interpretation of these models with respect to theories in psycholinguistics and the neurobiology of language. The goal of these simulations is to assess whether any effects found in the data could be attributable to properties of the data, or the linear model, that are unrelated to the linguistic phenomenon under consideration - in other words, this Chapter is a characterization of the prism through which we may view neural activity during spoken language comprehension. The prism of TRFs allows us to interpret neural activity as a function of various linguistic features, but with this power, also comes potential distortion of the (non-linear) neural signal by the linear model. As such, the results of these simulations are intended to help situate the interpretation, and the strength and limitations therein, of the findings in the rest of the dissertation. In addition, this Chapter provides general guidelines of what is important to keep in mind when designing an experiment for analysis with TRF-models.

The Chapter will provide an answer the following questions. (1) How does the interstimulus interval (ISI) affect the reconstruction accuracy of the TRF model? (2) If a feature enhances reconstruction accuracy in one frequency band, but not the other, does that mean that the response is in this frequency band? (3) Are different feature values able to extract the same TRF waveform? (4) Is the TRF suitable to model interactions between features *in time*? The Chapter has the following structure. In the first section, the model system is described in detail. In the subsequent second, third and fourth section, the questions are addressed with simulations. The fifth and sixth sections provide a summary of the findings and a discussion of the possibilities and limits of the TRF for neurolinguistic research.

6.2 The TRF: model description, estimation & scoring

6.2.1 Model description

The TRF is an analysis technique for examination of data with high temporal precision, such as MEG and EEG data, which has recently been used a lot in the study of language. Being a form of *time-resolved multiple regression*, which makes it highly suitable for the investigation of (naturalistic) language comprehension: the TRF allows for simultaneous modelling of different levels of linguistic representation, such as phonemes, words, and phrases and provide insight into the time-course of the response each of these highly stimulus specific features individually (Brodbeck, Hong, & Simon, 2018; Brodbeck, Presacco, & Simon, 2018; Brodbeck & Simon, 2020; Broderick et al., 2018; Crosse, Di Liberto, Bednar, & Lalor, 2016; Di Liberto et al., 2015; Drennan & Lalor, 2019; Gillis et al., 2021; Hale et al., 2022; Heilbron et al., 2022; Huizeling et al., 2022; Lalor & Foxe, 2010; Lalor, Power, Reilly, & Foxe, 2009; Sassenhagen, 2019; Slaats et al., 2023; Tezcan et al., 2023; Weissbart et al., 2019; Zioga et al., 2023).

Table 6.1: The fitted encoding models in the interaction effects-analyses.

Term	Unit*	Description
β	a.u.	Coefficient estimated with the linear model. Convolves with the stimulus to form the output
y	T	Real neural signal used for coefficient estimation
$\hat{y}$	T	Output of the TRF model. Here: the predicted neural signal
$\bar{y}$	T	Mean of the neural signal
x	(variable)	Input to the TRF model. Here: the stimulus
η	T	Error
τ		Time-lag
t	S	Time in seconds
N		Number of samples of x and y
M		Dimension of y: number of sensors in the measured output signal
P		Dimension of x: number of features
K		Number of discretized lags
dt	s	Time-step between lags, calculated with $\mathrm{dt} = 1/\mathrm{F_s}$
$\mathrm{F_s}$	Hz	Sampling frequency in Hz
X		Matrix of the time-lagged feature time series, vectorized. $X \in \mathbb{R}^{N \times kp}$
λ		Regularization parameter in ridge regression

* Where applicable.

The system we are working with when we create TRF-models is a *linear time-invariant system*. In a linear system, the relationship between the input (x) and the output (y) is a linear mapping: we get the output by multiplying the input with a constant. This constant is the coefficient, or β-weight, that we estimate when we fit a linear model. In a time-invariant system, it does not matter when

a particular input is applied. This system will output *y(t)* when the input is *x(t)*, and given the input *x(t+σ)*, the output will be *y(t+σ)*.

In essence, the temporal response function is a multivariate linear regression approach that estimates a set of coefficients β (the kernel) describing the *linear* relationship between the neural signal and some predictive features – in our case, linguistic annotations of the stimulus. The TRF method uses not just a perfect temporal alignment between the stimulus features and the neural signal; the regression is performed simultaneously over a pre-defined time window of *lags,* essentially aligning the stimulus features to the neural signal at different moments in time: *time-resolved regression*. Instead of a single coefficient β this approach yields a set of coefficients: one for each feature, at each point in time (and, in the case of neural data, at every sensor or source). Together, these coefficients capture a neural response much like an ERP.

In neuroimaging, this linear model is often referred to as an *encoding model* or a *forward model* if the output (y) represents the brain response, and the input (x) the stimulus – in other words, if the model predicts the neural signal using the stimulus information. In contrast, if y denotes the stimulus, and x the brain signal, the linear model is referred to as a *decoding model* or a *backward model.* In this dissertation, I have used exclusively encoding models: I used the stimuli to model the neural data, and not the other way around.

The equation of the TRF model is presented in (1) below. y refers to the output (the neural signal) at time t. β is the set of coefficients (kernel) estimated by the linear model. τ is the time-lag. Brain signals will always carry measurement noise that the model is not capable of representing: they are not (linearly) related to the stimulus features. This variance not captured by the linear model is denoted by η in the equation in 6.1 below.

$$y(t) = \sum_{\tau} \beta(\tau) x(t-\tau) d\tau + \eta(t) \tag{6.1}$$

This equation describes a TRF-model with *one* feature. However, we might perform this procedure with multiple features simultaneously: a *multivariate* regression. Each feature will receive its own kernel (set of β) over which will be summed to produce the output. This yields the following equation in 6.2.

$$y(t) = \sum_{i} \sum_{k} \beta_i(\tau) x_i(t-\tau) d\tau + \eta(t) \tag{6.2}$$

We can do a reduction of this equation by vectorizing the features and concatenating them along the dimension of the features (summation on kernels; the dummy index i above). This reduces the equation to the following:

$$y = X_{lagged}\beta + \eta \tag{6.3}$$

We solve this equation for a given sensor in our M/EEG data. Instead, we can also concatenate these sensor equations along a new dimension. This means that y becomes a matrix, and so does β. X_{lagged} does not change. This is referred to as *multiple* regression: the sensor-specific solutions are independent from each other.

6.2.2 Model estimation & evaluation

Estimating the kernel coefficients from the data is an ill-posed problem, as often in neuroimaging: there are more samples than lags and features: $N > kp$. This means that the system of equation is *over-determined:* there are more equations than variables, and there is not a unique solution. In other words, X is not squared and not invertible. It is possible to circumvent this problem by turning the estimation problem in to an optimization problem, in which we try to find the parameter that leads to the lowest *cost* according to a predetermined function. In our case, we use a least-squares solution. This solution finds β coefficients that minimize the sum of the squares of the difference between the predicted version of y ($\hat{y}$) and the real y, as is shown in equation 6.4 below. This is our *cost function*.

$$J(\beta) = \sum(\hat{y}[n] - y[n])^2 \tag{6.4}$$

In practice, this means that we solve the equation in 6.5 for β (shown in 6.6. This is the *closed-form solution* of the minimization problem.

$$X^T X\beta = X^T y \tag{6.5}$$

$$\beta = (X^T X)^{-1} X^T y \tag{6.6}$$

Often, a feature in X is to some degree correlated with one or more other features (multi-collinearity), or with lagged versions of itself (auto-correlation). Auto-correlation occurs for example in the speech envelope; the value of the next sample is always more similar to the current sample than any randomly

picked sample. This means that if we shift the speech envelope in time just slightly, these two shifted variants of the speech envelope will be correlated to each other. This can pose a problem for our model estimation.

Notice it is necessary to *invert* matrix X^TX to obtain β (the inverted matrix of matrix A is A^{-1}). When two features in X are correlated, the matrix X^TX will have eigenvalues close to zero. This is not a problem in and of itself – the matrix can still be inverted – but inverting the matrix will lead to the small values in the original matrix being very large. Any numerical inaccuracies in the eigenvalues that are close to zero in X^TX will blow up in the inversion, leading to a model that is far from the truth. To avoid this and obtain more reliable estimates of the coefficients, we sum a constant positive value λ to the diagonal of X^TX. This is called *ridge regression*.

We evaluate the model by computing a predicted neural signal ($\hat{y}$) and correlating this with the real neural signal (y). This is done on a held-out portion of the dataset – that is, a set of stimuli and neural responses that were not used to estimate the coefficients. Using the estimated β and a held-out stimulus matrix X, we compute a predicted neural signal $\hat{y}$. Then y and $\hat{y}$ are compared for similarity using the coefficient of determination r^2, which essentially is the division between the explained variance and the total variance (see 6.7) or Pearson's correlation coefficient (in this case the square root of 6.7). Both of these metrics quantify how much variance is explained by the model.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{6.7}$$

6.2.3 General experimental setup

All simulations described below were run in Python 3.8 using the packages SciPy, Scikit.Learn (sklearn), PyEEG, NumPy, Pandas, Matplotlib, Seaborn, MNE-Python and Statsmodels (Gramfort et al., 2013; Harris et al., 2020; Hunter, 2007; Seabold & Perktold, 2010; Virtanen et al., 2020; Waskom, 2021). The code is available on `https://github.com/sslaats/trf-simulations/`. As in many TRF models in the literature, and definitely those in this dissertation, each 'stimulus' feature is a spike-train of zeros with a value (here: sampled from a random distribution) inserted at the desired sample index. Further details are provided in each individual section.

6.3 Simulation set 1: Interstimulus interval & band-pass filtering

In this first set if simulations, we set out to answer questions that arose in the analysis of Chapter 2: namely, whether the interstimulus interval (ISI) *alone* could affect the reconstruction accuracy of the neural signal; and how band-pass filtering can potentially affect our results. To do this, we simulated MEG data with responses at different interstimulus intervals. In addition, we varied the signal-to-noise ratio (SNR). The duration of the signal was kept constant.

6.3.1 Experimental setup

The impulse responses were a multiplication of a hamming window and a sine function. Each of the impulse responses had the same shape, but could differ in duration, i.e., number of samples the response spanned. The responses were generated such that their spectral power was located in the delta band (here: 1-3 Hz) or the theta band (here: 4-8 Hz). The two impulse responses and their spectral power is plotted in Figure 6.1.

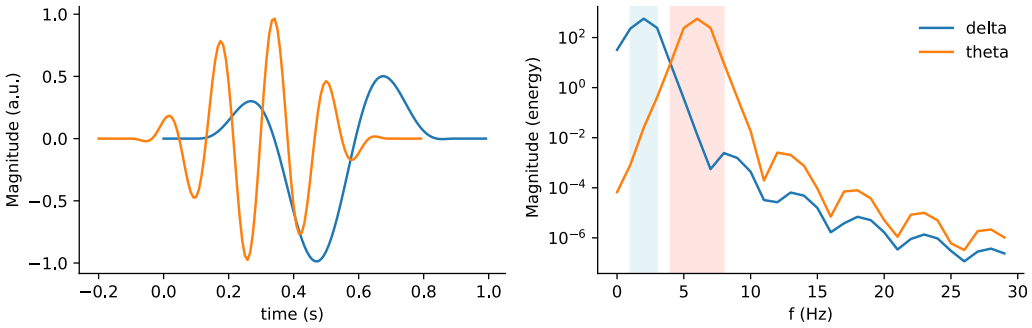


Figure 6.1: The impulse responses. A (left). The impulse responses used to generate the data in this section (ground truth for TRF estimation). B (right). Power spectral density of the impulse responses used in this section. Blue shaded area indicates the delta frequency range used in this section (3-6 Hz); orange shaded area indicates the theta frequency range used in this section (4-8 Hz).

These kernels were convolved with two spike-trains of random values (the 'stimuli'; x), which yielded the signal (y). A white noise component which was scaled according to the standard deviation of the response was summed to the clean signal to yield a signal with the desired signal-to-noise ratio. We generated two instances of both 'x' and 'y'; one served as the training set, and the other

as the test set. We used the TRF estimation and model evaluation approaches described in sections 6.2.1 and 6.2.2 using the generated training- and test sets.

6.3.2 Results: Interstimulus interval & signal length

These simulations were performed with the delta-band kernel exclusively. To evaluate the effect of the ISI, we created instances of x and y with a varying time-window between the spikes, ranging from 100 to 900 milliseconds. Since the signal length was kept constant, the shorter the ISI was, the longer part of the signal that contained exclusively noise at the end. To effectively evaluate whether any effects due to the ISI are caused by the ISI itself, we must disentangle it from the relative amount of noise in the signal. To do this, the models trained on the different ISIs and different SNRs were evaluated twice:

1. With a constant signal length (the full signal)
2. With a 'truncated' signal, where the noise-part at the end of the signal is removed.

The results of this simulation are displayed in Figure 6.2. Figure 6.2A shows that the reconstruction accuracy decreases with the SNR: for an SNR of -3, the reconstruction accuracy value is around 0.45, while the predictive power of the model reaches a ceiling when there is no noise added (the red line). What's striking is that the ISI itself (plotted on the x-axis) does not affect the reconstruction accuracy. Instead, the effect of ISI on the reconstruction accuracy appears to be caused by the length of the signal on which we estimate the reconstruction accuracy: the longer the signal, the lower the reconstruction accuracy. In essence, this effect is similar to the effect of SNR. After all, all signals contain the same number of responses, meaning that the longer signals (as caused by the longer ISI) contain a larger portion of signal that contains only noise than shorter signals. Indeed, as can be seen in figure 6.2B, the reconstruction accuracy is constant for a noiseless signal, despite the signal length being a function of the ISI.

In Chapter 2, the results revealed a higher reconstruction accuracy for word lists – the signal with a longer ISI – than for the sentences – the signal with a short ISI. These simulations were conducted to evaluate whether this difference could be caused by the difference in ISI. Based on these simulations, we can safely conclude that any differences in reconstruction accuracy due to ISI should drive the effects in the opposite direction: the overall reconstruction accuracy should be lower for word lists than for sentences.

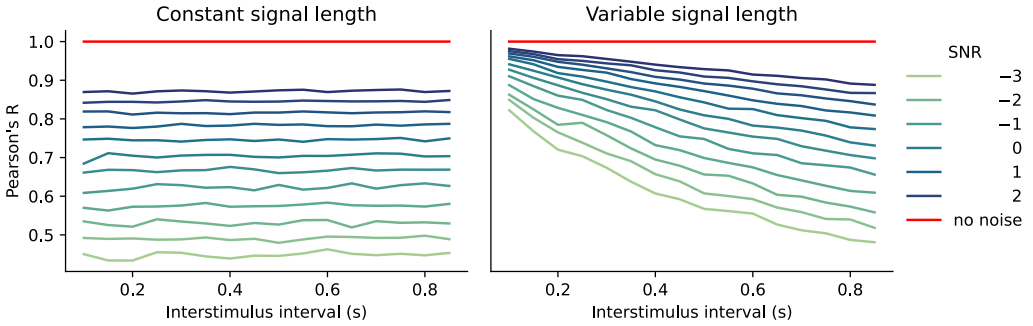


Figure 6.2: Reconstruction accuracy & SNR. A (left). Reconstruction accuracy values for different signal-to-noise ratios at interstimulus intervals from 0.1 to 0.9 seconds. The signal length is kept constant. B (right). Reconstruction accuracy values for different signal-to-noise ratios at interstimulus intervals from 0.1 to 0.9 seconds. The signal is truncated at the end of the last impulse response, meaning that the signal length increases for larger interstimulus intervals.

6.3.3 Results: Filtering

In this simulation, we wished to evaluate whether an increase in reconstruction accuracy in a specific frequency band can be interpreted as reflecting that the response captured by the feature has spectral power in this frequency bands. This is related to another finding from Chapter 2, namely that the response to word frequency increases reconstruction accuracy more in the delta band than in the theta band. Does this mean that the spectral energy of the response is indeed located in the delta band?

To evaluate this question, we simulated an electrophysiological signal (M/EEG-like) with two stimulus-dependent responses that had most spectral power in two neighboring frequency bands: delta and theta (see Figure 6.1). These responses were driven by separate stimuli: one stimulus predicted the delta-band response – we will call this the 'delta-stimulus' –, and the other one the theta-band response – the 'theta-stimulus'. We filtered the signal into the delta- and the theta band (1-3Hz and 4-8Hz for simulation purposes) using a FIR-filter. Of each band-pass filtered signal, we estimated several TRF models: a model that contained only the mismatching feature (i.e., the theta-stimulus in the case of a delta-band filtered signal, and the other way around); a model that contained the correct feature (i.e., the delta-stimulus in case of a delta-band filtered signal); and a model that contained both of these features.

We then estimated the relative increase in reconstruction accuracy that was driven by the addition of the *other* feature. This means that in the case of

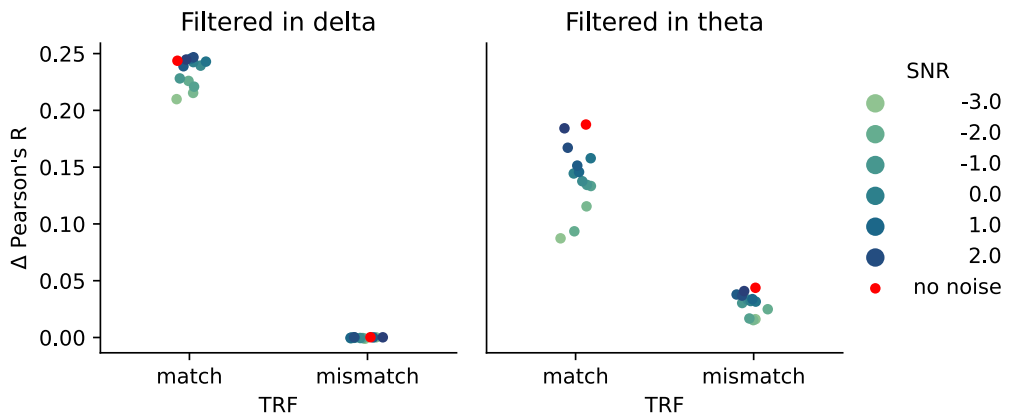


Figure 6.3: Reconstruction accuracy and band-pass filtering. Increases in reconstruction accuracy when adding a feature that predicts a response that has spectral power in the frequency spectrum of the data ('match') versus when adding a feature that predicts a response that does not have power in the frequency spectrum of the data ('mismatch'). Left: filtered in the delta-band (1-3 Hz). Right: filtered in the theta band (4-8 Hz).

the delta-band filtered signal, we computed the difference between the delta-stimulus only model and the both-stimuli model to get an estimate of the benefit from adding the theta-stimulus feature (the mismatch). Similarly, we computed the difference between the theta-stimulus only model and the both-stimuli model to get the benefit from adding the delta-stimulus feature (the match). These values are displayed in Figure 6.3 below.

We observed that the increase in reconstruction accuracy is indeed related to the frequency band in which the added response has most spectral power. That is, if a response has most spectral power in the delta band, adding the feature for this response to a model of the delta-band filtered signal, the reconstruction accuracy will improve; this is not the case if we add the feature for a model that has more spectral power in a neighboring frequency band. Any small improvement observed may be caused by the small overlap between the responses (see Figure 6.1).

6.3.4 Conclusions

In the first set of simulations, we evaluated (1) the effect of the interstimulus interval; and (2) the potentially artefactual effects of our band-pass choices. The simulations showed the following: ISI does not have an effect on reconstruction accuracy; rather, the signal-to-noise ratio does, both in terms of the amplitude

of the signal relative to the noise, and in terms of the number of samples that do not contain a signal but do contain noise. Because of this, if the signal length is different between conditions, a larger interstimulus interval will lead to lower reconstruction accuracy. This relationship between signal length and reconstruction accuracy can, thus, end up distorting the estimation of the contribution of a given cognitive process to neural activity, inasmuch as there may be differences in signal length by condition due to time, timing, or temporal resolution and dynamics. This constraint on estimation does not mean that the interstimulus interval does not affect how the brain handles the input; this may still be relevant, and is discussed in more detail in Chapter 3 itself. But differences in ISI do not alone, or by default, result in differences in reconstruction accuracy.

Secondly, our simulations showed that band-pass filtering itself does not cause the differences in reconstruction accuracy; properties of the responses do. Specifically, the frequency band which has the most power in the response – drives the reconstruction accuracy of the neural signal. This means that if we find improvements for the addition of a given feature in the delta band, the response captured by the feature has power in this frequency band. Of course, in contrast to in this simulation, in actual data we do not know beforehand in which frequency band the effects will show up. Finding effects on the reconstruction accuracy in the delta band means that a response is captured reliably in this frequency band.

6.4 Simulation set 2: Feature values

In this second set of simulations, we address some concerns related to the analysis performed in Chapter 5. In this Chapter, we split features into two separate features on the basis of a second variable. Specifically, in order to examine the effect of distributional measures on the neural response to syntactic information (node count), we split the syntactic features into two on the basis of the median of the distributional measures. Doing so yields a syntactic feature for words that are low in entropy (or surprisal), and one for words that are high in entropy (or surprisal). While this works in theory, the feature values in both groups do not have the same mean. If we then find differences between the two responses, does this mean that the responses themselves are different, or is this a result of the different feature values? Given the properties of the linear model, the responses should be extracted and reconstructed the same. In this small set of simulations, we show that this is indeed the case. We do this by simulating a response and a feature and splitting the feature in two such that the two resulting

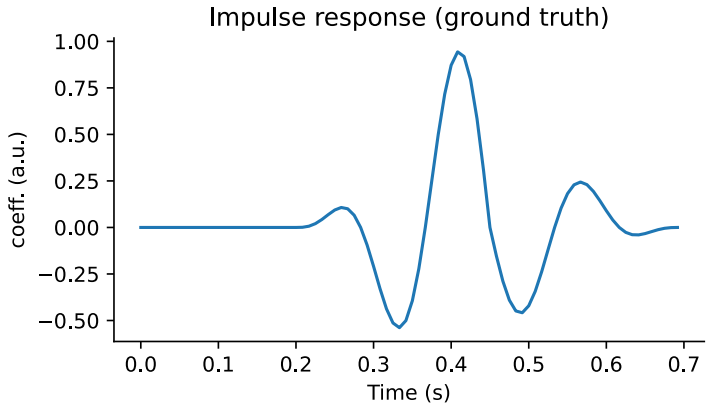


Figure 6.4: The impulse response used to create the data in this section (ground truth for TRF estimation).

features either do or do **not** have the same mean or standard deviation. In this way, we can observe the effects on TRF estimation (i.e., the shape and timing of the estimated response) and the reconstruction accuracy.

6.4.1 Experimental setup

We created an impulse response and two sets of 100 stimulus values (the features). To generate the data, we convolved the impulse response with one of the sets of stimulus values. The impulse response was a kernel generated by multiplying a Hanning window of 500 milliseconds with a sine wave. The impulse response is the 'ground truth': this is the response in the data that we want to extract using the TRF. The two sets of stimulus values were generated according to the effect we wanted to simulate and is described in more detail below. The ISI was 600 milliseconds. At this point, we did not add noise to the data, meaning that time points that do not contain a response are equal to zero. The data had a total length of 7320 samples, i.e. approx. one minute of data at 120 Hz sampling frequency. The impulse response is visible in Figure 6.4.

6.4.2 Results

Splitting a feature such that the two resulting sets of feature values are identical should not lead to different TRFs or reconstruction accuracies. We checked this as a baseline. To assure this, we created one set of 100 normally distributed random values (set A), and two identical sets of 50 normally distributed random

values (set B1 and B2). Set B1 was aligned to all the values above the median from set A, and set B2 to all the values below the median from set A.

Firstly, we used the full feature of sets B1 and B2 combined and computed the TRF on 80% of the data. The feature values and the resulting TRF are plotted in the top row of Figure 6.5. The TRF captured the response perfectly. Secondly, we created a model with two separate features (B1 and B2) and, again, computed the TRF. In the bottom row of Figure 6.5 the two sets of feature values are displayed, as are the TRFs. Observe that the TRFs overlap perfectly. In other words, the response is extracted efficiently in both cases. (This should not come as a surprise.) In both cases, we used the held-out 20% of the data to compute the reconstruction accuracy. This is done by convolving the resulting TRF with the held-out feature values and correlating the resulting signal to the actual data. In both cases, the reconstruction accuracy was 1.0, the highest value – meaning that the signal was perfectly reconstructed. (N.B., this is only possible because the signal was clean; consider the red lines in Figure 6.5.)

To assess whether the feature values affect TRF estimation, we made a minor change: B1 and B2 are no longer identical. Both are normally distributed, but in set B2 the standard deviation and mean are increased by 3.5 and 4.02, respectively. The values from set B2 were aligned to the values below the median from feature A (i.e., 'low surprisal node count'), and the values from set B1 were aligned to the values above the median from feature A (i.e., 'high surprisal node count'). Again, we first computed the TRF using the full B feature. As is shown in the top row in Figure 6.5 and the blue bar in Figure 6.6, here, too, the impulse response was captured perfectly. When we computed the TRF for B1 and B2 separately, we observed that there was no difference here, either. As before, the reconstruction of the signal was perfect in both cases.

From these simulations we can safely conclude that a difference between the distributions of a feature does not affect the TRF estimation, nor the reconstruction of the signal. This means that if we observe differences between 'B1' and 'B2' in our study, this is *not* due to the distributions of B1 and B2 being different; rather, it will be due to an interaction between factors A and B.

To test whether an interaction between factors A and B will *indeed* be captured by a split feature, we repeated the two simulations above. In this simulation, the true response (ground truth) was driven by an interaction between factors A and B. As before, we evaluate what happens when we model the response using factor B intact, or when we split factor B into B1 and B2 dependent on the values

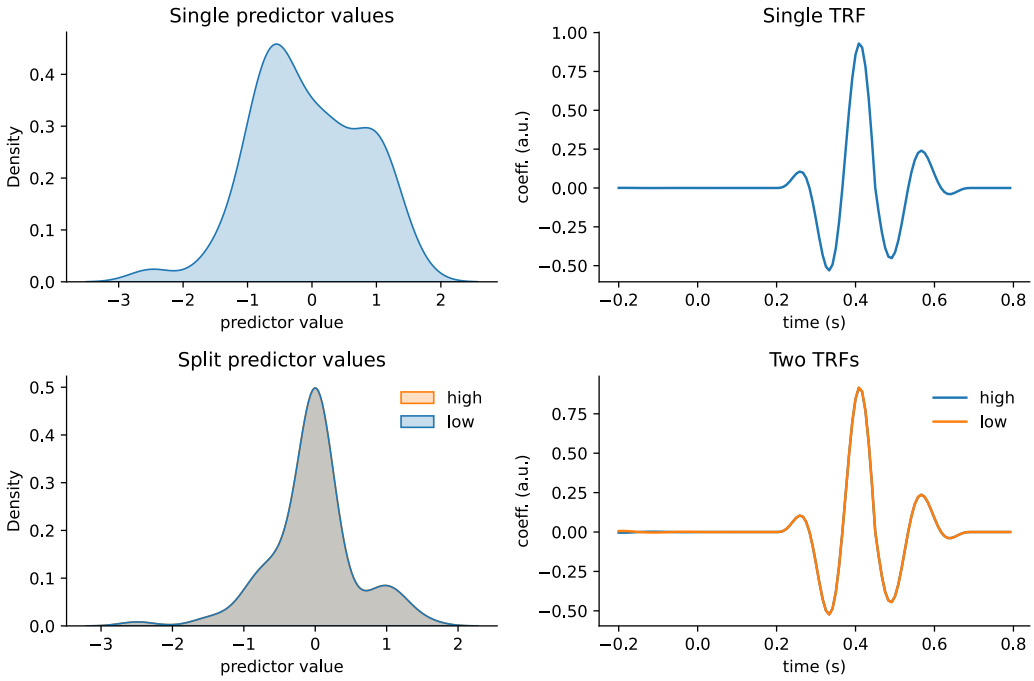


Figure 6.5: Feature B does not interact with feature A. A (top left). Kernel density plot of all feature values when combined into a single predictor. B (top right). TRF extracted using a single feature. C (bottom left). The same feature values as in (A) divided over two distributions according to Feature A such that the resulting distributions are identical (hence the gray color). D (bottom right). TRFs extracted using the two features presented in (C). Each of these features contained half of the values included in (A). Notice that the two TRFs overlap perfectly.

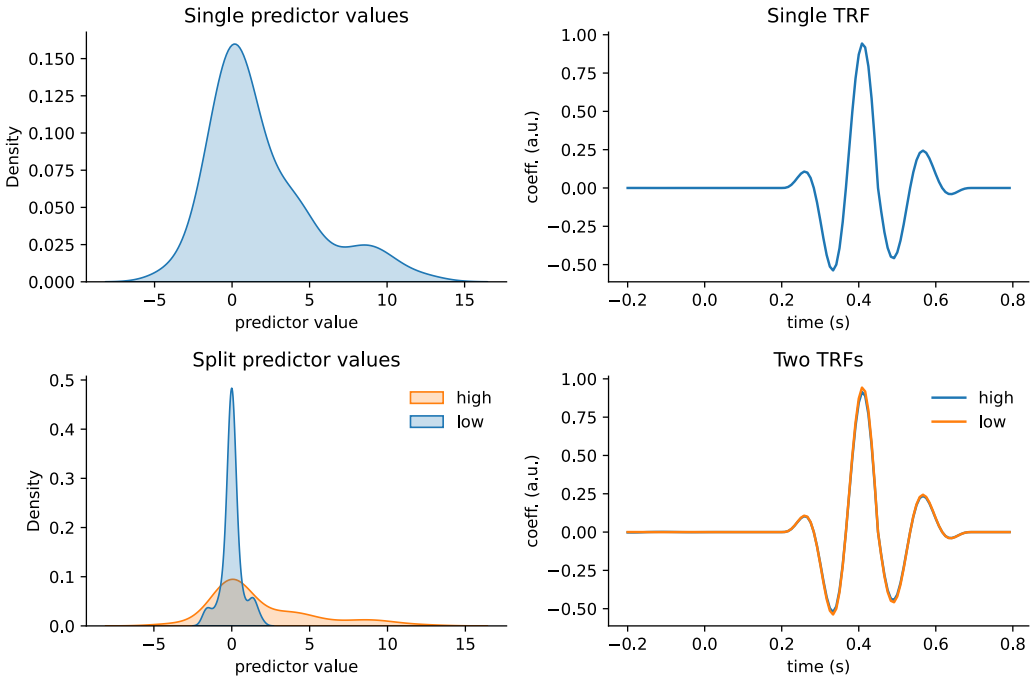


Figure 6.6: Feature B does not interact with feature A. A (top left). Kernel density plot of all feature values when combined into a single predictor. B (top right). TRF extracted using a single feature. C (bottom left) The same feature values as in (A) divided over two distributions according to feature A such that the resulting distributions are different from each other in their mean and standard deviation. D (bottom right) TRFs extracted using the two features presented in (C). Each of these features contained half of the values included in (A). Notice that the two TRFs overlap perfectly.

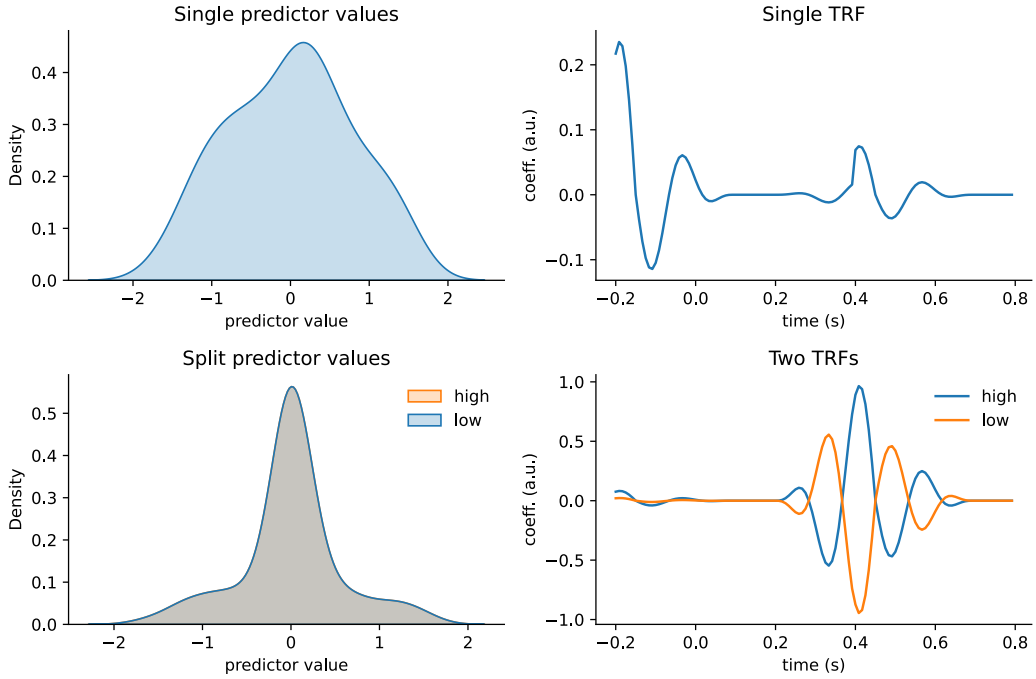


Figure 6.7: Feature B interacts with feature A. A (top left). Kernel density plot of all feature values when combined into a single predictor. B (top right). TRF extracted using a single feature. C (bottom left). The same feature values as in (A) divided over two distributions according to feature A such that the resulting distributions are identical (hence the gray color). D (bottom right). TRFs extracted using the two features presented in (C). Each of these features contained half of the values included in (A). Notice that the two TRFs differ from each other, even though the feature values do not.

from factor A. We do this for both identical and different distributions for B1 and B2.

We observe that the signal is reconstructed (much) better when we use a split predictor to capture (part of) the interaction (reconstruction accuracy intact: -0.40; split: 0.70). On the TRF waveform we also observe that the interaction is indeed captured by the split predictor (see Figure 6.7D below). As for the first two simulations, the observations are identical when the distributions underlying B1 and B2 are different (see Figure 6.8).

6.4.3 Conclusions

To sum up, from these simulations we can conclude that any differences between the TRFs for our high- and low node counts in Chapter 5 are caused by actual

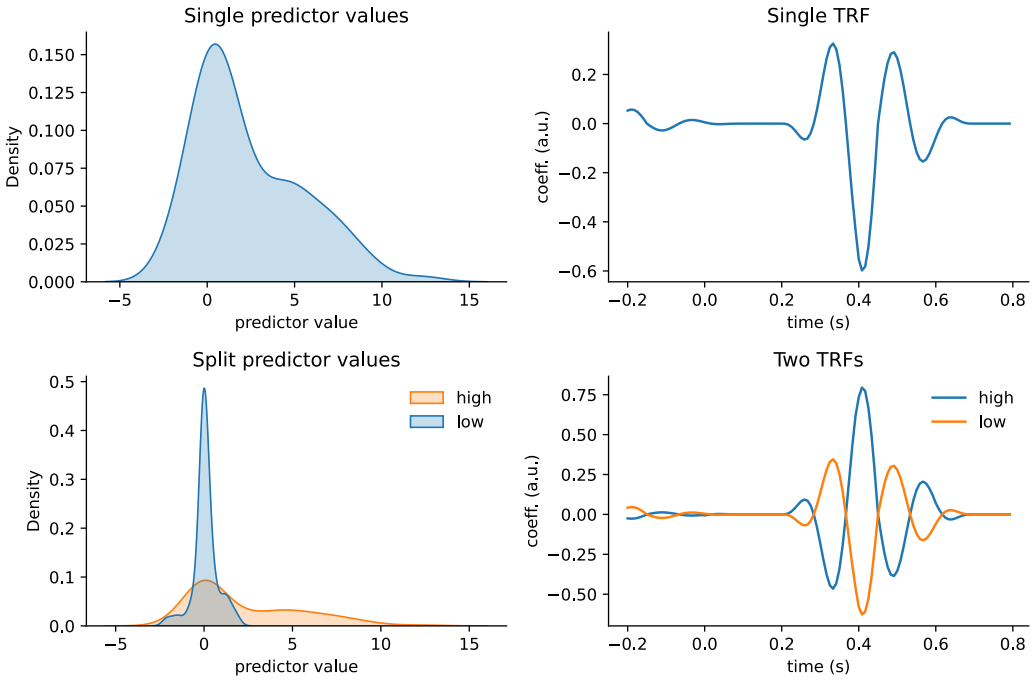


Figure 6.8: Feature B interacts with feature A. A (top left). Kernel density plot of all feature values when combined into a single predictor. B (top right). TRF extracted using a single feature. C (bottom left). The same feature values as in (A) divided over two distributions according to feature A such that the resulting distributions differ from each other in their mean and standard deviation. D (bottom right) TRFs extracted using the two features presented in (C). Each of the features contained half of the values included in (A). Notice that the two TRFs differ from each other.

differences in the data, and are not the result of different feature values when dividing the node counts over two separate features.

6.5 Simulation set 3: Timing effects

The third set of simulations concerns other questions that arose during the analysis presented in Chapter 5 The analyses revealed effects on the TRF-waveform, suggesting an interaction between factor A (surprisal/entropy) and factor B (node count) as per the simulations in set 2. A large difference between the simulations in set 2 and the findings in the data is that the interaction effect modelled in set 2 is an interaction of *amplitude*: the amplitude (and the sign) of the response to feature B is dependent on the values from feature A. In the data, however, there appears to be rather an effect of *latency*. Being a time-invariant system, the TRF model cannot capture an interaction between features that occurs in time. In this set of simulations, we aim to investigate what effects of splitting a predictor has on the reconstruction of the signal when the latency of the response to feature B is affected by feature A; and how such a latency effect can show up in different models made of the data.

6.5.1 Experimental setup

To this end, we model a signal using a kernel that is driven in amplitude by feature B (a normal distribution with a mean of 5 and a sd of 2; mimicking *node count*) and feature A, which drives the onset latency (a log-normal distribution with a mean of 2.3 and a sigma of 0.41; mimicking *surprisal*). The kernel had a width of 500ms or 800ms. An example of such a set of responses is shown in Figure 6.9 below. The interaction works as follows: the response is moved forward (i.e., to the right) the number of samples that equals the surprisal value. I.e., if surprisal is 17, 17 zeros are added to the onset of the response. To examine the effect of the strength of the latency shift, we multiplied surprisal with 20 multiplication factors linearly spaced between 0 (no influence of surprisal) and 2 (latency = 2 x surprisal in samples). For each of these multiplication factors, we modeled 1000 responses at a sampling rate of 120 Hz. For statistical comparison between conditions, we randomly generated these distributions 1000 times.

To be able to gain insight into how *categorically* splitting a feature of which the timing of the response is *parametrically* dependent on the other variable affects the reconstruction accuracy of a model, we estimated four TRF models:

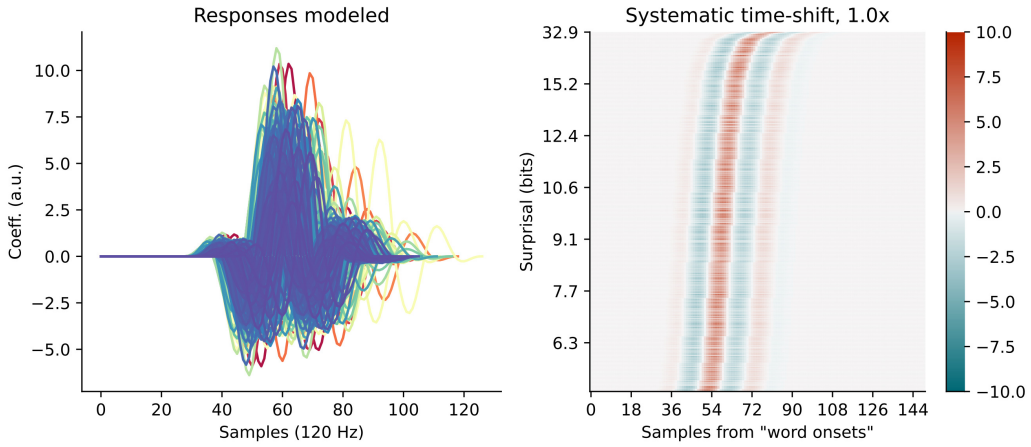


Figure 6.9: Example data for simulation set 3. A (left): Example of the responses in the data (y) for a kernel width of 500ms and a multiplication factor of 1.05. The amplitude of the response is a function of feature B 'node count'; the latency of the response (visible as temporal jitter) is a function of feature A 'surprisal' multiplied by 1.05. B (right): the same responses as in A, ordered by surprisal value.

(1) feature B, 'node count' – the feature that drives the response; (2) feature B 'node count' and feature A 'surprisal' – the feature that drives response latency; (3) feature B1 ('node count – high surprisal') and feature B2 ('node count – low surprisal'); and (4) a random split of feature B 'node count'. The difference between models (1), (2) and (3) on the one hand, and (1), (2) and (4) on the other hand show how including two features with half the number of values affects reconstruction accuracy. At the same time, the difference between models (3) and (4) will tell us whether the systematic split of the feature will lead to a better description of the signal when compared to a model with the same number of features and number of values per feature, but distributed randomly. In other words, this will tell us whether the split contains information (as it should), without the confound of having features with a different number of values.

In addition to these four models, we wished to test whether the 'surprisal' feature will add to a description of the signal, despite not being the direct cause for a response itself. To this end, we estimated models (5) with features B1, B2, and feature A 'surprisal' and (6) with a random split of feature B 'node count' and feature A 'surprisal' in addition to (2), for comparison with models (1), (3) and (4). Furthermore, we were interested in the general ability of TRF models to capture interactions in time with an interaction term, so we created also model (7) which contained an interaction term between features A and B; and model (8) which contained feature B, A, and the interaction term. All models are

summarized in table 6.2. All of these models were fitted to the data as described above, and to a version of the data in which the latency shift was not determined by surprisal, but by another, unmodeled feature that was drawn from the same log-normal distribution as the surprisal feature.

Table 6.2: TRF models and the included features in Simulation set 3.

	Model name	Features
1	Node	*Node count* (feature B)
2	Node + surprisal	*Node count* (feature B) + *surprisal* (feature A)
3	Node split	$Node\ count_{high}$ + $node\ count_{low}$ (node count split by median of surprisal)
4	Node random split	$Node\ count_{random}$ + $node\ count_{random}$ (node count split randomly)
5	Node split + surprisal	$Node\ count_{high}$ + $node\ count_{low}$ + *surprisal*
6	Node random split + surprisal	$Node\ count_{random}$ + $node\ count_{random}$ + *surprisal*
7	Interaction	*Node count * surprisal*
8	Node + surprisal + interaction	*Node count + surprisal + node count * surprisal*

6.5.2 Results

The observations were as follows. We first wanted to know whether splitting feature B – node count - on the basis of the median of feature A – surprisal – can capture the effect of latency on the TRF waveforms. Indeed, this appears to be the case. We observed that splitting feature B according to the median of feature A – which drives the latency of the response – revealed the temporal shift on the TRF waveform. In addition, the size of the latency difference is reflected in the temporal shift expressed by the TRFs, as can be seen in Figure 6.10 below.

Secondly, we looked at the effect of splitting feature B according to the median of feature A on the reconstruction accuracy. As is displayed in Figure 6.11A, doing so increased the reconstruction accuracy relative to most models, but not all (at all multiplication values). We also fit all models on data in which the interaction between surprisal and node count did not exist; rather, node count interacted with a third, unmodelled variable. The results of this are presented in Figure 6.11B. As can be observed from the plot alone, the 'split' models perform better than most models. The only model that outperforms the split models when the influence of surprisal is relatively small, is the model that contains node count, surprisal, and a multiplicative interaction (i.e., node count $\times$ surprisal). This is clearly represented in Figure 6.12 below: the t-values of the contrast between the Node split model and the Node + surprisal + interaction model are negative at first, and later turn positive. This means that the Node + surprisal + interaction model outperforms the Node split model at first, and that this relationship later reverses.

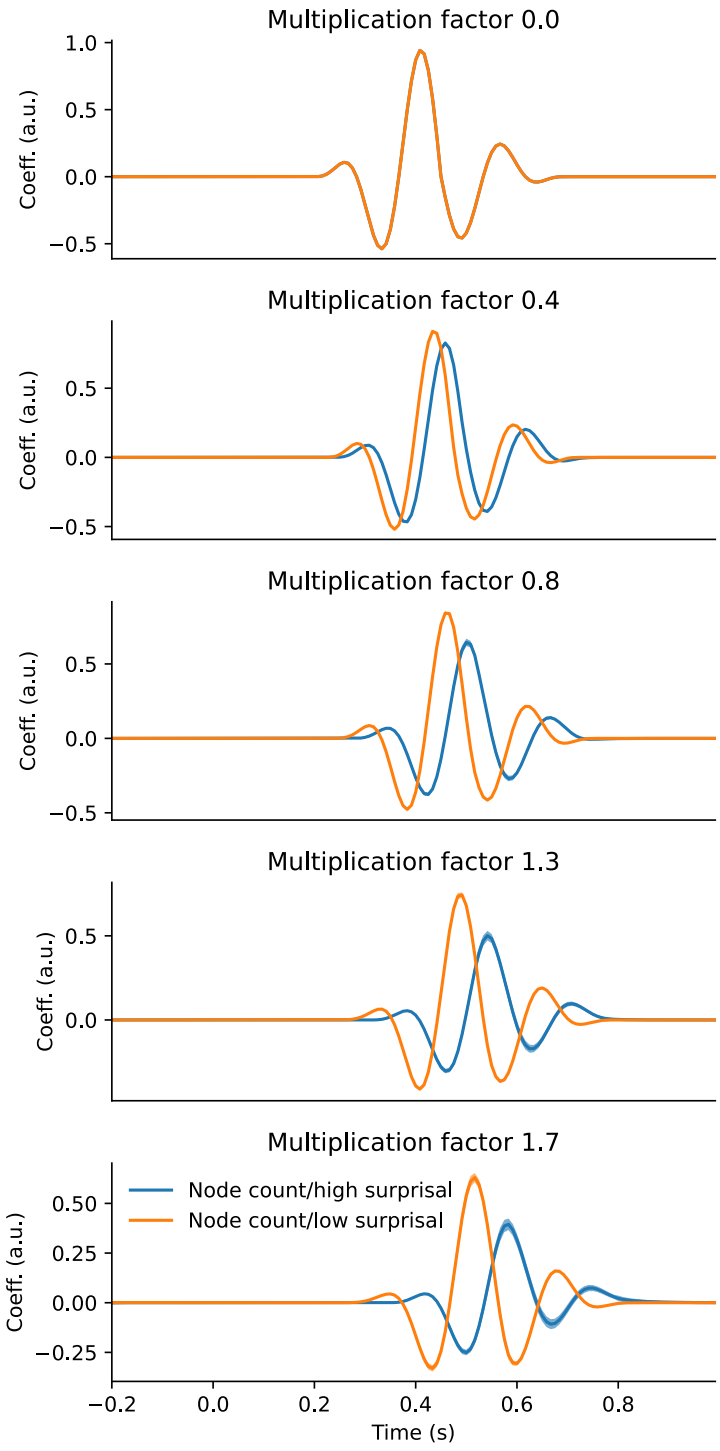


Figure 6.10: Average TRFs extracted using model 3 (feature B1 and B2) for various multiplication factors. High surprisal node count TRF in blue, low surprisal node count TRF in orange. Standard deviation across the 1000 repetitions is shown in shaded areas; however, this was so small that it is barely visible.

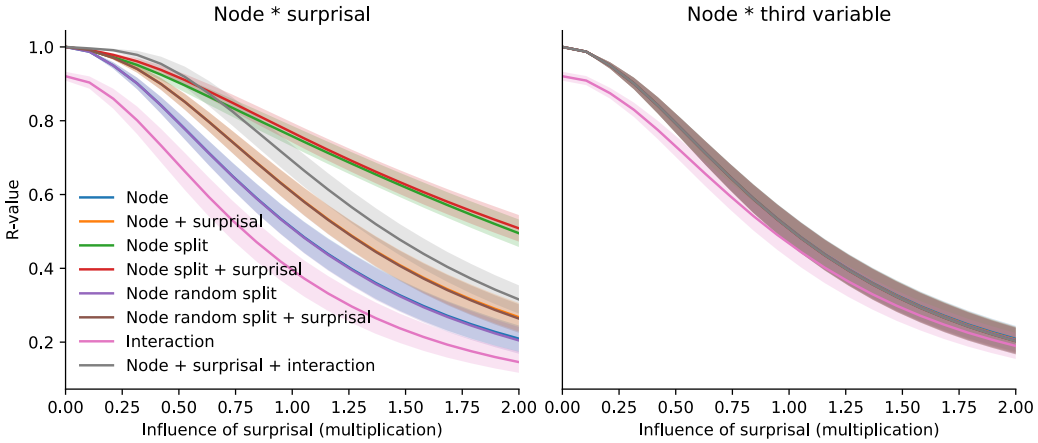


Figure 6.11: Reconstruction accuracy values for temporal interaction. A (left). Reconstruction accuracy for data that contains a temporal interaction between 'node count' and 'surprisal'. B (right). Reconstruction accuracy for data that does not contain a temporal interaction between 'node count' and 'surprisal'; rather, the temporal interaction is between 'node count' and a third, unmodelled, feature.

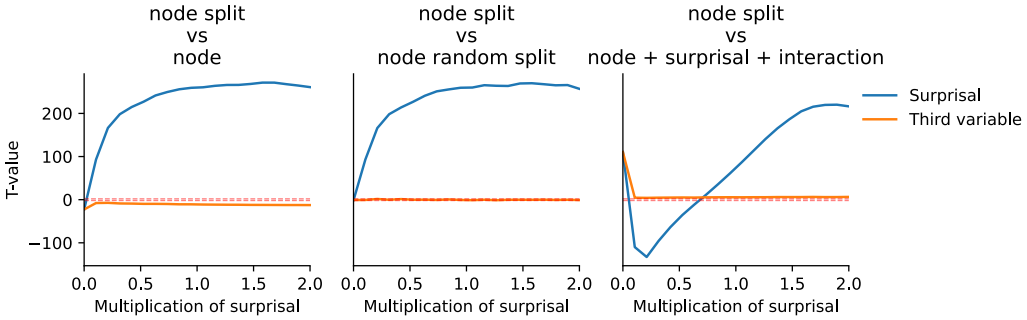


Figure 6.12: Model comparison for temporal interactions. T-values for the comparison of the Node split model against the Node model, the Node random split model, and the Node + surprisal + interaction model when fit to (1) the data in which surprisal determined the latency shift (in blue; left panel from Figure 6.11) and (2) the data in which a third, unmodelled variable determined the latency shift (in orange; right panel from Figure 6.11). The dotted red line marks the alpha-values of 1.96 and -1.96 – i.e., the significance threshold in a simple t-test.

Of course, each of these models is an *approximation* of the data. With this in mind, it becomes logical that several aspects of the data affect which model performs best when it comes to reconstructing the data. We repeated the simulations with a longer kernel – of 800 milliseconds – and noticed that the influence of surprisal needs to be larger for the Node split model to outperform the large

Node + surprisal + interaction model. This suggests that the effect size of surprisal (as represented by the multiplication) relative to the response duration is what drives part of the reconstruction accuracy patterns. The results from the analysis with the 800ms-kernel are displayed in Figures 6.13 and 6.14.

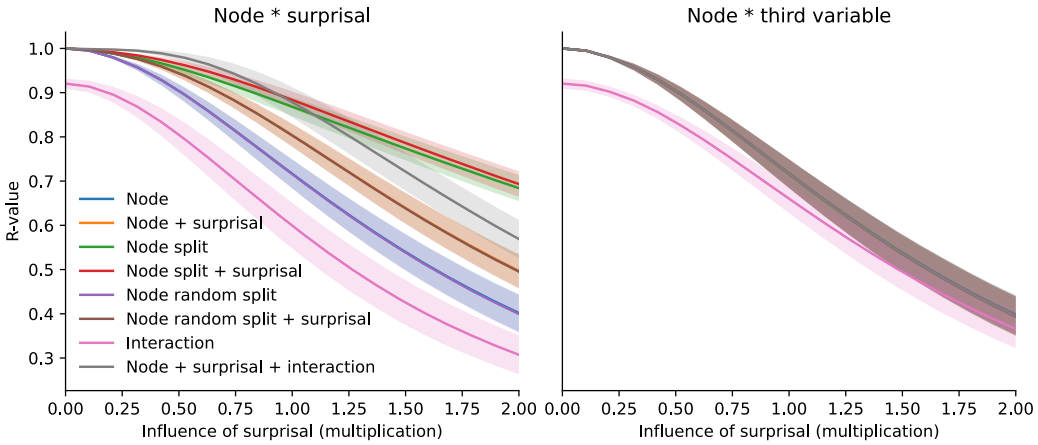


Figure 6.13: Kernel: 800ms. A (left): reconstruction accuracy for data that contains a temporal interaction between 'node count' and 'surprisal'. B (right): reconstruction accuracy for data that does not contain a temporal interaction between 'node count' and 'surprisal'; rather, the temporal interaction is between 'node count' and a third, unmodelled, feature. Reconstruction accuracy values for model 7 "Interaction" not shown; see Figure 6.16 below.

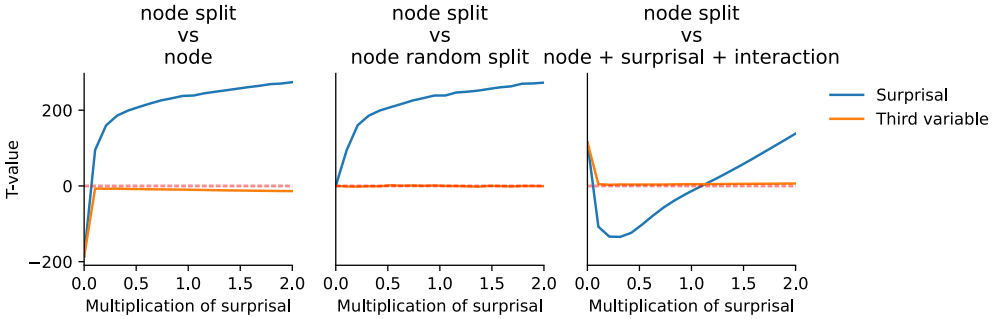


Figure 6.14: Kernel: 800ms. T-values for the comparison of the Node split model against the Node model, the Node random split model, and the Node + surprisal + interaction model when fit to (1) the data in which surprisal determined the latency shift (in blue; left panel from Figure 6.11) and (2) the data in which a third, unmodelled variable determined the latency shift (in orange; right panel from Figure 6.11). The dotted red line marks the alpha-values of 1.96 and -1.96 – i.e., the significance threshold in a simple t-test.

There are likely many other variables that contribute to whether a model with a split predictor will outperform models with intact features. This indicates the interaction in time may not show up as an increase of the reconstruction accuracy depending on the model that is chosen. A model with an interaction may better capture the patterns in the data, but that does not mean that there is no interaction in time. On the other hand, the difference between a random split and a systematic split (i.e., using the second variable to determine when a value is part of feature B1 or B2) always leads to an increase in reconstruction accuracy – if the interaction is in the data. The comparison between a random split feature and a systematic split is therefore a reliable statistical baseline to check whether a temporal interaction exists.

Thirdly, we consider the effect of adding the temporal modulator 'surprisal' to our models. We observe that generally, if a temporal modulation exists, surprisal does increase reconstruction accuracy (Figure 6.15, left and middle panels). In fact, when we consider the TRFs from the strongest model (surprisal influence = 2.0), displayed in Figure 6.15 below, we can appreciate that the 'surprisal' response is a (smaller) phase-shifted response from the node-count. This can have important consequences for how we interpret our models, as will be discussed further below.

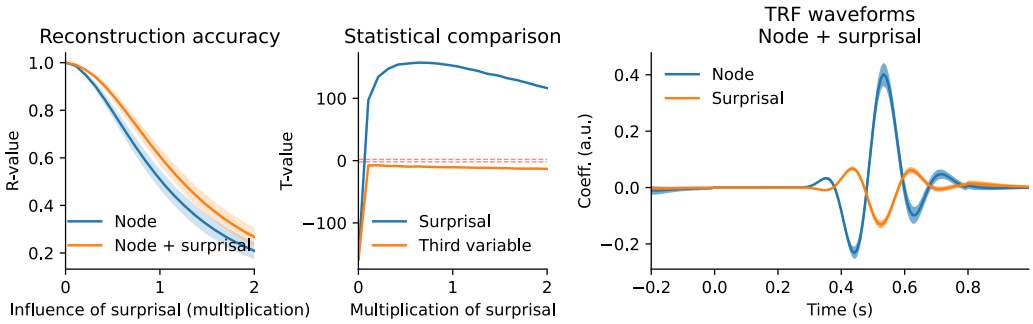


Figure 6.15: A (left). Reconstruction accuracy values for the models Node and Node + surprisal. B (middle). R-values for the contrast Node + surprisal model / Node model when fit to (1) the data in which surprisal determined the latency shift and (2) the data in which a third, unmodelled variable determined the latency shift (in orange). C (right). TRFs extracted using the Node + surprisal model for multiplication = 2.0 (maximal influence of surprisal).

Fourthly, the interaction term *alone* performs worse than a node feature or the combination of node and surprisal (see 6.16A). This is also the case when the interaction is computed with a variable that does not interact with it (see 6.16B). When the interaction term is added to a model that contains both sur-

prisal and node count, the reconstruction of the signal appears to be best when the influence of surprisal is small.

Fifth and finally, a general observation that is nonetheless important: the reconstruction accuracy values decrease in *all* models as the influence of surprisal increases, both when the feature driving the temporal jitter (surprisal) is modelled in some way (6.11A) and when it is not (6.11B). Given that there is no noise in the signal, this means that none of the models are able to fully capture the temporal interaction, and the temporal jitter is modeled as noise (η). The time-invariance of TRF models means that the only way to model temporal interactions is by splitting a feature. This is why split-feature models may perform worse than some other ones when the effect of time is small – having two features with fewer samples is worse for response estimation than one feature with twice as many samples -, but eventually outperform the other models. These models 'suffer less' from the increased temporal variance. This is observable on the less steep slopes of the reconstruction accuracy values as the temporal variance increases. Though not examined here, from this reasoning follows that using *more* than two splits will lead to more division of the temporal jitter over the features, and therefore to a less steep slope.

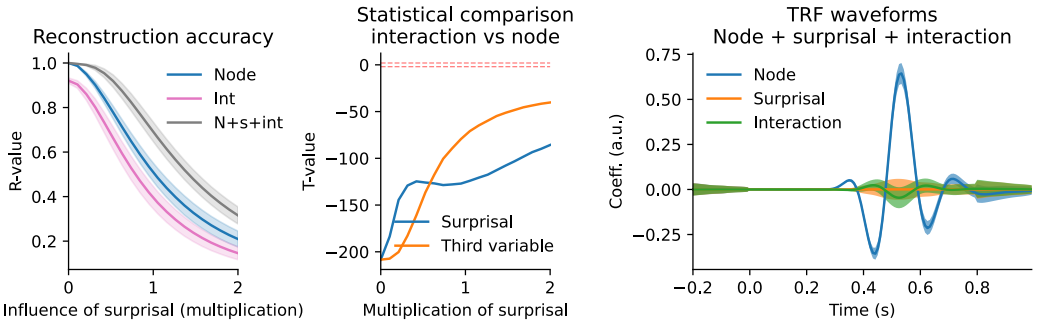


Figure 6.16: A (left). Reconstruction accuracy values for the models Node, Interaction, and Node + surprisal + interaction. B (middle). T-values for the contrast Interaction model / Node model when fit to (1) the data in which surprisal determined the latency shift (in blue) and (2) the data in which a third, unmodelled variable determined the latency shift (in orange). C (right). TRFs extracted using the Node + surprisal + interaction model for multiplication = 2.0 (maximal influence of surprisal).

6.5.3 Conclusions

In this set of simulations, we considered TRF models in the case of temporal interactions. The simulations showed that systematically splitting a feature that predicts a response of which the onset is parametrically related to a second feature *using* that second feature will lead to an increase relative to (1) a model with that feature intact and (2) a model with a random split of that feature. For comparison with other types of models, most notably models with all main features and their interaction, a higher reconstruction accuracy for the split model will may only appear when the influence from the second feature is large enough relative to the duration of the response.

Furthermore, the simulations revealed that a feature which functions as a temporal modulator and does not predict its own response *can* increase reconstruction accuracy when added to a TRF-model, and also have a visible TRF-response (the beta-weights are not zero). This is an effect that one should keep in mind when extracting responses to – especially – linguistic stimuli. Temporal effects as a result of linguistic manipulations are paramount in brain and behavior (Brysbaert, Mandera, & Keuleers, 2018; Donhauser & Baillet, 2020; Kaufeld, Ravenschlag, et al., 2020; Linzen & Jaeger, 2016; Martin & Doumas, 2017, 2019a, 2019b; Tanner et al., 2014; Ten Oever & Martin, 2021). It is possible that some of the responses that one models in their TRF model does not exist in the brain as a separate response – rather, it is an effect of temporal modulation on other responses.[1]

Finally, we have seen that none of the models evaluated in this set of simulations is able to capture the temporal interaction to its full extent – not even the interaction models. This means that in a TRF model, temporal jitter to a large extent is *noise*. These two final points clearly highlight the inability of the temporal response function to capture effects of latency, and potentially mask them – a crucial missing feature for language research.

[1] Interestingly, an effect that is often found to be of temporal nature is (lexical) surprisal (e.g., Futrell et al., 2020; Levy & Gibson, 2013; Monsalve et al., 2012). When using surprisal to model our signal using TRFs, it is entirely possible that the response that we extract is not one of its own, but rather an effect of temporal modulation stemming from a wide array of other linguistic aspects of the stimulus and processes in the human brain – such as structure building, semantic composition, and lexicosemantic association. See also Chapter 1 of this dissertation.

6.6 General discussion

In this Chapter, I presented simulations that explore some properties of the TRF that served to determine whether some of the effects found in other Chapters of this dissertation could be attributable to properties linear model and as such are unrelated to the linguistic phenomenon under consideration. The Chapter answered the three following questions: (1) How does the interstimulus interval (ISI) affect the reconstruction accuracy of the TRF model? (2) If a feature enhances reconstruction accuracy in one frequency band, but not the other, does that mean that the response is in this frequency band? (3) Are different feature values able to extract the same TRF waveform? (4) Is the TRF suitable to model interactions between features *in time*?

In the first set of simulations, we evaluated the effect of the interstimulus interval and the potentially artefactual effects of our band-pass choices in order to answer questions (1) and (2). The simulations showed that ISI does not have an effect on reconstruction accuracy; rather, the signal-to-noise ratio does. This finding can be nicely explained by some properties of the model itself and the model evaluation we have described in the introduction. Recall the equation in 6.7: we evaluate the model by comparing the explained variance with the total variance. Our stimuli are arrays of zeros with nonzero values inserted at the onset of the response. Importantly, when there is no information contained in the stimulus matrix (they are zero), the variance of the predicted signal $\hat{y}$ is *also* zero – essentially, the values in these positions will be equivalent to the mean – while the variance of the real signal y is dominated by noise. Zero divided by a high number will be zero. For our purposes, that means that if we were to compute the reconstruction accuracy on separate parts of the signal, the reconstruction accuracy is zero in all places where there is *not* a nonzero value, which is everywhere outside of the minimum and maximum lag window. In these places, the reconstruction accuracy will be driven by the noise in the actual data. In other words, the interstimulus interval itself does not matter as long as the lengths of the two compared signals are matched.

Secondly, the simulations showed that band-pass filtering itself does not cause the differences in reconstruction accuracy; properties of the responses do. Specifically, in which frequency band the response has most power drives the reconstruction accuracy of the neural signal. When we consider further properties of the temporal response function from a signal processing perspective, this finding makes perfect sense. Convolution in the time domain corresponds to pointwise multiplication in the frequency (Fourier) domain, and vice versa (the *convolu-*

tion theorem). This means that we can rewrite the time-domain equation for our models to a version in spectral space, displayed in 6.8 below.

$$\hat{Y}(f) = X(f) \cdot B(f) \tag{6.8}$$

In this equation, $\hat{Y}$(f), X(f) and B(f) denote the Fourier spectra of $\hat{y}$, x, and β, respectively. Each frequency contributes independently to the outcome of the equation (pointwise multiplication: frequency A in X is multiplied with frequency A in B). This means that linear systems as the one we use here *cannot* generate a response with a frequency component that is not in the input, and *will* extract responses with frequency components that are in the input. In our simulations, therefore, the extracted β coefficients will have the frequency components that are in the data y, and, by proxy, so will $\hat{y}$.

When we then consider a model that is estimated on a narrow-band signal, such as the delta- or theta bands as is the case in this dissertation, we show the relevance of the extracted band-specific response to the band-specific signal. Since the linear time-invariant system acts independently on each frequency we should find a higher reconstruction accuracy in a given frequency band if and only if y and x share power in the frequency band being analyzed. In the case of multiple kernels and multiple features, as is almost always the case in language research, if we find an effect for one feature in one specific frequency band (but not another), that means that the extracted response indeed has most spectral energy in that particular frequency band.

In the second set of simulations, we answered question (3). We concluded that different feature values will extract the same response if they were indeed taken from a single linear feature-response relationship. This is not surprising at all. Let us consider a highly simplified example. Imagine we are trying to solve the equations in 6.9 and 6.10 for A (the beta weight). We already know the error (8), we have our y (-7 in 6.9 and 1523 in 6.10). The feature values (our x) are wildly different: -1 and 101, but because the relation between y and x is identical, we find the same value for A in both cases: 15. Both of these are points on the line y = 15x + 8.

$$-7 = -1A + 8 \tag{6.9}$$

$$1523 = 101A + 8 \tag{6.10}$$

If we create models with the same feature on two separate conditions, and the feature values are different between conditions, that does not impede us from extracting the same response. On the flip side, if there *are* differences between the responses, that is not a result of our differing feature values. To make this concrete for language research, imagine that in Chapter 2 we wanted to study the effect of surprisal between word lists and sentences. Obviously, surprisal is a multiword estimate, so the resulting surprisal values will be different in our two conditions. If the extracted surprisal response differs between conditions, that is crucially *not* due to different surprisal values. Rather, it is an effect that is in the data; the brain might respond to multiword probability differently depending on whether the words are combined into sentences or not. (N.B., this is hypothetical – we did not test this.)

In the final set of simulations, we considered TRF models in the case of temporal interactions. This set answers question (4): is the TRF suitable to model interactions in time? The simulations showed that modelling a response A – node count – of which the onset is parametrically related to a second feature B – surprisal – by systematically splitting the feature of A on the basis of the values of B will capture the time-shift in the TRF waveform, providing insight into the existence potential temporal interaction (which may not be known beforehand). To all of the models evaluated in these simulations, the temporal jitter was *noise (η)*. In the models with the systematically split features, the unexplained variance becomes smaller relative to unsplit models when the error is large enough, because the variance will then be centered around two features. The distance between the actual responses and two separate extracted responses is then smaller than distance between the actual responses and a single extracted response – that is, as long as the temporal jitter is large enough.

Finally, the simulations revealed that a feature which functions as a temporal modulator and is not directly associated with a separable response *can* increase reconstruction accuracy when added to a TRF-model, and also have a detectable TRF-response (the beta-weights are not zero). This is an effect that one should keep in mind when extracting responses to – especially – linguistic stimuli. Temporal effects as a result of (linguistic) manipulations abound (Brysbaert et al., 2018; Donhauser & Baillet, 2020; Kaufeld, Ravenschlag, et al., 2020; Linzen & Jaeger, 2016; Martin & Doumas, 2017, 2019a, 2019b; Tanner et al., 2014; Ten Oever & Martin, 2021). Many event-related responses are likely subject to temporal changes as a function of other latent variables (Martin, 2020). If these latent variables are included as features in a linear time-invariant system, they

can create the illusion of a response where there is none (at least, not a main effect).

6.7 Conclusions

This Chapter has provided insight into the possibilities and the limits of the temporal response function in the following ways. Firstly, comparing reconstruction accuracy values between conditions is most reliable when the signals from the two conditions have the same duration and share the same number of responses. The simulations showed that if this is *not* the case as a result of different interstimulus intervals, the interstimulus interval by itself does not drive differences in the reconstruction accuracy. Rather, the stretch of data that contains noise is of importance. For language research, that means that conditions should be balanced for data duration when the time interval between impulses is not balanced. That is, if we present participants with two conditions that differ in the summed duration of the pauses between the stimuli, we must make sure to record some resting state before or after the shortest sequence of stimuli to ensure the same temporal proportion of the recording contains the response we are modelling in our TRF-models. Secondly, any differences between frequency bands resulting from band-specific TRF-models can be interpreted reliably. For example, if we find that a given feature has an effect on the reconstruction accuracy in the delta band, but not in the theta band (or any other combination of frequency bands) this means that the response associated with that feature has power in the delta band. Thirdly, in the case of two different (linguistic) conditions with unbalanced feature values, the same TRF can be extracted if the true responses in the data are indeed identical in the two conditions. If the extracted TRFs do differ, this means the response to that particular feature differs between the conditions: this difference is not attributable to the different feature values. Fourthly, and finally, the TRF can capture effects in time (i.e., response delays or time-shifts), but only in a categorical fashion. This means that temporal effects can only be captured by having separate conditions (one in which the response is delayed relative to the other one). Such separate conditions can provide general insight into whether a given factor influences the delay or time-shift of a response. The simulations revealed that models with a categorical split are reliably evaluated for reconstruction accuracy by comparing the systematic categorical split to a random categorical split. When creating such models, it is important to keep in mind that the TRF cannot directly model temporal inter-

actions of a continuous nature. That is, the TRF cannot be used to estimate the coefficient of the delay of one response as a function of the other. This is a direct consequence of the time-invariant linear system. Instead, the TRF model will capture this temporal effect as noise, and potentially a separate response (e.g., surprisal).

7 | General discussion

Language is one of the core capacities that make us human. The incredible generative power of human language lets us produce and understand unique sentences: we build structure from sequences of words in order to understand what we hear, read, or see. Yet, how we do it remains an intriguing open question in the fields of linguistics, psychology, cognitive science, and cognitive neuroscience. One strand of research has focused on our ability to build syntactic structure as the result of learning and using sequential statistics, such as transitional probabilities between different units (Frank & Bod, 2011; Frank & Christiansen, 2018; Frost et al., 2019; McCauley & Christiansen, 2019). Another strand has modeled the role of syntactic structure as a separate level of representation that is hierarchically structured and abstracts away from the lexical items itself (Brennan & Hale, 2019; Lo et al., 2022; Martin, 2016, 2020; Matchin & Hickok, 2020). In this dissertation, I approached our combinatorial capacity from the perspective that human brains are both probabilistic engines, and abstract structure-driven computers. Specifically, I investigated how lexical distributional information, such as surprisal and word frequency, and syntactic information jointly shape the process of language comprehension.

The studies presented in this dissertation answer various questions surrounding this topic, and raise even more new ones. In what follows, I will first provide a summary of the main findings from each Chapter. After that, I will address some of the questions that arise when interpreting the findings of **Chapters 3, 4 and 5** in light of the theoretical position of **Chapter 2** (surprisal values reflect and capture variation from a wide array of latent linguistic factors) and some of the findings from the simulations in **Chapter 6**. Then, I will interpret the findings in two computational models that leverage *time* in computation to model language processing, and provide a verbal description of a model that combines insights from both of these. This model contains a natural way in which lexical probabilistic information can inform structure building. I close this dissertation with a summary.

7.1 Summary of main findings

In **Chapter 2**, I asked two main questions. Firstly, I asked why lexical surprisal works well as a predictor for human behavioral and neural data. I argued that surprisal is a great predictor for data because it is "representationally agnostic": it captures variance from all potential sources, including syntactic structure. I showed this through simulation with a toy grammar and recurrent neural networks (RNN), varying both word frequency values and the grammar of the input language. Secondly, given the answer to the first question, I asked what the results from studies that used lexical surprisal as a predictor can tell us about language processing. I concluded that effects of surprisal themselves do not directly inform theories of language comprehension because they lack discriminative insight into the latent variables (such as word frequency and syntactic structure) driving the surprisal values. This is not a problem if the study exclusively aims to predict data; it becomes a problem in the development of a theory of language comprehension.

In **Chapter 3**, I presented results from an analysis project of magnetoencephalography (MEG) data. I investigated whether the presence of syntactic structure affects how delta- and theta band (<4Hz and 4-10Hz, respectively) neural activity represents lexical information. I did this by extracting a purely lexical response from two different conditions: sentences and word lists (scrambled versions of the sentences). The TRF approach allowed me to disentangle signatures of lexical processing from other processes, such as the response to the acoustics of the stimulus. I modeled the responses to lexical information with word frequency and compared these responses between the sentence and word list conditions in sensor space and in source space. The results revealed that responses to words are affected overwhelmingly in the temporal domain, though also spatially, by the presence of syntactic structure. The responses to words were delayed by approximately 350 milliseconds in the word list condition relative to the sentence condition. Furthermore, the response to word frequency was more strongly represented in the signal when the word was embedded in a sentential context. Finally, only in the sentence condition did the lexical information reach the left inferior frontal gyrus. Taken together, this suggests that lexical information is propagated to the left inferior frontal gyrus when the information is to be integrated with the prior context, and as a consequence lexical information is represented more robustly in the delta-band (but not theta-band) neural signal in sentences than in word lists.

In **Chapter 4**, I approached the interplay between lexical distributional information and syntactic information from another perspective: instead of investigating whether lexical information is processed differently given the availability, or lack, of syntactic information, here, I investigated whether the probability of a word in context affects the use of syntactic information. This question is interesting from two perspectives. The first perspective is language comprehension as an instantiation of cue-based inference, in which statistical knowledge and syntactic knowledge both function as cues. According to this perspective, the statistical probability of a word and grammatical knowledge of the receiver should both affect the process of comprehension. The second perspective is the recent view that surprisal from various statistical language models can capture all sorts of psycholinguistic effects. With this in mind, in this study I investigated whether lexical surprisal affects the computation of subject-verb agreement in an online self-paced reading paradigm. The results provided no clear evidence for an interaction between surprisal and the grammaticality of the target on reading times. However, the results did indicate that the best model of the data requires an explicit specification of grammaticality; only lexical surprisal is not enough. This suggests that while lexical surprisal contains some information about grammaticality through the input the model has received (as explained in Chapter 2), it is not enough as a model of language comprehension in humans. The results of this study suggested that language comprehension is strongly guided by grammaticality.

In **Chapter 5**, I asked again whether lexical probabilistic information affects syntactic processing, but this time using the same approach as in **Chapter 3**. In this study, I analyzed MEG data from a naturalistic listening paradigm: participants were listening to an audiobook in the scanner. We created several annotations of the audiobooks, among which a minimalist syntactic parse. Using TRF-models, I extracted responses to those syntactic annotations for words that were associated with high- or low surprisal values. As in **Chapter 3**, I compared the resulting responses to each other. The results showed that the probability of a word given the context affects the time-course of structure building: the response associated with structure building is delayed by as much as 150 milliseconds for words that are unexpected given the context (high surprisal) compared to words that are relatively more expected given the context (low surprisal).

Chapter 6 presents an overview of several sets of simulations that complemented and informed the analyses presented in **Chapters 3** and **5**. The goal of the simulations was to assess whether any effects found in the analyses from

Chapters 3 and **5** could be attributable to properties of either the data or the linear model that were unrelated to the theoretical phenomenon under consideration. These simulations help situate the interpretation of the findings presented in the thesis. In this Chapter, I specifically address the following four questions. (1) How does the interstimulus interval (ISI) affect the reconstruction accuracy of the TRF model? The simulations showed that the interstimulus interval itself does not affect the reconstruction accuracy of the TRF model; rather, interactions between the data length and the proportion of noise do. (2) If a feature enhances reconstruction accuracy in one frequency band, but not the other, does that mean that the response is in this frequency band? The presented simulations show that results restricted to specific frequency bands are reliable; this is a direct consequence of pointwise multiplication. (3) Are different feature values able to extract the same TRF waveform? Indeed, the simulations suggest that this possible. (4) Is the TRF suitable to model interactions between features *in time*? Perhaps the most interesting question given its direct relation to the theoretical implications of Chapter 5, the simulations indicate that the TRF cannot model continuous interactions between features in time. The only way to capture time-shifts is by modelling separate conditions.

Considering these findings together, two key aspects of language system in the process of comprehension emerge. The first aspect suggests that grammatical knowledge strongly influences both behavior and neural dynamics. We can see this in the large differences between lexical processing in word lists and in sentences in **Chapter 3**, the large influence of grammaticality on reading times in **Chapter 4**, and the significant contribution of node-count features to TRF models of the data in **Chapter 5**. At the same time, the simulations from **Chapter 2** suggest that the effects of surprisal capture variability that is driven not by statistical processing, but by the use of grammatical knowledge. Moreover, as shown in **Chapter 6**, besides containing information of other latent variables, a response to a surprisal predictor in a TRF model may reflect temporal variation in higher-level computations – such as structure building – that is a result of the predictability of a word, rather than a separable response to the predictability of the word itself. These findings raise several questions concerning the nature of lexical probability and the status of syntactic knowledge. These will be addressed in section 7.2.

The second aspect indicates that knowledge from the internal language model - of structure and probabilistic information alike - modulates the temporal dynamics of neural correlates of linguistic processes, specifically of lexical process-

ing and syntactic structure building. The lack of structural context induces a delay of ~350 milliseconds in lexical processing, and high surprisal can introduce a delay of ~150 milliseconds in the process of structure building. The simulations presented in **Chapter 6** showed that the frequently used TRF-models are not able to capture effects of temporal shifts. This means that the effects of time shown in **Chapters 3** and **5** of this dissertation could only be shown by modeling the neural signal twice depending on the presentation conditions. Time may be a crucial factor in the process of language comprehension. How time can be of use in an account of language comprehension that leverages both distributional and structural information will be discussed in section 7.3 below.

7.2 Syntax and surprisal – a tension, trade-off, or collaboration?

Chapters 2, **3**, **4** and **5** reveal in different ways that abstract linguistic knowledge is important for models of the neural data and behavior, and that lexical distributional information such as surprisal (and, perhaps, entropy) is not sufficient to model human data; we need predictors that derive from our knowledge of syntax to adequately model the observed values (Bai et al., 2022; Brennan & Hale, 2019; Brennan & Martin, 2020; Coopmans, 2023; Coopmans et al., 2022; Lo et al., 2022; Ten Oever, Carta, et al., 2022; Weissbart & Martin, 2023). This is directly in line with formal analyses of language (e.g., Chomsky, 1965; Everaert et al., 2015; Jackendoff, 1972; Rizzi, 1997), which highlight aspects of linguistic structure that cannot be explained by statistics alone. What, then, *is* the role of lexical distributional information in the process of comprehension?

Some of the Chapters in this dissertation may suggest that the effects of lexical distributional information observed in the literature (e.g., Heilbron et al., 2022; Lowder et al., 2018; Monsalve et al., 2012; Weissbart et al., 2019) are of correlational nature rather than causal. **Chapter 2**, for example, showed that (lexical) surprisal values arise and change as a consequence of various sources of variation, among which word frequency and the syntactic structure underlying the string of words. Furthermore, **Chapter 6** showed that the effect of surprisal in TRF-models can be the result of time-shifts in a different response, rather than a separate neural response as frequently found in the literature (Weissbart et al., 2019). These observations may suggest that effects of sequential statistics are not associated with a separate, potentially domain-general (Conway & Christiansen, 2005; Daltrozzo & Conway, 2014; Frost, Armstrong, Siegelman, &

Christiansen, 2015) cognitive process of probabilistic processing. This is not the case. The literature from statistical learning cited in some parts of this dissertation unambiguously demonstrates that humans (and, indeed, other animals) are sensitive to the statistics of the sequential input in the absence of any latent variable that could causally underlie the sequential output (Armstrong, Frost, & Christiansen, 2017; Aslin et al., 1998; Bai et al., 2022; Batterink & Paller, 2017, 2019; Frost et al., 2019; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Santolin & Saffran, 2018). What the suggestions from **Chapters 2** and **6** do mean, is that we cannot be certain that the effects we observe for statistical predictors in the models of our data are the result of statistical processing alone. They may stem from other latent variables that are causal to the observed statistical patterns – such as our knowledge of the structure of language (Martin, 2016, 2020). This does not apply only to the literature that (fully) relies on these distributional metrics (e.g., Armeni et al., 2019; Frank & Bod, 2011; Heilbron et al., 2022, 2019; Monsalve et al., 2012), but also the literature that aims at disentangling these factors, such as **Chapters 3** and **5** of this thesis. Especially **Chapter 5** is subject to these concerns. When one uses surprisal of any language model as a feature in a model, this feature will capture effects that are driven by other latent variables. In that sense, it is possible, if not likely, that the delay in the neural correlate of structure building for low surprisal relative to high surprisal words is not driven exclusively by processes that are statistical in nature. For example, factors like semantic coherence (the extent to which the meaning of a word is related to the context) are related to distributional information but not identical. While words that are semantically coherent with the context likely have low surprisal values, there are situations in which a word may be probabilistically unexpected, but semantically coherent with the context. Such factors may play a role in this process, too.[1]

If distributional estimates such as surprisal reflect variation from many underlying variables, what does this interpretation mean for any effects of surprisal, entropy, transitional probabilities, and other metrics in models of neural data, such as TRF models? This dissertation does not provide a definite answer to the question, but on the basis of this work it is likely that these effects are a sum of several underlying processes. The effects may be partially caused by temporal modulation of higher-level computations as shown in the simulations in **Chapter 6** and the data of **Chapter 5**, as a result of distributional informa-

[1] In fact, semantic coherence has been found to help learners to learn distributional information (Ouyang, Boroditsky, & Frank, 2017).

tion informing the process of comprehension as a cue. The effects may also be caused by predictability as a result of latent linguistic variables, such as knowledge of grammar or semantic acceptability, which are captured by surprisal but not purely statistical in nature in the sense that transitional probabilities in the statistical learning literature are. And, finally, part of the response may be actual "pure" probabilistic processing – the process that is tapped into with statistical learning experiments. Future research should aim to disentangle the individual contributions of these sources of variance.

In order to distinguish between the representations of probabilistic information and the role of other factors, a few open questions remain. For example, how does the brain represent probabilistic information? Before we can answer this question, it is necessary to have clear theoretical models of *what* the brain computes probabilistic information over. After all, for the brain to be sensitive to the probability of an item, that item must be somehow represented. For language, it appears likely that most levels of representation receive some probabilistic treatment (phonemes: Gwilliams, Linzen, et al., 2018; Gwilliams, Poeppel, et al., 2018; Ten Oever et al., 2024; Tezcan et al., 2023; words: Armeni et al., 2019; Heilbron et al., 2019; Slaats et al., 2023; Weissbart et al., 2019; phrases and sentences: Arnon & Snider, 2010; Linzen & Jaeger, 2016). But how does one separate the representation from its probability? A potential thesis is that this cannot be done, and that the neural representations themselves are, to some degree, probabilistic (Martin, 2016, 2020; Norris & McQueen, 2008). However, the mental representation of language must be discrete at some point during processing, otherwise there cannot be a stable interpretation of the input (see also **Chapter 2** of this dissertation, and Fodor & Pylyshyn, 1988). With this in mind, a way to separate a representation from its probability while maintaining the large role for probabilistic information is to model the relationship between the cue and the target representation as a probabilistic one, as is done in the models by Martin (Martin, 2016, 2020).

Given this interpretation of effects of probability, it is difficult to determine what effects lexical distributional information can have on the process of language comprehension broadly, and structure building in particular. There are several options that may exist simultaneously, of which I will mention two. The first one, seen in **Chapters 3**, **4**, and **5** of this dissertation, is that probabilistic information can affect the timing of linguistic computations. This will be discussed in more detail in section 7.3. A second option is that the probability of a representation – be this a phoneme, a word, or a syntactic category – can

qualitatively affect the process of inference by disambiguating the input (Hale, 2001, 2006; Jurafsky, 1996) by sensitizing the system for a particular grammatical construction on the basis of probability (e.g., Linzen & Jaeger, 2016). Both of these effects can result from the use of probability as a *cue* for the inference of the structure and meaning of an utterance (Martin, 2016, 2020).

7.3 The role of time in language processing

The results from **Chapters 3** and **5** suggest that the timing of neural responses is strongly dependent on the linguistic information available to the listener. This finding is the mirror image of studies that show that timing of the input (relative to the phase of the neural signal) influences what the listener perceives (Kaufeld, Ravenschlag, et al., 2020; Kösem et al., 2018; Ten Oever & Sack, 2015). Below, I will discuss the findings from these Chapters from the perspective of two theoretical frameworks that leverage time in language processing and the computation of structure: *STiMCON* (Ten Oever & Martin, 2021, 2024), which aims to explain how isochronous oscillations can track pseudo-rhythmic speech and can explain some of the effects of time found in this dissertation; and *time-based binding* (Martin & Doumas, 2017, 2019a, 2019b), which aims to provide a mechanistic account of compositionality in a neural system. When combined with STiMCON, the latter model can explain how both probabilistic and syntactic processing can shape the process of language comprehension and its read-outs.

7.3.1 STiMCON

The brain displays intrinsic oscillatory activity. This activity arises as a result of and/or exists as a modulator of the excitability of neural populations (Buzsáki, 2004). At the same time, the dynamics of ongoing oscillations are moderated by cognitive processes and stimulus processing (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008). Neural oscillations have been hypothesized to track the temporal dynamics of speech for optimal processing (Keitel et al., 2018; Keitel, Ince, Gross, & Kayser, 2017; Lakatos, Gross, & Thut, 2019; Zion Golumbic et al., 2013). This has been suggested to lie at the root of the phenomenon that the timing of the stimulus can affect what the listener perceives mentioned above (Kaufeld, Ravenschlag, et al., 2020; Kösem et al., 2018; Ten Oever & Sack, 2015). However, while oscillators produce an isochronous rhythm, speech is not purely

rhythmic, and it is not clear how isochronous oscillations can track input that is at best pseudo-rhythmic.

Ten Oever and Martin (2021; 2024) show that the temporal dynamics in speech are dependent on the predictability of words in a sentence and suggest that these systematic temporal dynamics – which lead to the pseudo-rhythmicity of speech – in fact carry information to the processor and need not be a problem for tracking of speech in an isochronous oscillator. They show how this could work in their computational model STiMCON (Speech Tracking in a Model Constrained Oscillatory Network; Ten Oever & Martin, 2021). In brief, the model consists of lexical nodes that have a certain threshold of activation, an isochronous oscillator at a 4Hz cycle that sensitizes and desensitizes the lexical nodes, lateral inhibition, and an internal language model (the individually acquired statistical and structural knowledge of language stored in the brain; in the model represented by simple transitional probabilities) that provides feedback. The ongoing oscillator determines a sensitive window for the lexical nodes. This means that the internal language model will sensitize a node that is predictable from the context, and will lead to an earlier supra-threshold activation of this particular node – on a less-excitable phase of the oscillation.

The simulations with STiMCON showed the following. Firstly, the presentation of words that had different predictability values indeed affected the activation time of the node corresponding to that particular word. In other words, the precise timing at which nodes reach their threshold of activation is dependent on the feedback that is coming from the internal language model: the nodes for words that are more predictable receive more feedback, and are therefore active earlier, than the nodes for less predictable words Secondly, varying the presentation time for words with different predictability values revealed that words that have higher levels of feedback (i.e., are more predictable) can be processed at earlier times relative to the isochronous theta cycle than words that are less predictable. This is a result of feedback and lateral inhibition. The simulations revealed that the difference in presentation timing do not directly map to activation timing differences: 130 milliseconds of presentation variation leads to only 19 milliseconds variation in the activation times of the word nodes. This means that the model provides a mechanism that can map isochronous neural oscillations onto the pseudo-rhythmic speech signal in a way that leverages the temporal variation in the input. The simulations indicate that non-isochrony in the input can lead to isochrony in the brain (i.e., activation of the word nodes

is more isochronous than the input) when the temporal variation in the input matches the predictions from the internal language model.

The results presented in this dissertation are in accordance with STiMCON from various perspectives. For example, in **Chapter 3** I showed that the presence of syntactic information in the input signal leads to an earlier response to words. Rephrasing this to fit the terminology of STiMCON, we could state that the participants' internal language model (which contains knowledge of the structure of language as well as probabilistic knowledge) provides feedback through syntactic and compositional semantic knowledge, sensitizing nodes that represent lexical input *earlier* than when this feedback is not available. This could be interpreted as an instantiation of the feedback mechanism proposed by Ten Oever and Martin (2021) at a higher level of abstraction. In a similar vein, **Chapter 5** showed that the bottom-up node count response appeared earlier when the word was predictable from the context (i.e., low surprisal). This can be interpreted as nodes that represent grammatical encoding of the input being sensitized and thus active earlier when a word is predictable relative to when it is not predictable from the context.

These findings are first and foremost in accordance with the authors' general proposal that 'what' (i.e., the linguistic content) and 'when' (of the input or the neural response) are not independent, at least not in the delta band. The presence or absence of syntactic structure modulates the timing of the delta-band response to words; and, conversely, the predictability of a word modulates the timing of the delta-band neural correlates of structure building. More specifically, the findings fit with the overall dynamics of activation exhibited by the STiMCON computational model. Full integration of the present results in the computational model would require an expansion of the internal language model, which is kept limited to transitional probability information for purposes of the simulation, namely with structural knowledge of language. The results presented in this dissertation suggest that the interaction between time and linguistic knowledge holds at the interface between lexical processing and structure building in addition to the purely lexical processes modelled in STiMCON.

A few aspects of the results from this dissertation complicate the comparison between those results and the proposal by Ten Oever and Martin (2021, 2024). Firstly, although the patterns observed on the TRFs match the timing of node activation in STiMCON qualitatively, the TRF-analysis used in this dissertation does not provide direct insight into oscillatory activity. As a consequence, it is unclear whether the findings from **Chapters 3** and **5** reflect phase-resets of on-

going oscillations or whether we are looking at impulse/evoked responses (or both). Secondly, the results presented here show temporal variation that is relatively large, with 150 milliseconds variation in the neural correlate of structure building, and as much as 300 milliseconds variation in the neural response to word frequency. That is quite a large effect if we suppose that an isochronous oscillator acts as a temporal filter; especially taking into consideration the findings from Ten Oever and Martin (2021), which suggest that the variation in word onset differences and word durations is in the order of magnitude of less than 100 milliseconds, that gets diminished to only 19 milliseconds variation in the activation of word nodes (the model for neural activation). If the mechanism proposed in STiMCON plays a role in the generation of these results, an open question remains how these temporal variations in the neural data can become so large.

Despite these aspects that complicate the comparison, the results from **Chapters 3** and **5** are in line with the general perspective put forward by Ten Oever and Martin: namely, that the brain state at the time of stimulus onset influences processing of the current stimulus – i.e., that the current stimulus is integrated with the ongoing process – and that this phenomenon is reflected in the timing of the neural response.

7.3.2 Time-based binding

Departing from the need for a mechanistic account of language processing, (Martin & Doumas, 2017, 2019a, 2019b) propose a computational model of *binding* (i.e., combining elements for further processing) that displays oscillatory activity as a natural consequence of its organization. A mechanistic account of language processing must satisfy several computational requirements. One of them is central to this dissertation: the mechanism must compute discrete, structured representations from unstructured continuous input (speech) in time. At the same time, the mechanism must maintain representations of the parts of the newly formed structure, such as the words, alongside computing and maintaining the structure itself. Several models abide to one of these criteria, but not both. Recurrent neural networks, for example, create conjunctive representations through what is sometimes called *synaptic binding*; the network learns by updating the weights between the nodes, with these connections representing synapses. This type of model is able to learn from unstructured continuous input based on statistical association, but the representations of the parts are not maintained. Instead, the model forms conjunctive represen-

tations from which the individual parts are indistinguishable: the representation of the word {cake} will not have any relation to the representation of the sentence {cool professors eat cake} because {cake} is not represented independently (Martin & Doumas, 2017), while it is in a compositional representation ($\{\{\{\{cool_{adj}\}\}_{AdjP}\{professors_{n}\}\}_{NP}\{\{eat_{v}\}\{\{cake_{n}\}_{NP}\}_{VP}\}\}_{IP}$).

Martin and Doumas (Martin & Doumas, 2017, 2019a, 2019b) propose a solution to this problem: instead of synaptic binding, representations are bound by *time*. The authors call this *time-based binding*. The basic idea is that the system uses the timing of the firing of nodes to carry information about the relationship between representations for further processing rather than combining the representations to form a conjoined representation. Many proposals that use time to bind representations rely on synchrony of activation (e.g., Senoussi, Verbeke, & Verguts, 2022): when two nodes are active at the same time, they are linked together for further processing. Two nodes that are out of synchrony can stay independent. Martin and Doumas' DORA (Discovery of Relations by Analogy), designed to represent predicate relations (Doumas et al., 2008; Martin & Doumas, 2017, 2019a, 2019b), employs a slight asynchrony between representations. In this way, the model can leverage closeness in time to bind representations, yet simultaneously maintain independence between levels of representation. DORA consists of a neural network with several layers of units with lateral inhibition. These layers are necessary to create hierarchical representations. The units are grouped in four separate banks; these are groups of units that serve a specific function, such as the current focus of attention and long-term memory. The system learns through Hebbian learning and is – crucially – sensitive to time.

Processing of an input sequence works as follows. The model encodes the elements that are bound in lower levels of a hierarchy directly from the sequential input and then uses slower dynamics to accumulate evidence for relations at higher levels of the hierarchy. Units on a higher layer of the network fire when two or more subunits, such as word nodes, fire within a certain time of each other. This results in a hierarchical representation that represents input values independently. Specifically, when the model is presented with $\{\{\{cool_{adj}\}\}_{AdjP}\{professors_{n}\}\}_{NP}$, this will be represented across two layers. In the bottom layer, the nodes representing the words {cool} and {professor} will fire at a slight asynchrony, and activate the phrase node in the second layer – also at an asychrony to the layer below. This asynchrony allows for a separate representation of the words and the phrase.

In this model, Martin and Doumas (2017) simulated processing of stimuli from Ding and colleagues (Ding et al., 2016). This seminal work had shown that the brain tracks not only words, but also phrases and sentences as indicated by a 2Hz peak (the frequency of the phrases) and a 1Hz peak (the frequency of the sentences) that were absent from the data when participants were listening to word lists (Ding et al., 2016). These findings suggested that some cortical populations (groups of neurons, red.) specifically code for higher-level representations. The simulations with DORA performed by Martin and Doumas (2017) revealed the same pattern as found in the data by Ding and colleagues (2016): the activity from DORA showed the same 1Hz and 2Hz peaks when the data contained sentence structure, but not when the model was presented with word lists. This oscillatory pattern was created by time-based binding. This pattern was not obtained from an RNN, which does not use time-based binding (but see also: Frank & Yang, 2018).

In sum, *time-based binding* to encode relations between units satisfies the criteria for a theory of language processing (discrete, structured representations from unstructured input; compositionality), and the mechanism is capable of giving rise to oscillatory dynamics found in response to structured linguistic input. How does this account relate to the findings from this dissertation? As was mentioned in section 7.3.1, the results from the MEG analyses presented in **Chapters 3** and **5** do not directly speak to oscillatory activity, nor do they provide insight into the phase of the response to individual words, phrases, or other units. As such, the findings do not directly fit into the framework presented above. One parallel appears to be that findings suggest that some asynchrony plays a role in processing, both at the lexical level and at the phrasal level. I will go into this in more detail in section 7.3.3 below.

7.3.3 BiMCON: How statistics can inform structure building

The presence of asynchrony is where we can make an interesting connection between time-based binding, STiMCON, and the results from this dissertation: what drives the asynchrony *within* a level of representation in time-based binding, i.e. the asynchrony that determines whether the nodes at the next level of representation fire? After all, when two lower-level nodes fire with too much time between them, the next level will not be activated. In other words, if two lexical representations are active with a large gap in between, the phrase node will not fire. Could the temporal gap between the firing of two words, and therefore the activation of the phrase node, be determined by a word's probability?

To build on these thoughts, one could conceive a model that combines time-based binding with STiMCON, which I will call *BiMCON* (Binding in a Model Constrained Oscillatory Network). This model combines the core features of the two models: the model-constrained oscillatory network from STiMCON, and time-based binding from DORA. Such a model is interesting because a combination of STiMCON and time-based binding could provide a mechanistic account of how statistical information can guide structure building. Namely, if word nodes are bound for further processing by time through time-based binding, and a statistical language model determines the time of activation of a word node, the probability of a word logically plays a role in determining whether this word and the preceding word are bound. In addition, BiMCON would obey the computational requirements for a mechanistic account of language processing through time-based binding as described above, and it would explain how pseudo-rhythmic input can be tracked by ongoing oscillations. There are many ways in which these two mechanisms could be combined. In what follows, I will outline one high-level theoretical possibility; further research is required to determine whether this is the correct way to merge the two proposals and how this could work in practice.

In BiMCON, the mechanism from STiMCON enables predictability to modulate activation time of the word nodes by combining an ongoing oscillator with feedback from a statistical language model. What kind of information this model should contain needs to be determined; it could be modelled as a strongly interconnected lexicon (TRACE (McClelland & Elman, 1986), Shortlist A (Norris, 1994), Shortlist B (Norris & McQueen, 2008)), though models that take sequential probability into account would be preferred, such as a simple n-gram model or an RNN (as in STiMCON originally). In this model, the probability of a word given the context plays a role in the temporal spacing of two adjacent words: the higher the probability of a word, the stronger the feedback, and the earlier it will be activated relative to the ongoing oscillator. The words are bound for further processing by the mechanism time-based binding. The interaction between these two mechanisms (STiMCON and time-based binding, respectively) will lead to how statistical information can inform structure building: two lexical nodes that fire close in time (because the second word is predictable) trigger firing of the phrase node. By contrast, lexical nodes that fire at a larger temporal distance (when the second word is unpredictable given the first word) will not trigger firing of the phrase node, as such effectively introducing a boundary between the two words.

The results of this dissertation consistently show that the presence of syntactic structure and/or grammatical information is an important determinant of the neural signal (**Chapters 3 and 5**) and reading times (**Chapter 4**). The combination of STiMCON and time-based binding in BiMCON provides a direct mechanism to model this strong influence. The simulation of time-based binding by asynchrony in DORA revealed that the system oscillates similarly to human brains during the processing of grammatically correct sentences. The oscillations produced by the network as a result of binding may be routed *back into the* system, such that the oscillator used in STiMCON is no longer an external oscillator, but the rhythm produced by the system itself.

An important issue that requires further research is how the rhythms produced by time-based binding should influence the activation of lexical features. If the oscillator were replaced by a read-out of the activity of the system in real-time (which may seem unrealistic, but is in line with findings showing that neural entrainment determines what we perceive Kösem et al., 2018), then the high excitability phase of word nodes might coincide with the moments *after* binding. This seems unwanted, because it could mean that words are recognized faster on phrase boundaries. The exact effects of such an implementation are difficult to predict without computationally implementing the model, however, because the result will depend on the relative influence of the feedback from the statistical language model and the oscillator on the excitability of the word nodes. Another possibility is to sensitize the isochronous oscillator to a high-level read-out of the activity of the model, for example by introducing a phase-reset of the isochronous oscillator as a function of the activity of the model (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Lakatos et al., 2019; Luo & Poeppel, 2007; Sauseng et al., 2007). The chosen mechanism will inevitably affect the timing of activation of word nodes, and as such binding later in the cycle. Simulation of these options is necessary to identify which system will be able to identify the correct relations between stimuli, as well as display activity that is reminiscent of the patterns observed in the literature and in this dissertation.

It is conceivable that a computational implementation of BiMCON can simulate the findings presented in **Chapters 3** and **5** by looking at the activation of the word- and phrase nodes. The later activation of word nodes when structure is fully absent relative to when it is present shown in **Chapter 3** will be a direct consequence of the "grammatical" oscillator that sensitizes the word nodes (a read-out of the binding system); in the absence of structure, this oscillator will not contain the peaks related to phrase-building, and as a consequence, sen-

sitize the word nodes to a lesser extent. The later activation of phrase nodes when words have high surprisal relative to when they have low surprisal shown in **Chapter 5** is a direct consequence of the lexical probability influencing the process of binding.[2]

In sum, BiMCON combines the core features of STiMCON (Ten Oever & Martin, 2021, 2024) and time-based binding (Martin & Doumas, 2017, 2019a, 2019b) to create a model of structure building that naturally allows statistical information to inform the process of structure building by leveraging time at multiple levels: the time of the input relative to the ongoing oscillator; the feedback from the statistical language model influences the activation time of a word node; and the time of activation of word nodes relative to the previous word affects the process of binding.

7.4 Conclusion

In this dissertation, I investigated how lexical distributional information, such as surprisal and word frequency, and syntactic information jointly shape the process of language comprehension. The studies have highlighted two key aspects of the language system during comprehension. Firstly, grammatical knowledge strongly determines both behavior and neural dynamics. Secondly, the influence of lexical distributional information and syntactic information alike is visible on the neural read-out as an effect of time. When the current linguistic representation does not fit well with the current state of the processor, which is defined by distributional and syntactic information of the internal language model, the neural response is delayed. These findings suggest that time is crucial in the combination of these two types of information. I proposed the model BiMCON (Binding in a Model Constrained Oscillatory Network), a combination of previous models STiMCON and time-based binding. This model leverages time to describe how lexical distributional information can affect the process of structural inference, and how both lexical distributional information and syntactic structure building can shape the neural readout.

[2] A problem is that this likely predicts that lexical processing is affected by the probability of the word, which was not found in **Chapter 5**.

References

Acuña-Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in spanish. *Lingua*, *143*, 108–128. doi: 10.1016/j.lingua.2014.01.013

Amenta, S., Hasenäcker, J., Crepaldi, D., & Marelli, M. (2023). Prediction at the intersection of sentence context and word form: Evidence from eye-movements and self-paced reading. *Psychonomic Bulletin & Review*, *30*(3), 1081–1092. doi: 10.3758/s13423-022-02223-9

Armeni, K., Willems, R. M., van den Bosch, A., & Schoffelen, J. M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, *198*(May), 283–295. doi: 10.1016/j.neuroimage.2019.04.083

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: past, present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *372*(1711). Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/27872366` doi: 10.1098/rstb.2016.0047

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. doi: 10.1016/j.jml.2009.09.005

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3), 170–176. doi: 10.1177/0963721412436806

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, *64*(SUPPL.2), 86–105. doi: 10.1111/lang.12074

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. doi: 10.1111/1467-9280.00063

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

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*, *81*(2), 571–589. doi: 10.3758/s13414-018-1626-4

Bai, F. (2022). *Neural representation of speech segmentation and syntactic structure discrimination* (Doctoral dissertation). Retrieved from `https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3429429_3`

Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. *PLOS Biology*, *20*(7), e3001713. doi: 10.1371/journal.pbio.3001713

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. doi: 10.1016/j.cognition.2012.06.003

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*, *31*(5-6), 378–412. doi: 10.1080/02643294.2014.895314

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. doi: 10.1016/j.cortex.2017.02.004

Batterink, L. J., & Paller, K. A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex*, *115*, 56–71. doi: 10.1016/j.cortex.2019.01.013

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. doi: 10.1016/J.TICS.2011.10.001

Blanco-Elorrieta, E., Ding, N., Pylkkänen, L., & Poeppel, D. (2020). Understanding requires tracking: Noise and knowledge interact in bilingual comprehension. *Journal of Cognitive Neuroscience*, *32*(10), 1975–1983. doi: 10.1162/jocn_a_01610

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. doi: 10.1016/j.neuroimage.2015.11.069

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in english number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. doi: 10.1080/01690969308406949

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, *43*(2), 83–128. doi: 10.1006/cogp.2001.0753

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. doi: 10.1016/0010-0285(91)90003-7

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer.* Retrieved from `http://www.praat.org/`

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America*, *84*(1), 101–114. doi: 10.1121/1.396976

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

Brehm, L., Hussey, E., & Christianson, K. (2020). The role of word frequency and morpho-orthography in agreement processing. *Language, Cognition and Neuroscience*, *35*(1), 58–77. doi: 10.1080/23273798.2019.1631456

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE*, *14*(1). doi: 10.1371/journal.pone.0207741

Brennan, J. R., & Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B*, *375*(1791), 77–83. doi: 10.1098/RSTB.2019.0305

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W. M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157-158*, 81–94. doi: 10.1016/j.bandl.2016.04.008

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., ... Simon, J. Z. (2023). Eelbrain, a python toolkit for time-continuous analysis with temporal response functions. *Elife*, *12*, e85012.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.e5. doi: 10.1016/j.cub.2018.10.042

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. doi: 10.1016/j.neuroimage .2018.01.042

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25–31. doi: 10.1016/j.cophys.2020.07.014

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. doi: 10.1016/j.cub.2018.01.080

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. doi: 10.1016/j.jml.2020.104174

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, *12*, 110. doi: 10.3389/FPSYG.2021.615538/BIBTEX

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. doi: 10.1177/0963721417727521

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... others (2013). Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature Neuroscience*, *7*(5), 446–451. doi: 10.1038/nn1233

Campanelli, L., Dyke, J. V., & Marton, K. (2018). The modulatory effect of expectations on memory retrieval during sentence comprehension. *Publications and Research*. Retrieved from `https://academicworks.cuny.edu/gc_pubs/440`

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... Raffel, C. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*, 2633–2650.

Carnie, A. (2013). *Syntax: a generative introduction* (3, Ed.). Oxford: Wiley-Blackwell.

Chen, J., & ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes*, *117*, 29–

34. doi: 10.1016/j.beproc.2014.09.004

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124. doi: 10.1109/TIT.1956.1056813

Chomsky, N. (1965). *Aspects of the theory of syntax* (Vol. 11). MIT Press.

Christiansen, M. H., & Chater, N. (2015). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*(2016). doi: 10.1017/S0140525X1500031X

Cinque, G. (2004). Issues in adverbial syntax. *Lingua*, *114*(6), 683–710.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(1), 24–39. doi: 10.1037/0278-7393.31.1.24

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, *103*(2), 163–179. doi: 10.1016/j.cognition.2006.03.003

Coopmans, C. W. (2023). *Triangles in the brain: The role of hierarchical structure in language use* (Doctoral dissertation). Retrieved from `https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3498010_6`

Coopmans, C. W., de Hoop, H., Hagoort, P., & Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiology of Language*, *3*(3), 386–412. doi: 10.1162/nol_a_00070

Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2021). Structure-(in)dependent interpretation of phrases in humans and lstms. *Proceedings of the Society for Computation in Linguistics (SCiL)*, 459–463.

Coopmans, C. W., Kaushik, K., & Martin, A. E. (2023). Hierarchical structure in language and action: A formal comparison. *Psychological Review*, *130*(4), 935–952. doi: 10.1037/rev0000429

Creemers, A., & Meyer, A. S. (2022). The processing of ambiguous pronominal reference is sensitive to depth of processing. *Glossa Psycholinguistics*, *1*(1). Retrieved from `https://escholarship.org/uc/item/39k99073` doi: 10.5070/G601166

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human*

Neuroscience, *10*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnhum.2016.00604

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(16), 5842–5847. doi: 10.1073/pnas.1320525111

Daltrozzo, J., & Conway, C. M. (2014). Neurocognitive mechanisms of statistical-sequential learning: what do event-related potentials tell us? *Frontiers in Human Neuroscience*, *8*, 437. doi: 10.3389/fnhum.2014.00437

Deacon, T. W. (1997). What makes the human brain different? *Annual Review of Anthropology*, *26*(1), 337–357. doi: 10.1146/annurev.anthro.26.1.337

de Vries, W., & Nissim, M. (2021). As good as new. how to successfully recycle english gpt-2 to make models for other languages. In (pp. 836–846). Retrieved from http://arxiv.org/abs/2012.05628 doi: 10.18653/v1/2021.findings-acl.74

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. doi: 10.1016/j.cub.2015.08.030

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. doi: 10.1016/j.jml.2013.04.003

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. doi: 10.1038/nn.4186

Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., & Zhang, J. (2018). Attention is required for knowledge-based sequential grouping: Insights from the integration of syllables into words. *Journal of Neuroscience*, *38*(5), 1178–1188. doi: 10.1523/JNEUROSCI.2606-17.2017

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. doi: 10.1016/j.neuroimage.2013.06.035

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393.e9. doi: 10.1016/J.NEURON.2019.10.019

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of

the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. doi: 10.1037/0033-295X.115.1.1

Drennan, D. P., & Lalor, E. C. (2019). Cortical tracking of complex sound envelopes: Modeling the changes in response with intensity. *eNeuro*, *6*(3). Retrieved from `https://www.eneuro.org/content/6/3/ENEURO.0082-19.2019` doi: 10.1523/ENEURO.0082-19.2019

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, *36*(2), 147–164. doi: 10.1006/jmla.1996.2484

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531–559. doi: 10.1037/0033-295X.112.3.531

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2), 195–225. doi: 10.1007/BF00114844

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. doi: 10.1016/j.tics.2004.02.002

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *39*(29), 5750–5759. doi: 10.1523/JNEUROSCI.1828-18.2019

Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. doi: 10.1016/j.tics.2015.09.008

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS ONE*, *8*(10), e77661. doi: 10.1371/journal.pone.0077661

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71. doi: 10.1016/0010-0277(88)90031-5

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*(1), 173–216.

doi: 10.1016/j.cognition.2005.10.003

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*, 475–494. doi: 10.1111/tops.12025

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. doi: 10.1177/0956797611409589

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531. doi: 10.1098/rspb.2012.1741

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*. Retrieved from `http://www.tandfonline.com/action/journalInformation?journalCode=plcp21` doi: 10.1080/23273798.2018.1424347

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PLOS ONE*, *13*(5), e0197304. doi: 10.1371/journal.pone.0197304

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392. doi: 10.1152/physrev.00006.2011

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268. doi: 10.1016/j.tics.2012.04.001

Friederici, A. D. (2015). Chapter 10 - white-matter pathways for speech and language processing. In M. J. Aminoff, F. Boller, & D. F. Swaab (Eds.), *Handbook of clinical neurology* (Vol. 129, pp. 177–186). Elsevier. Retrieved from `https://www.sciencedirect.com/science/article/pii/B978044462630100010X`

Friston, K. (2012). The history of the future of the bayesian brain. *NeuroImage*, *62*(2), 1230–1233. doi: 10.1016/j.neuroimage.2011.10.004

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 1–87. doi: 10.1037/bul0000210

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends Cogn Sci*, *19*(3). doi: 10.1016/j.tics.2014.12.010

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814. doi: 10.1111/cogs.12814

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691. doi: 10.1111/j.1467-9280.2006.01767.x

Gervain, J. (2014). Early rule-learning ability and language acquisition. In F. Lowenthal & L. Lefebvre (Eds.), *Language and recursion* (pp. 89–99). New York, NY: Springer. Retrieved from `https://doi.org/10.1007/978-1-4614-9414-0_7`

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*(MAR), 1–5. doi: 10.3389/fpsyg.2013.00138

Ghitza, O., Giraud, A. L., & Poeppel, D. (2012). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*(DEC), 4–7. doi: 10.3389/fnhum.2012.00340

Giglio, L., Ostarek, M., Sharoh, D., & Hagoort, P. (2024). Diverging neural dynamics for syntactic structure building in naturalistic speaking and listening. *Proceedings of the National Academy of Sciences*, *121*(11), e2310766121.

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *41*(50), 10316–10329. doi: 10.1523/JNEUROSCI.0812-21.2021

Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint arXiv:2103.04469*. Retrieved from `http://arxiv.org/abs/2103.04469` doi: 10.48550/arXiv.2103.04469

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, *7*(267), 1–13. doi: 10.3389/fnins.2013.00267

Greco, M., Cometa, A., Artoni, F., Frank, R., & Moro, A. (2023). False perspectives on human language: Why statistics needs linguistics. *Frontiers in Language Sciences*, *2*. Retrieved from `https://www.frontiersin.org/articles/10.3389/flang.2023.1178932`

Grodzinsky, Y., Pieperhoff, P., & Thompson, C. (2021). Stable brain loci for the processing of complex syntax: A review of the current neuroimaging evidence. *Cortex*, *142*, 252–271. doi: 10.1016/j.cortex.2021.06.003

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. doi: 10.3758/BF03204386

Grosjean, F., & Itzler, J. (1984). Can semantic constraint reduce the role of word frequency during spoken-word recognition? *Bulletin of the Psychonomic Society*, *22*(3), 180–182.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. doi: doi.org/10.1177/1745691620970585

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational brain & Behavior*. doi: 10.31234/osf.io/tbmcg

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599. doi: 10.1523/JNEUROSCI.0065-18.2018

Gwilliams, L., Poeppel, D., Marantz, A., & Linzen, T. (2018). Phonological (un)certainty weights lexical activation. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 29–34). doi: 10.18653/v1/w18-0104

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. doi: 10.1111/1467-9280.00476

Hagoort, P. (2013). Muc (memory, unification, control) and beyond. *Frontiers in Psychology*, *4*(JUL). doi: 10.3389/FPSYG.2013.00416

Hagoort, P. (2015). Muc (memory, unification, control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of language* (pp. 339–347). Elsevier Inc.

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), 439–483. doi: 10.1080/01690969308407585

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model.. Retrieved from `https://aclanthology.org/N01-1021`

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672. doi: 10.1207/s15516709cog0000_64

Hale, J. T. (2016). Information-theoretical complexity metrics. *Language and*

Linguistics Compass, *10*(9), 397–412. doi: 10.1111/lnc3.12196

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, *8*(1), 427–446. doi: 10.1146/annurev-linguistics-051421-020803

Harnad, S. (2003). Categorical perception. Nature Publishing Group: Macmillan. Retrieved from https://eprints.soton.ac.uk/257719/

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. doi: 10.1038/s41586-020-2649-2

Hasson, U. (2017). The neurobiology of uncertainty: Implications for statistical learning. *Phil. Trans. R. Soc. B*, *372*(1711).

Hasson, U., & Tremblay, P. (2015). *Neurobiology of statistical information processing in the auditory domain*. Elsevier Inc. Retrieved from http://dx.doi.org/10.1016/B978-0-12-407794-2.00043-2

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. doi: 10.1073/pnas.2201968119

Heilbron, M., Ehinger, B., Hagoort, P., & de Lange, F. P. (2019). Tracking naturalistic linguistic predictions with deep neural language models.. Retrieved from http://arxiv.org/abs/1909.04400 doi: 10.32470/CCN.2019.1096-0

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, *26*(3), 376–405. doi: 10.1080/01690965.2010.492642

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. doi: 10.1080/23273798.2015.1072223

Huizeling, E., Arana, S., Hagoort, P., & Schoffelen, J. M. (2022). Lexical frequency and sentence context influence the brain's response to single words. *Neurobiology of Language*, *3*(1), 149–179. doi: 10.1162/NOL_A_00054

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2022). Statistically based chunking of nonadjacent dependencies. *Journal of Experimental Psychology: General*. Retrieved from https://psycnet.apa.org/

psycarticles/2022-55323-001 doi: 10.1037/XGE0001207

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*. The MIT Press. Retrieved from https://eric.ed.gov/?id=ED082548

Jafarian, M., & De Persis, C. (2015). Formation control using binary information. *Automatica*, *53*, 125–135. doi: 10.1016/j.automatica.2014.12.016

Johnson, R. L., & Rayner, K. (2007). Top-down and bottom-up effects in pure alexia: Evidence from eye movements. *Neuropsychologia*, *45*(10), 2246–2257. doi: 10.1016/j.neuropsychologia.2007.02.026

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194. doi: 10.1207/s15516709cog2002_1

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. doi: 10.1037/0033-295X.87.4.329

Kapteijns, B., & Hintz, F. (2021). Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics. *PLOS ONE*, *16*(7), e0254546. doi: 10.1371/journal.pone.0254546

Katz, L., Boyce, S., Goldstein, L., & Lukatela, G. (1987). Grammatical information effects in auditory word recognition. *Cognition*, *25*(3), 235–263. doi: 10.1016/S0010-0277(87)80005-7

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience*, *40*(49), 9467–9475. doi: 10.1523/JNEUROSCI.0302-20.2020

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (2020). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(3), 549–562. doi: 10.1037/xlm0000744

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology*, *16*(3), e2004473. doi: 10.1371/JOURNAL.PBIO.2004473

Keitel, A., Ince, R. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, *147*, 32–42. doi: 10.1016/j.neuroimage.2016.11.062

Keuleers, E., Brysbaert, M., & New, B. (2010). Subtlex-nl: A new measure for dutch word frequency based on film subtitles. *Behavior Research Methods*

2010 42:3, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. doi: 10.1016/j.csl.2017.01.005

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 169–181. doi: 10.1037/0278-7393.22.1.169

Krauska, A., & Lau, E. (2023). Moving away from lexicalism in psycho- and neuro-linguistics. *Frontiers in Language Sciences*, *2*. Retrieved from `https://www.frontiersin.org/articles/10.3389/flang.2023.1125127`

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *3798*. doi: 10.1080/23273798.2015.1102299

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611. doi: 10.1016/j.jml.2012.04.003

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. *arXiv preprint arXiv:2205.11463*. Retrieved from `http://arxiv.org/abs/2205.11463` doi: 10.48550/arXiv.2205.11463

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current biology: CB*, *28*(18), 2867–2875.e3. doi: 10.1016/j.cub.2018.07.023

Lago, S., Acuña Fariña, C., & Meseguer, E. (2021). The reading signatures of agreement attraction. *Open Mind*, *5*, 132–153. doi: 10.1162/opmi_a_00047

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279–292. doi: 10.1016/j.neuron.2006.12.011

Lakatos, P., Gross, J., & Thut, G. (2019). A new unifying account of the roles of neuronal entrainment. *Current Biology*, *29*(18), R890–R905. doi: 10

.1016/j.cub.2019.07.075

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–113. doi: 10.1126/science.1154735

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193. doi: 10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. doi: 10.1152/jn.90896.2008

Lam, N. H., Schoffelen, J. M., Uddén, J., Hultén, A., & Hagoort, P. (2016). Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. *NeuroImage*, *142*, 43–54. doi: 10.1016/j.neuroimage.2016.03.007

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in cognitive sciences*, *18*(9), 472–479. doi: 10.1016/j.tics.2014.05.001

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

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In (p. 234). Honolulu, Hawaii: Association for Computational Linguistics. Retrieved from `http://portal.acm.org/citation.cfm?doid=1613715.1613749` doi: 10.3115/1613715.1613749

Levy, R., & Gibson, E. (2013). Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*(MAY), 229. doi: 10.3389/FPSYG.2013.00229/BIBTEX

Lewis, R. L., Vasishth, S., & Dyke, J. A. V. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454. doi: 10.1016/j.tics.2006.08.007

León-Cabrera, P., Rodríguez-Fornells, A., & Morís, J. (2017). Electrophysiological correlates of semantic anticipation during speech comprehension. *Neuropsychologia*, *99*, 326–334. doi: 10.1016/j.neuropsychologia.2017.02.026

Li, J., & Hale, J. T. (2019). Grammatical predictors for fmri time-courses.

In R. C. Berwick & E. P. Stabler (Eds.), *Minimalist parsing* (p. 0). Oxford University Press. Retrieved from https://doi.org/10.1093/oso/9780198795087.003.0007

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382–1411. doi: 10.1111/COGS.12274

Linzen, T., Siegelman, N., & Bogaerts, L. (2017). Prediction and uncertainty in an artificial language. In *Cogsci 2017 - proceedings of the 39th annual meeting of the cognitive science society* (pp. 2592–2597).

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th international conference on language resources and evaluation*.

Liu, Y., Shu, H., & Wei, J. (2006). Spoken word recognition in context: Evidence from chinese erp analyses. *Brain and Language*, *96*(1), 37–48. doi: 10.1016/J.BANDL.2005.08.007

Lo, C.-W., Tung, T.-Y., Ke, A. H., & Brennan, J. R. (2022). Hierarchy, not lexical regularity, modulates low-frequency neural synchrony during language comprehension. *Neurobiology of Language*, *3*(4), 538–555. doi: 10.1162/nol_a_00077

Loerts, H., Stowe, L. A., & Schmid, M. S. (2013). Predictability speeds up the re-analysis process: An erp investigation of gender agreement and cloze probability. *Journal of Neurolinguistics*, *26*(5), 561–580. doi: 10.1016/j.jneuroling.2013.03.003

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*, 1166–1183. doi: 10.1111/cogs.12597

Lu, Y., Jin, P., Pan, X., & Ding, N. (2022). Delta-band neural activity primarily tracks sentences instead of semantic properties of words. *NeuroImage*, *251*, 118979. doi: 10.1016/j.neuroimage.2022.118979

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology*, *7*(2), 183–188. doi: 10.1037/1089-2680.7.2.183

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. doi: 10.1016/j.cogpsych.2016.06.002

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010.

doi: 10.1016/j.neuron.2007.06.004

Maheu, M., Meyniel, F., & Dehaene, S. (2022). Rational arbitration between statistics and rules in human sequence processing. *Nature Human Behaviour 2022*, 1–17. doi: 10.1038/s41562-021-01259-6

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. doi: 10.1016/j.cognition.2012.09.010

Mai, G., & Wang, W. S.-Y. (2023). Distinct roles of delta- and theta-band neural tracking for sharpening and predictive coding of multi-level speech features during spoken language processing. *Human Brain Mapping*, *44*(17), 6149–6172. doi: 10.1002/hbm.26503

Mancini, S., Postiglione, F., Laudanna, A., & Rizzi, L. (2014). On the person-number distinction: Subject-verb agreement processing in italian. *Lingua*, *146*, 28–38. doi: 10.1016/j.lingua.2014.04.014

Marcus, G., Vijayan, S., Bandi Rao, S., & Vishton, P. (1999). Rule learning by seven-month old infants. *Science*, *283*, 77–80. doi: 10.1126/science.283.5398.77

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg- and meg-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/J.JNEUMETH.2007.03.024

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102. doi: 10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. doi: 10.1016/0010-0277(80)90015-3

Marslen-Wilson, W. D., & Tyler, L. K. (2007). Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 823–836. doi: 10.1098/rstb.2007.2091

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. doi: 10.1016/0010-0285(78)90018-X

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*(February), 1–17. doi: 10.3389/fpsyg.2016.00120

Martin, A. E. (2018). Cue integration during sentence comprehension: Electro-

physiological evidence from ellipsis. *PLOS ONE*, *13*(11), e0206616. doi: 10.1371/journal.pone.0206616

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427. doi: 10.1162/jocn_a_01552

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology*, *15*(3), 1–23. doi: 10.1371/journal.pbio.2000663

Martin, A. E., & Doumas, L. A. (2019a). Predicate learning in neural systems: using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, *29*, 77–83. doi: 10.1016/j.cobeha.2019.04.008

Martin, A. E., & Doumas, L. A. A. (2019b). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190306. doi: 10.1098/rstb.2019.0306

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., & McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of experimental psychology. Learning, memory, and cognition*, *35*(5), 1231–1239. doi: 10.1037/a0016271

Martin, A. E., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: Evidence from sluicing. *Journal of memory and language*, *64*(4), 327–343. doi: 10.1016/j.jml.2010.12.006

Martin, A. E., Monahan, P. J., & Samuel, A. G. (2017). Prediction of agreement and phonetic overlap shape sublexical identification. *Language and Speech*, *60*(3), 356–376. doi: 10.1177/0023830916650714

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), 663–678. doi: 10.1002/hbm.24403

Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*(3), 1481–1498. doi: 10.1093/CERCOR/BHZ180

Matchin, W., Liao, C.-H., Gaston, P., & Lau, E. (2019). Same words, different structures: An fmri investigation of argument relations and the angular gyrus. *Neuropsychologia*, *125*, 116–128. doi: 10.1016/j.neuropsychologia.2019.01.019

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recog-

nition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. doi: 10.1080/01690965.2012.705006

Mazerolle, M. J. (2020). *Aiccmodavg: Model selection and multimodel inference based on (q)aic(c)*. Retrieved from `https://cran.r-project.org/package=AICcmodavg`

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51. doi: 10.1037/REV0000126

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi: 10.1016/0010-0285(86)90015-0

McGinnies, E., Comer, P. B., & Lacey, O. L. (1952). Visual-recognition thresholds as a function of word length and word frequency. *Journal of Experimental Psychology*, *44*(2), 65–69. doi: 10.1037/h0063142

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. doi: 10.1126/science.1245994

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621. doi: 10.1111/ejn.13748

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, *27*(9), 4293–4302. doi: 10.1093/cercor/bhw228

Meyer, L., Sun, Y., & Martin, A. E. (2020a). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089–1099. doi: 10.1080/23273798.2019.1693050

Meyer, L., Sun, Y., & Martin, A. E. (2020b). "entraining" to speech, generating language? *Language, Cognition and Neuroscience*, *35*(9), 1138–1148. doi: 10.1080/23273798.2020.1827155

Molinaro, N., Barraza, P., & Carreiras, M. (2013). Long-range neural synchronization supports fast and efficient reading: Eeg correlates of processing expected words in sentences. *NeuroImage*, *72*, 120–132. doi: 10.1016/j.neuroimage.2013.01.031

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650. doi: 10.1111/ejn.13811

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In (p. 398–408). USA: Association for Computational Linguistics.

Monte-Ordoño, J., & Toro, J. M. (2017). Early positivity signals changes in an abstract linguistic pattern. *PLoS ONE*, *12*(7), 1–14. doi: 10.1371/journal.pone.0180727

Moore-Cantwell, C. (2013). Syntactic predictability influences duration. *Proceedings of Meetings on Acoustics*, *19*(1), 060206. doi: 10.1121/1.4801075

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. doi: 10.1037/H0027366

Nelson, M. J., Dehaene, S., Pallier, C., & Hale, J. T. (2017). Entropy reduction correlates with temporal lobe activity. In T. Gibson, T. Linzen, A. Sayeed, M. van Schijndel, & W. Schuler (Eds.), (p. 1–10). Valencia, Spain: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W17-0701` doi: 10.18653/v1/W17-0701

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., ... Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), E3669–E3678. doi: 10.1073/pnas.1701590114

New, B., ferrand, L., pallier, C., & brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. doi: 10.3758/BF03193811

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance ii. statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *49*(2), 85–117. doi: 10.1016/j.cogpsych.2003.12.002

Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, *36*(4), 569–587. doi: 10.1006/jmla.1996.2497

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. doi: 10.1016/0010-0277(94)90043-4

Norris, D., & McQueen, J. M. (2008). Shortlist b: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. doi: 10.1037/0033-295X.115.2.357

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. (2011). Fieldtrip: Open source software for advanced analysis of meg, eeg and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*. doi: doi:10.1155/2011/156869

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773. doi: 10.1006/jmla.1995.1033

Ouyang, L., Boroditsky, L., & Frank, M. C. (2017). Semantic coherence facilitates distributional learning. *Cognitive Science*, *41*(S4), 855–884. doi: 10.1111/cogs.12360

Pascanu, R., & Jaeger, H. (2011). A neurodynamical model for working memory. *Neural Networks*, *24*(2), 199–207. doi: 10.1016/j.neunet.2010.10.003

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*(3), 427–456. doi: 10.1006/jmla.1999.2653

Peña, M., & Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience*, *24*(5), 1149–1164. doi: 10.1162/jocn_a_00144

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. doi: 10.1073/pnas.1012551108

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R., & Cotterell, R. (2022). On the effect of anticipation on reading times. *arXiv*. Retrieved from `http://arxiv.org/abs/2211.14301` doi: 10.48550/arXiv.2211.14301

Pollock, J.-Y. (1989). Verb movement, universal grammar, and the structure of ip. *Linguistic Inquiry*, *20*(3), 365–424.

Postman, L., & Adis-Castro, G. (1957). Psychophysical methods in the study of word recognition. *Science*, *125*, 193–194. doi: 10.1126/science.125.3240.193

Pulvermüller, F., & Assadollahi, R. (2007). Grammar or serial order?: Discrete combinatorial brain mechanisms reflected by the syntactic mismatch neg-

ativity. *Journal of Cognitive Neuroscience*, *19*(6), 971–980.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. doi: 10.1126/science.aax0050

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences*, *22*(10), 870–882. doi: 10.1016/j.tics.2018.08.003

Rizzi, L. (1997). The fine structure of the left periphery. In *Elements of grammar* (pp. 281–337).

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 324–333). Singapore: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D09-1034`

Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition*, *125*(1), 49–63. doi: 10.1016/j.cognition.2012.06.008

Ryu, S. H., & Lewis, R. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In (p. 61–71). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.cmcl-1.6` doi: 10.18653/v1/2021.cmcl-1.6

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515. doi: 10.1006/jmla.2000.2759

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–8. doi: 10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. doi: 10.1006/jmla.1996.0032

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52–63. doi: 10.1016/j.tics.2017.10.003

Sassenhagen, J. (2019). How to analyse electrophysiological responses to nat-

uralistic language with time-resolved multiple regression. *Language, Cognition and Neuroscience*, *34*(4), 474–490. doi: 10.1080/23273798.2018.1502458

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of meg/eeg data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6). doi: 10.1111/psyp.13335

Sauseng, P., Klimesch, W., Gruber, W. R., Hanslmayr, S., Freunberger, R., & Doppelmayr, M. (2007). Are event-related potential components generated by phase resetting of brain oscillations? a critical discussion. *Neuroscience*, *146*(4), 1435–1444. doi: 10.1016/j.neuroscience.2007.03.014

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fmri study on two-word phrasal processing. *Cortex*, *96*, 105–120. doi: 10.1016/j.cortex.2017.09.002

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*(6), 1270–1281. doi: 10.3758/BF03201199

Schoffelen, J. M., Oostenveld, R., Lam, N. H., Uddén, J., Hultén, A., & Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, *6*(1). doi: 10.1038/s41597-019-0020-y

Schuberth, R. E., & Eimas, P. D. (1977). Effects of context on the classification of words and nonwords. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 27. doi: 10.1037/0096-1523.3.1.27

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In (pp. 92–96). Austin, Texas. Retrieved from `https://conference.scipy.org/proceedings/scipy2010/seabold.html` doi: 10.25080/Majora-92bf1922-011

Senoussi, M., Verbeke, P., & Verguts, T. (2022). Time-based binding as a solution to and a limitation for flexible cognition. *Frontiers in Psychology*, *12*. Retrieved from `https://www.frontiersin.org/articles/10.3389/fpsyg.2021.798061`

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sharpe, V., Reddigari, S., Pylkkänen, L., & Marantz, A. (2018). Automatic access to verb continuations on the lexical and categorical levels: evidence from meg. *Language, Cognition and Neuroscience*, *34*(2), 137–150. doi: 10

.1080/23273798.2018.1531139

Sheather, S. J. (2009). Diagnostics and transformations for multiple linear regression. In S. Sheather (Ed.), *A modern approach to regression with r* (pp. 151–225). New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-0-387-09608-7_6

Simpson, G. B., Peterson, R. R., Casteel, M. A., & Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 88. doi: 10.1037/0278-7393.15.1.88

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

Slaats, S., Weissbart, H., Schoffelen, J.-M., Meyer, A. S., & Martin, A. E. (2023). Delta-band neural responses to individual words are modulated by sentence processing. *Journal of Neuroscience*, *43*(26), 4867–4883. doi: 10.1523/JNEUROSCI.0964-22.2023

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. doi: 10.1016/j.cognition.2013.02.013

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), 8443–8453. doi: 10.1523/jneurosci.5069-11.2012

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In (pp. 901–904). ISCA. Retrieved from https://www.isca-speech.org/archive/icslp_2002/stolcke02_icslp.html doi: 10.21437/ICSLP.2002-303

Tanner, D., & Bulkes, N. Z. (2015). Cues, quantification, and agreement in language comprehension. *Psychonomic Bulletin & Review*, *22*(6), 1753–1763. doi: 10.3758/s13423-015-0850-3

Tanner, D., Grey, S., & van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the p600 during sentence comprehension. *Psychophysiology*, *54*(2), 248–259. doi: 10.1111/psyp.12788

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215. doi: 10.1016/j.jml.2014.07.003

Tavano, A., Blohm, S., Knoop, C. A., Muralikrishnan, R., Fink, L., Scharinger, M.,

... Poeppel, D. (2022). Neural harmonics of syntactic structure. *bioRxiv*, 2020.04.08.031575. doi: 10.1101/2020.04.08.031575

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. doi: 10.1126/science.1192788

Ten Oever, S., Carta, S., Kaufeld, G., & Martin, A. E. (2022). Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. *eLife*, *11*, e77468. doi: 10.7554/eLife.77468

Ten Oever, S., Kaushik, K., & Martin, A. E. (2022). Inferring the nature of linguistic computations in the brain. *PLOS Computational Biology*, *18*(7), e1010269. doi: 10.1371/journal.pcbi.1010269

Ten Oever, S., & Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *eLife*, *10*, e68066. doi: 10.7554/eLife.68066

Ten Oever, S., & Martin, A. E. (2024). Interdependence of "what" and "when" in the brain. *Journal of cognitive neuroscience*, *36*(1), 167–186.

Ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences*, *112*(52), 15833–15837. doi: 10.1073/pnas.1517519112

Ten Oever, S., Titone, L., Te Rietmolen, N., & Martin, A. E. (2024). Phase-dependent word perception emerges from region-specific sensitivity to the statistics of language. *Proceedings of the National Academy of Sciences*, *121*(23), e2320489121.

Tezcan, F., Weissbart, H., & Martin, A. E. (2023). A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. *eLife*, *12*, e82386. doi: 10.7554/eLife.82386

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. doi: 10.1080/15475440709336999

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249–267. doi: 10.1016/J.WOCN.2018.09.004

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2011). Generalizing linguistic structures under high attention demands. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(2), 493–501. doi: 10.1037/a0022056

Traxler, M. J. (2005). Plausibility and verb subcategorization in temporarily

ambiguous sentences: Evidence from self-paced reading. *Journal of Psycholinguistic Research*, *34*(1), 1–30. doi: 10.1007/s10936-005-3629-2

Trecca, F., McCauley, S. M., Andersen, S. R., Bleses, D., Basbøll, H., Højen, A., ... Christiansen, M. H. (2019). Segmentation of highly vocalic speech via statistical learning: Initial results from danish, norwegian, and english. *Language Learning*, *69*(1), 143–176. doi: 10.1111/lang.12325

Tung, T.-Y., & Brennan, J. R. (2023). Expectations modulate retrieval interference during ellipsis resolution. *Neuropsychologia*, 108680. doi: 10.1016/j.neuropsychologia.2023.108680

Tyler, L. K., Voice, J. K., & Moss, H. E. (2000). The interaction of meaning and sound in spoken word recognition. *Psychonomic Bulletin & Review*, *7*(2), 320–326. doi: 10.3758/BF03212988

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics 1983 34:5*, *34*(5), 409–420. doi: 10.3758/BF03203056

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*(4), 562–579. doi: 10.1037/0012-1649.22.4.562

Vallat, R. (2018). Pingouin: statistics in python. *The Journal of Open Source Software*, *3*, 1026. doi: 10.21105/joss.01026

van Alphen, P., & McQueen, J. M. (2001). The time-limited influence of sentential context of function word identification. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1057–1071. doi: 10.1037/0096-1523.27.5.1057

van den Bosch, A., & Berck, P. (2009). Memory-based machine translation and language modeling. *Prague Bulletin of Mathematical Linguistics*, *91*, 17–26.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. doi: 10.1016/j.jml.2006.03.007

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988. doi: 10.1111/cogs.12988

van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In (pp. 1597–1605). Denver, Colorado: Association for Compu-

tational Linguistics. Retrieved from http://aclweb.org/anthology/N15-1183 doi: 10.3115/v1/N15-1183

Vasishth, R. L. L., Shravan. (2001). An activation-based model of sentence processing as skilled memory retrieval. In *Dictionary of world philosophy*. Routledge.

Verga, L., Sroka, M. G. U., Varola, M., Villanueva, S., & Ravignani, A. (2022). Spontaneous rhythm discrimination in a mammalian vocal learner. *Biology Letters*, *18*(10), 20220316. doi: 10.1098/rsbl.2022.0316

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00002

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... van Mulbregt, P. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272. doi: 10.1038/s41592-019-0686-2

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, *45*(6), 1611–1617. doi: 10.1037/a0016134

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. doi: 10.1016/j.jml.2009.04.002

Wang, L., Zhu, Z., & Bastiaansen, M. (2012). Integration or predictability? a further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in Psychology*, *3*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00187

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. doi: 10.1126/science.167.3917.392

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi: 10.21105/joss.03021

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2019). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. doi: 10.1162/jocn_a_01467

Weissbart, H., & Martin, A. E. (2023). The structure and statistics of language jointly shape cross-frequency dynamics during spoken language comprehension. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/2023.10.06.561087v1 doi: 10.1101/2023.10.06

.561087

Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, *129*, 104400. doi: 10.1016/j.jml.2022.104400

Zioga, I., Weissbart, H., Lewis, A. G., Haegens, S., & Martin, A. E. (2023). Naturalistic spoken language comprehension is supported by alpha and beta oscillations. *Journal of Neuroscience*, *43*(20), 3718–3732. doi: 10.1523/JNEUROSCI.1500-22.2023

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980–991. doi: 10.1016/j.neuron.2012.12.037

Nederlandse samenvatting

Een belangrijk aspect van het menselijk taalvermogen is de syntaxis: ons vermogen om woorden zo te combineren dat de resulterende combinatie een specifieke betekenis heeft. Hierdoor betekent “Roos zoekt Willem-Jan” iets anders dan “Willem-Jan zoekt Roos”. Onze woorden kunnen op oneindig veel manieren worden gecombineerd: we kunnen hele korte zinnen begrijpen, en hele lange, en zinnen die we nog nooit eerder hebben gehoord. Het menselijk vermogen om zinsstructuren te creëren en te analyseren is – voor zover we weten – ongeëvenaard. Maar hoe doen we het?

Er zijn meerdere theorieën over welke (neurale) mechanismen aan dit vermogen ten grondslag liggen. In mijn proefschrift focus ik op twee van deze theorieën. De eerste beschouwt ons vermogen om syntactische structuren te creëren als het resultaat van het leren en gebruiken van statistische informatie, zoals hoe waarschijnlijk een woord is in zijn context. Volgens deze theorie is het gebruik van een abstracte structuur niet (altijd) nodig om te begrijpen wat er wordt gezegd; we kunnen dit met kennis over de waarschijnlijkheid van woorden alleen. De andere theorie modelleert ons syntactische vermogen als de creatie van een afzonderlijk representatieniveau dat een hiërarchische structuur heeft en abstraheert van de woorden zelf. Volgens deze theorie is kennis van de syntaxis niet gebonden aan de specifieke woorden of morfemen. In plaats daarvan zijn de regels van toepassing op syntactische categorieën, zoals het zelfstandig naamwoord of het werkwoord. De interpretatie van een zin hangt af van dit abstracte systeem.

Deze twee opvattingen lijken tegenover elkaar te staan in de literatuur. Tegelijkertijd hebben talloze experimenten bewijs geleverd voor de relevantie van beide soorten kennis in het proces van taalbegrip. Een goede theorie over hoe we taal begrijpen moet daarom aspecten bevatten van beide theorieën: menselijke hersenen zijn ontzettend goed in statistiek én ze zijn in staat om abstracte representaties te produceren. In dit proefschrift benader ik taalbegrip vanuit dit perspectief. Ik heb onderzocht hoe statistische informatie over woorden en syntactische informatie gezamenlijk het proces van taalbegrip vormgeven. Het onderzoeken van deze vraag kan ons helpen uit te vinden welke mechanismen een

rol spelen in het menselijk vermogen om zinsstructuren af te leiden uit spraak en tekst.

Ik heb dit vraagstuk van verschillende kanten benaderd. In **hoofdstuk 2** bekeek ik statistische informatie vanuit een theoretisch perspectief. Ik keek naar 'lexical surprisal'. Dit is een getal dat kwantificeert hoe verrassend (of onwaarschijnlijk) een woord statistisch gezien is in een bepaalde context. In dit hoofdstuk onderzocht ik ten eerste waarom lexical surprisal goed werkt als voorspeller van allerlei soorten data, bijvoorbeeld hoe snel mensen woorden lezen. Ik beargumenteerde dat dit komt doordat statistische informatie over woorden variatie kan reflecteren uit allerlei verschillende bronnen ('latente variabelen'), waaronder de syntactische structuur. Ik heb dit laten zien door middel van een simulatie met een simpele grammatica. Ten tweede vroeg ik me af wat de resultaten van onderzoeken die lexical surprisal als voorspeller gebruikten ons kunnen vertellen over het proces van taalbegrip. Ik concludeerde dat deze resultaten geen directe theoretische inzichten geven, juist omdat lexical surprisal geen onderscheid maakt in de latente variabelen (zoals woordfrequentie en syntactische structuur) die de surprisal-waarden bepalen. Dit is geen probleem als het onderzoek uitsluitend gericht is op het voorspellen van de data. Het wordt wel een probleem als we de resultaten willen gebruiken om een theorie over taalbegrip te maken.

In **hoofdstuk 3** onderzocht ik of de aanwezigheid van syntactische structuur invloed heeft op de manier waarop het brein reageert op woorden. Ik deed dit door hersenscans gemaakt met magneto-encefalografie (MEG) te analyseren van mensen die luisterden naar zinnen (met syntactische structuur) en woordenlijsten (zonder syntactische structuur). Door een specifieke implementatie van lineaire regressie kon ik uit deze scans de hersenresponsen op individuele woorden onderscheiden van de rest van de hersenactiviteit. Ik vond dat de respons op woorden in woordenlijsten met ongeveer 350 milliseconden vertraagd was ten opzichte van de respons op woorden in zinnen. Bovendien was de informatie over deze woorden beter vertegenwoordigd in het signaal wanneer het woord in zin stond. Dit betekent dat we het makkelijker vinden om woorden te herkennen in een gestructureerde zinscontext.

In **hoofdstuk 4** benaderde ik de relatie tussen statistische informatie en structuur andersom: ik onderzocht of de waarschijnlijkheid van een woord in context het gebruik van syntactische informatie beïnvloedt. Om specifieker te zijn, onderzocht ik of lexical surprisal invloed had op de berekening van de relatie tussen het onderwerp en het werkwoord als mensen zinnen lezen. Dit deed ik door

de surprisal en de grammaticaliteit van het onderwerp te variëren in vier condities. De resultaten leverden geen duidelijk bewijs voor een interactie tussen surprisal en grammaticaliteit: de moeilijkheid van het lezen van een ongrammaticaal onderwerp was niet verminderd als het onderwerp zeer voorspelbaar was. De resultaten gaven echter wel aan dat het beste model van de gegevens een expliciete specificatie van grammaticaliteit vereist; alleen lexical surprisal is niet genoeg. De resultaten van dit onderzoek suggereerden dat taalbegrip sterk wordt bepaald door grammaticaliteit.

In **hoofdstuk 5** vroeg ik opnieuw of lexicale statistische informatie de syntactische verwerking beïnvloedt, maar deze keer met dezelfde aanpak als in **hoofdstuk 3**. In deze studie analyseerde ik MEG-gegevens van mensen die in de scanner naar verhalen luisterden. Zoals in **hoofdstuk 3**, heb ik met behulp van regressiemodellen de respons van het brein op de syntactische structuur van een zin geïsoleerd. Deze heb ik verdeeld over twee groepen: de respons op woorden die erg voorspelbaar waren, en de respons op woorden die niet voorspelbaar waren. Deze hersenresponsen heb ik toen met elkaar vergeleken. De resultaten toonden aan dat de waarschijnlijkheid van een woord (gegeven de context) het tijdsverloop van het bouwen van een structuur beïnvloedt: de reactie die gepaard gaat met het bouwen van een structuur wordt met maar liefst 150 milliseconden vertraagd voor woorden die gezien de context onverwacht zijn vergeleken met woorden die gezien de context waarschijnlijker zijn. Dit betekent dat we de structuur van een zin makkelijker kunnen creëren als het woord dat we horen statistisch voorspelbaar was.

Hoofdstuk 6 geeft een overzicht van verschillende reeksen simulaties die de analyses uit de **hoofdstukken 3** en **5** zowel leidden als aanvulden. Het doel van de simulaties was om te beoordelen of eventuele effecten gevonden in de analyses uit de **hoofdstukken 3** en **5** toe te schrijven waren aan eigenschappen van de gegevens of het lineaire model die geen verband hielden met het theoretische fenomeen in kwestie. Deze simulaties helpen bij de interpretatie van de resultaten van de andere hoofdstukken. De simulaties laten zien dat de resultaten betrouwbaar zijn, maar ook dat de regressiemethode geen vertragingen of versnellingen van een hersenrespons door interacties tussen variabelen kan modelleren. De enige manier om tijdsverschuivingen vast te leggen is door afzonderlijke condities te creëren, zoals ik gedaan heb in dit proefschrift.

In **hoofdstuk 7** breng ik de resultaten uit dit proefschrift samen. De onderzoeken in dit proefschrift hebben twee belangrijke aspecten van het proces van taalbegrip laten zien. Ten eerste bepaalt grammaticale kennis hoe we reageren

op taal, zowel als het gaat om de hersenrespons als om hoe snel we woorden lezen. Ten tweede is de invloed van zowel statistische informatie als syntactische informatie in het brein zichtbaar als een vertraging of versnelling van de reactie. Wanneer nieuwe taalkundige informatie niet goed aansluit bij de huidige status van het brein, die onder andere wordt bepaald door statistische en syntactische informatie van het interne taalmodel, wordt de neurale respons vertraagd. Deze bevindingen suggereren dat de dimensie van tijd cruciaal is voor de combinatie van deze twee soorten informatie. Daarom heb ik in dit hoofdstuk het model BiMCON ('Binding in a Model Constrained Oscillatatory Network') voorgesteld. Dit model is een combinatie van de eerdere modellen STiMCON en time-based binding. Het maakt gebruik van de tijdsdimensie om te beschrijven hoe lexicale statistische informatie het proces van de opbouw van syntactische structuur kan beïnvloeden, en hoe beide soorten kennis invloed hebben op de staat van het brein.

English Summary

An important aspect of human language ability is syntax: our ability to combine words in such a way that the resulting combination has a specific meaning. This means that "Roos is looking for Willem-Jan" means something different from "Willem-Jan is looking for Roos". Our words can be combined in infinitely many ways: we can understand very short sentences, very long ones, and sentences we have never heard before. The human ability to create and analyze sentence structures is – as far as we know – unparalleled. But how do we do it?

There are several theories about which (neural) mechanisms underlie this ability. In my dissertation I focus on two of these theories. The first views our ability to create syntactic structures as the result of learning and using of statistical information, such as how likely a word is in its context. According to this theory, the use of an abstract structure is not (always) necessary to understand what is being said; we can do this with statistical information alone. The other theory models our syntactic ability as the creation of a separate level of representation that has a hierarchical structure and abstracts away from the words themselves. According to this theory, knowledge of syntax is not tied to the specific words or morphemes. The interpretation of a sentence depends on this abstract system. These two views are often opposed to each other in the literature. At the same time, numerous experiments have provided evidence for the relevance of both types of knowledge in the process of language comprehension. A good theory about how we understand language must therefore contain aspects of both theories: human brains are very good at statistics, and they are able to produce abstract representations. In this dissertation, I approach language understanding from this perspective. I have investigated how statistical information about words and syntactic information jointly shape the process of language understanding. Investigating this question can help us find out what mechanisms are involved in the human ability to infer sentence structures from speech and text.

I have approached this issue in different ways. In **Chapter 2** I looked at statistical information from a theoretical perspective. I looked at 'lexical surprisal'. This is a number that quantifies how surprising (or unlikely) a word statistically is, in a given context. In this Chapter I first investigated why lexical surprisal

works well as a predictor of all kinds of data, for example, how fast people read words. I argued that this is because statistical information about words can reflect variation from many different sources ('latent variables'), including syntactic structure. I have shown this through a simulation with a simple grammar. Second, I wondered what the results of studies that used lexical surprisal as a predictor can tell us about the process of language comprehension. I concluded that these results do not directly provide theoretical insights, precisely because lexical surprisal does not distinguish between the latent variables (such as word frequency and syntactic structure) that determine the surprisal values. This is not a problem if the research is exclusively aimed at predicting the data. It does become a problem if we want to use the results to build a theory about language comprehension.

In **Chapter 3** I investigated whether the presence of syntactic structure influences the way the brain responds to words. I did this by analyzing brain scans made with magnetoencephalography (MEG) of people who listened to sentences (with syntactic structure) and lists of words (without syntactic structure). Through a specific analysis technique, an implementation of linear regression, I was able to distinguish the brain responses to individual words from the rest of the brain activity. I found that the response to words in word lists was delayed by about 300 milliseconds compared to the response to words in sentences. Furthermore, the information about these words was better represented in the signal when the word was in the sentence. This means that we find it easier to recognize words in a structured sentence context.

In **Chapter 4** I approached the relationship between statistical information and structure the other way around: I investigated whether the probability of a word in context influences the use of syntactic information. To be more specific, I investigated whether lexical surprisal influenced the computation of the relationship between the subject and the verb when people read sentences. I did this by varying the surprisal and grammaticality of the subject in four conditions. The results did not provide clear evidence for an interaction between surprisal and grammaticality: the difficulty of reading an ungrammatical subject was not reduced when the subject was highly predictable. However, the results did indicate that the best model of the data requires an explicit specification of grammaticality; lexical surprisal alone is not enough. The results of this study suggested that language comprehension is strongly determined by grammaticality.

In **Chapter 5** I asked again whether lexical statistical information influences syntactic processing, but this time using the same approach as in **Chapter 3**. In this study I analyzed MEG data from people listening to stories in the scanner. As in **Chapter 3**, I used regression models to isolate the brain response to the syntactic structure of a sentence. I divided these responses into two groups: the response to structure for words that were very predictable, and the response to structure for words that were not predictable. I then compared these brain responses with each other. The results showed that the probability of a word (given the context) influences the time course of building a sentence structure: the response associated with building a structure is delayed by as much as 150 milliseconds for words that are unexpected given the context are compared to words that are more likely given the context. This means that we can create the structure of a sentence more easily if the word we hear was statistically predictable.

Chapter 6 provides an overview of several sets of simulations that both guided and complemented the analyzes in Chapters **3** and **5**. The purpose of the simulations was to assess whether any effects found in the analyzes of Chapters **3** and **5** were due to properties of the data or the linear model that were unrelated to the theoretical phenomenon in question. These simulations help interpret the results of the other Chapters. The simulations show that the results are reliable, but also that the regression method cannot model delays or accelerations of a brain response due to interactions between variables. The only way to capture time shifts is to create separate conditions, as I did in this dissertation.

In Chapter **7** I bring together the results from this dissertation. The studies in this dissertation have revealed two important aspects of the language comprehension process. Firstly, grammatical knowledge determines how we respond to language, both in terms of brain response and how quickly we read words. Secondly, the influence of both statistical information and syntactic information in the brain is visible as a slowing down or speeding up of the response. When new linguistic information does not match well with the current state of the brain, which is determined, among other things, by statistical and syntactic information from the internal language model, the neural response is delayed. These findings suggest that the dimension of time is crucial for the combination of these two types of information. That is why I have presented the BiMCON ('Binding in a Model Constrained Oscillatatory Network') model in this Chapter. This model is a combination of the previous model *STiMCON* and time-based binding models. It uses the temporal dimension to describe how lexical statistical information

can influence the process of building syntactic structure, and how both types of knowledge influence the state of the brain.

Research data management

Data availability

Three Chapters in this thesis contain experimental data, and two others contain simulated data. The behavioral data of Chapter 4 were acquired at the Max Planck Institute for Psycholinguistics. This dataset has been archived at the MPI for Psycholinguistics Archive. I provide the persistent identifier to the corresponding collection below. The MEG data of Chapters 3 and 5 were acquired at the Donders Centre for Cognitive Neuroimaging. These datasets have been archived at the Donders Repository. I provide the identifiers under the corresponding Chapters. The simulated data from Chapters 2 and 6 were archived on the Open Science Framework, and the code to generate the data is shared on GitHub. I provide links to these Repositories.

Chapter 2 Code: `https://github.com/sslaats/surprisal`. Data: `https://osf.io/xp3r7/`.

Chapter 3 Code: `https://osf.io/ky9bj/`. Data: `https://data.ru.nl/collections/di/dccn/DSC_3011020.09_236`.

Chapter 4 Code: `https://github.com/sslaats/surprisal-agreement`. Data: `https://hdl.handle.net/1839/fb1854a4-af77-4a27-bdcc-f8ec80a8ac82`.

Chapter 5 Data: `https://data.ru.nl/collections/di/dccn/DSC_3027007.01_206?0`.

Chapter 6 Code: `https://github.com/sslaats/trf-simulations`. Data: `https://osf.io/kwexj/`.

Acknowledgements

This PhD was a *process*: a move to a town that has zero houses for starters was followed by a multi-year global health crisis. The fact that you are reading this dissertation right now is made reality by many people other than myself; many people that kept me happy, (somewhat) sane and/or motivated during all those years. In this brief chapter I want to thank them - or most of them, at least. It is impossible to mention everyone by name. Considering that you are reading this, I want you to know that I am ever so grateful for your support.

First and foremost, I would like to express my gratitude towards my supervisors **Andrea** and **Antje**. **Andrea**, I am incredibly grateful for your support throughout the whole process. It has been a true privilege to work with you. Thank you for the inspiring meetings and for your advice on all things academia. Thank you for fostering communication in the Language and Computation in Neural Systems group when we were working from home; those lab meetings had a profound impact on my weeks. Most importantly: thank you for expressing your confidence in my abilities when I lacked it the most. **Antje**, thank you for your down-to-earth guidance, your great feedback on manuscripts, and for taking on the additional role as copromotor during Andrea's absence. Thank you for always being available for a chat, and for sharing my enthusiasm about walking. Also, not unimportant: thank you for all the PoL pancake hikes. I greatly enjoyed every single one.

I would also like to thank **Hugo** here. **Hugo**, you have no idea how much you helped me. Without you, I am absolutely certain this thesis would literally not exist. When COVID hit, my project was in crisis (as was our society), and you helped it get back on the rails. Thank you for telling me that I should not hedge my questions with "maybe this is a stupid question, but", for teaching me how to perform TRF-analysis from scratch, for meeting me in person when we were stuck at home, and for telling me over and over again that our project made sense. Thank you so much.

Thank you, **Hans Rutger**, for supporting me in the initial stages of my project, for always including me as a member of the TEMPOS-group, and for allowing me to meet your wonderful daughters.

I would also like to thank the members of my reading committee, **prof. Caroline Rowland**, **prof. David Poeppel**, and **dr. ir. Robert Oostenveld**, for the time they invested in evaluating this dissertation.

Cecília, **Caitlin**, **Orhun**, my paranymphs and dear dear friends, I am so happy to have met you. **Cecília**, thanks for being a fantastic office mate and friend. You had large shoes to fill (more on this later) and did this effortlessly from the very beginning; even our small post-it conversations before we met made me happy. The office was painfully empty when you were on leave. So thank you so much. Thanks for being the voice of reason when it came to responding to reviewers, and for all the brainstorms about your projects, and mine. I will never forget our lovely coffee walks on campus, nor those few longer hikes in the woods (or carrying a stroller through a muddy field). I will miss you dearly, and really hope we will work and walk together again. **Caitlin**, thanks for being my friend since the very first week I was at the institute. You made me feel so welcome in the PoL-group. I still remember our first good conversation, in a bar in Bottendaal, many years ago. You might not remember it, but from this moment, I felt part of the group of PoL PhDs. Thanks for all the (virtual) coffees and long chats, for words of advice about the workings of the institute when I first arrived, and for creating a warm atmosphere of trust. Not many people are capable of doing this. You are a great colleague and friend. I will miss you so much. **Orhun**, thanks for being the glue of the group. You joined PoL right after most previous PhDs left, and with your arrival a whole new wind blew through the group. I have so much to thank you for, starting with the pizza in the park and the Turkish dinners with rakıand beers at your house together with **Ronny** during COVID times, and of course all the Italian lunches. You brought everybody together. Thanks for inviting me to Wednesday drinks every week, even when I declined very often (I appreciated it every time). Thanks for our long hikes, the supply of a general meme vibe which I always appreciate, and the rants about Dutch politics. We should set up that Minecraft server soon.

Greta, thank you for becoming my friend basically immediately after we met. I vividly remember talking to you after one of the group lunches about the things that were bothering me at the start of my PhD, and you made me feel so understood. When the other spot in (y)our office opened up, I did not hesitate to ask if I could join you. What a good decision that was. I thoroughly enjoyed our office time: talking about language and neural oscillations, the social dynamics at the institute, and life in general. Of course we also did some work. Unfortunately our time together in the office was cut short by COVID, but I am happy we ma-

naged to keep in touch ever since. I look forward to seeing you in Switzerland! **Eirini**, thank you for letting me take over your super cute house when you moved. I am also eternally grateful for the numerous coffee breaks over Zoom with **Eirini**, **Greta** and **Caitlin**. These moments of connection kept me sane at home. **Laurel**, thank you for comforting me that time you found me panicking in the forest behind the institute, and for vigorously nodding during my first lab meeting presentations. **Morgane**, thanks for turning down the motor on your e-bike when we cycled together, and the chats about cats and Dutch men. **Ruth**, thanks for all the Reels and pictures of your dog(s), they have brought many smiles to my days. **Candice**, thanks for being such an understanding friend at the institute. **Sandra**, thank you for the screaming goat and all our chats. **Veerle**, thanks for your great energy. Thank you **Thy**, for checking in every morning when we were the only ones in the hallway. **Stan**, thank you for listening to my rants and for your advice as a response to them. Thanks to everyone, mentioned and not mentioned, in the past and present **Psychology of Language department** for being generally amazing. Thanks for the great vibes, the ice cream breaks, the pancake walks, your great feedback and questions during lab meetings, and for being a willing (and sometimes not-so-willing) audience for my stupid jokes (dad energy, **Kyla**?). I will miss you all so much.

A great round of thanks to all the members of the **Language and Computation in Neural Systems group**. I feel incredibly lucky to have worked with such kind and unbelievably intelligent people. Our virtual contact kept me sane(*ish*) during those years at home. **Ioanna**, **Filiz**, **Rong**, **Ryan**, **Noémie** and **Anna**, thank you for the great time at SNL 2023 in Marseille, and for the incredibly sweet collection and card after my backpack was stolen. I will never forget that you did this for me. **Sanne**, thanks for the good times supervising together and for answering all the questions I have asked you. Thank you **Cas**, for doing syntactic annotations together and for our other TRF-meetings. I am sorry for the confusing results, but it was (and will continue to be) interesting to talk about them.

Many other people from the MPI and the academic world outside of it have contributed to my PhD experience. **Laura Giglio**, thanks for the fun times writing for the Levelt grant and the AMLaP chats about the future. **Julia von der Fuhr**, thanks for the good times at MPI TalkLing. **Naomi**, thanks for the creative times and moments of connection. I would have loved to spend more time with you. **Limor**, thanks for our random evenings. And for saying "We should do this again! Next month?" every time. And we should do this again - for real!

Teun, thank you for the great time developing the Python course and the joint suffering of the UWV-madness. Thanks **Bissera** for becoming my friend after spending a few conferences together. Your enthusiasm is contagious. I hope we meet many more times! **Laura Fernandez Merino**, thank you for your friendship at the BCBL and beyond. Lots of thanks also to "my" students **Yangyi**, **Sofia**, **Julia**, and **Ilse**. You have taught me so much.

I would also like to mention some of my friends who were not my colleagues - who have always reminded me that there is more to life than academia. My dear *Utrechtse poepjes* **Victoria**, **Tessa**, **Anja**, **Mirjam** and **Carlijn**. Thank you for, well, pretty much everything. For the *gorgeous* drawings on Zoom, our Christmas dinners and New Years' celebrations. For MadNes festival. For sharing life in general. I am so happy that we remain friends, despite being scattered all over Europe. Dear **Lin An** and **Anne**, thank you for sticking around after the BA Linguistics, for painting my house, building a snowman, and the best pajama I own. **Chris** y **Anouk**, hola! Gracias for all the pizza and sangria (*jarra de neus*) and the Pitbull classics, *dale*. **Celeste** and **Roos**, the best housemates I could have ever imagined, thank you for our warm conversations and for always providing great advice over an Arrabiata with a Barbera. Every time we see each other feels like coming home to our 65m^2 in the center of Utrecht. Thank you, sweet **Fleur**, for stalking me at our pole dance school until we were friends. Luckily this was not a long process. Our friendship has been a great support for me during my PhD time in Nijmegen. Thank you for all our down-to-earth conversations about absolutely everything. Hopefully there will be many more. **Lisa**, thank you for the great times at MadNes festival, the countless coffees at the Maria Montessori building, and all (desperate) PhD conversations. **Lana**, *mi donostiarra holandesa*, thank you for all our moments of reflection in Donosti, Leiden, Nijmegen, Utrecht and Almere. You are a truly wonderful friend. Mi querida **Zoe**, *mi musa griega*. Thank you for believing in me from the moment I applied for this position. Thank you for being a constant in my life despite living so far away. Thanks for the letters, watching movies together over Zoom, and pizzas in bed in a hostel in Donosti. *Te echo de menos*.

My dearest sister **Seline**, I was not sure if I should add you to the list of family members or the list of friends; I think I can assign you to both at this point. We have been able to see each other so often the past few years: working together at my house during the first lockdown, painting your studio, bouldering, mining, and going for a high tea at the Millinger Theetuin. Thank you for your endless patience listening to my complaints. Ik hoop dat onze band voor altijd zo sterk

blijft. **Vati** en **Mutti**, my dear parents, I have absolutely loved living in the same city as you. It was great to be able to come over on any rainy afternoon, just for a cup of tea. See, I don't even know where to begin when thanking you, because you have *always* been there for me. So I will keep it simple: thank you for inspiring me to always pursue my interests. Thank you for listening and providing down-to-earth advice, about anything. Thank you for always walking with me (figuratively, and literally, sometimes as far as 52 kilometers from Nijmegen to Zutphen). I hope Tante Tok will frequently camp outside of Geneva. I would also like to express my gratitude to my grandparents. Dankjulliewel **opa Wil** en **oma Jet**, **oma Vera**, en natuurlijk ook **Theo**, voor alle liefde, steun en interesse in mijn studie en werk. *Prrrrr* and *brokjes* as a thank you to our cat **Nes** for all the serotonin she provided by simply existing.

Of course I have to save the best for last: my partner **Jeroen**. Thank you for being by my side throughout this whole process. When we met, on my third day back in the Netherlands, I never even imagined that we would live together and adopt a cat, let alone move to Switzerland together, but here we are. *Doing life* with you is absolutely wonderful. You have supported me in so many ways during my PhD that I cannot possibly list everything, but I will name a few things. Thank you for listening to the endless stories about my day, for convincingly admiring countless "squiggly lines" (a.k.a., temporal response functions), for telling me to go for a little walk (whoops, 830 kilometers), for reminding me that *I can*, for the cute notes and drawings, and for all our fantastic camping trips. Thank you for believing in me, always. Ik hou van je.

Curriculum Vitae

Sophie Slaats was born in Nijmegen, the Netherlands, in 1993. She obtained a bachelor degree in Linguistics and Phonetics, and a bachelor degree in Spanish Language and Culture from Utrecht University, the Netherlands, in 2016. She then obtained a Master's degree in Linguistics from Utrecht University in 2018, and a Master's degree in Cognitive Neuroscience of Language from the University of the Basque Country, Spain, in 2019. In 2019, she started her PhD project in the Psychology of Language department at the Max Planck Institute for Psycholinguistics, funded by a four-year PhD Fellowship from the IMPRS for Language Sciences. She is currently working as a postdoc at the University of Geneva, Switzerland.

Publications

Slaats, S., A.S. Meyer & A.E. Martin (*submitted*). Surprisal is not enough: Additive effects of grammaticality and lexical surprisal in self-paced reading.

Slaats, S. & A.E. Martin (*submitted*). What's surprising about surprisal.

Slaats, S., A.S. Meyer & A.E. Martin (*in press*). Lexical surprisal affects the time course of syntactic structure building. *Neurobiology of Language*.

Coopmans, C. W., A. Mai, **S. Slaats**, H. Weissbart & A.E. Martin (2023). What oscillations can do for syntax depends on your theory of structure building. *Nature Reviews Neuroscience*, 24(11), 723-723. https://doi.org/10.1038/s41583-023-00734-5

Slaats, S., H. Weissbart, J.M. Schoffelen, A.S. Meyer & A.E. Martin (2023). Delta-band neural responses to individual words are modulated by sentence processing. *Journal of Neuroscience*, 43 (26). https://doi.org/10.1523/JNEUROSCI.0964-22.2023

González Alonso, J., J. Alemán Bañón, V. DeLuca, D. Miller, S. M. Pereira Soares, E. Puig Mayenco, **S. Slaats** & J. Rothman (2020). Event related potentials at initial exposure in third language acquisition: Implications from an artificial mini-grammar study. *Journal of Neurolinguistics* (56). https://doi.org/10.1016/j.jneuroling.2020.100939